





Prediction of soil water retention curve based on physical characterization parameters using machine learning

Enzo Aldo Cunha Albuquerque¹ , Lucas Parreira de Faria Borges¹ ,
André Luís Brasil Cavalcante^{1#} , Sandro Lemos Machado² 

Article

Keywords

Soil water retention curve
Physical characterization
parameters
Machine learning
Python

Abstract

This paper explores the potential of machine learning techniques to predict the soil water retention curve based on physical characterization parameters. Results from 794 water retention and suction points obtained from 51 different soils were used in the algorithm. The soil properties used are the percentages of gravel, sand, silt, and clay, the plasticity index, the porosity, and the relation between the volumetric water content and total suction. The data were used as input for machine learning estimators to predict the volumetric water content of a soil with specified physical characterization parameters and suction, the techniques of artificial intelligence were developed in python. Results show that an extremely randomized trees' estimator can reach a coefficient of determination of 0.99 in the training dataset, with a coefficient of 0.90 in the cross-validation and testing dataset, which measures the generalization capacity. Furthermore, a continuous function can be obtained by fitting a model such as Cavalcante & Zornberg, or van Genuchten, or Costa & Cavalcante (bimodal) to the predictions of the machine learning for use in numerical methods. These results indicate that the proposed machine learning estimator can become an interesting alternative to estimate the soil water retention curve in engineering practice. This work is in progress and the predictions can be improved with the addition of new data. Know how to participate at the end of the paper.

1. Introduction

The soil water retention curve is a fundamental soil property that governs many agricultural, environmental, and engineering applications (Khlosi et al., 2016). However, conducting laboratory tests for soil water retention curve (SWRC) determination can be expensive and time-consuming, mainly in places located away from universities and research centers (Achieng, 2019; Haghverdi et al., 2015; Khlosi et al., 2016). An alternative for estimating this curve using physical characterization parameters such as percentage of gravel, sand, silt, clay, plasticity index (PI), and porosity (n) could contribute to a preliminary assessment of the soil water retention curve requiring less laboratory cost and time.

Other researchers developed procedures and techniques to assess or estimate the properties of unsaturated soils. Costa (2017) developed a model capable of representing the centrifuge permeameter test in order to facilitate obtaining

the hydraulic properties of a soil. He obtained the SWRC and the hydraulic conductivity of the soil requiring a shorter time compared to the filter paper and pressure plate tests and using a single moisture sensor. Arya & Paris (1981), Fredlund et al. (2002) and Vanapalli & Catana (2005) proposed models for estimating the SWRC using grain-size distribution curve and volume-mass properties.

Artificial intelligence is a novel technique that can be used for estimating the SWRC. It has been widely applied in geotechnics as examples are given. Ozelim et al. (2022) proposed a methodological framework to monitor internal erosion in dams based on artificial intelligence, which consist of processing the acoustic data obtained by geophones through artificial intelligence techniques in order to identify anomalies and classify the health status of the dam. Belcher et al. (2015), Fisher et al. (2016, 2017) investigated the erosion events, crack detection and anomaly detection in an experimental earth embankment seismic data using unsupervised, semi-supervised

#Corresponding author. E-mail address: albrasilc@gmail.com

¹Universidade de Brasília, Department of Civil and Environmental Engineering, Brasília, DF, Brasil.

²Universidade Federal da Bahia, Department of Materials Science and Technology, Salvador, BA, Brasil.

Submitted on January 6, 2022; Final Acceptance on June 14, 2022; Discussion open until November 30, 2022.

<https://doi.org/10.28927/SR.2022.000222>



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

and supervised machine learning techniques. Marjanović et al. (2011) and Tien Bui et al. (2016) used machine learning approach to assess problems of landslide susceptibility.

Machine learning techniques have been used in geotechnical engineering to predict engineering properties of soils based on previously known index properties. Some studies were performed in this subject to predict the compression index of soft soils from the Brazilian coast (Oliveira Filho et al., 2020), to define soil classes by the similarity of the CPT measurements (Carvalho & Ribeiro, 2019, 2020), to predict friction capacity of driven piles in cohesive soil (Prayogo & Susanto, 2018), to construct a site-specific multivariate probability distribution model of soil characteristics using Bayesian machine learning and hybridization between site-specific and generic data (Ching & Phoon, 2019).

Concerning the SWRC prediction using characterization results, one can cite the use of k-nearest-neighbors and sand, silt, and clay percentages and soil bulk density on samples from Belgium and UNSODA database to derive a pedotransfer function (PTF) for van Genuchten’s (1980) equation (Haghverdi et al., 2015); support vector machine (SVM) was also employed using a variety of parameters based on 72 samples from Syria to predict PTF at some SWRC points (Khlosi et al., 2016) and SVM and neural networks were applied to suction data to predict the SWRC for one loamy sand soil (Achieng, 2019).

This paper explores the potential use of machine learning techniques, such as extremely randomized trees, random forest, decision trees, logistic regression, support vector machine, multi-layer perceptron, and k-nearest neighbors, to predict the soil water retention curve for different soil types using physical characterization parameters. The database contains 794 measured SWRC points (main drying branch) and related soil characterization properties carried out on a wide variety of soils compiled by the authors. This dataset is divided into training, cross-validation, and test sets used, respectively, to fit, select and evaluate the model. Then, Cavalcante’s & Zornberg’s (2017), van Genuchten’s (1980) and Costa & Cavalcante (2021) functions are fitted to machine learning prediction to obtain a continuum function that can be used in other applications, such as numerical calculus. This study fits the Ordinance number

1.122 from Ministry of Science, Technology, Innovations and Communications from Brazil (Brasil, 2020), which establishes priorities for research, development, and innovation projects to enable technologies such as artificial intelligence to contribute to the innovation base on intensive products in scientific and technological knowledge.

2. Materials and methods

2.1 Materials

All the compiled dataset containing the SWRC points and the corresponding characterization properties was filtered from the Environmental Geotechnics Laboratory (GEOAMB) (UFBA, 2022) of the Federal University of Bahia (UFBA). Table 1 presents variables’ statistical properties. Most of the soil samples are sandy soils, with low percentages of gravel. There are some samples with high clay and silt content. The mean plasticity index is near 13% and only few soils present plasticity above 21%. The samples have porosity between 0.24 and 0.69. The suction covers from 0 up to 4.10^4 kPa. Figure 1 illustrates all the SWRC points of the dataset. It can be noted that most points are in the central region of the graph.

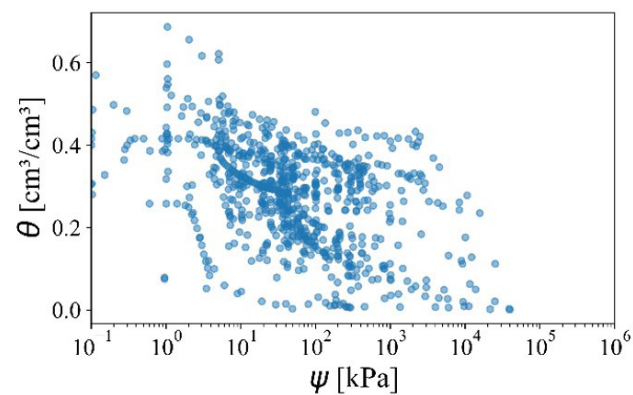


Figure 1. All dataset points in one graph of volumetric water content against suction.

Table 1. Statistical description of the selected 794 samples.

Variable	Mean	Standard deviation	Minimum	25% ^a	50% ^a	75% ^a	Maximum
Sand (%)	52.22	27.48	1.00	32.00	55.00	73.00	100.00
Clay (%)	26.97	23.00	0.00	8.25	22.00	45.17	89.00
Silt (%)	21.14	16.55	0.00	9.00	19.00	29.30	85.00
Gravel (%)	0.62	1.24	0.00	0.00	0.00	0.08	5.00
Plasticity index (%)	12.78	12.84	0.00	3.00	9.00	21.00	58.00
Porosity	0.48	0.10	0.24	0.42	0.49	0.54	0.69
Volumetric water content, θ (cm ³ /cm ³)	0.28	0.12	0.00	0.20	0.30	0.37	0.69
Suction, ψ (kPa)	617.29	2856.69	0.00	10.74	42.17	199.35	39605.60

^aQuantiles.

The dataset doesn't have all features available. In some experiments, the granulometry, or the plasticity index weren't measured. Within the entire database of the 794 points, there were missing 11 points without percentages of sand, and silt, 37 points without plasticity index, filling with the average value of the attribute was done to solve this. More information about the dataset is available in: <https://geofluxo.com/geoapps/swrc-ai/report/>

2.2 Methods

The potential of machine learning to predict the soil water retention curve was investigated by developing estimators in Python using the scikit-learn library (Pedregosa et al., 2011). Other tools used to subsidize were: pandas (McKinney, 2010), NumPy (Oliphant, 2006), matplotlib (Hunter, 2007), jupyter notebook (Kluyver et al., 2016), and anaconda navigator (Anaconda, 2016).

An overview of the method is shown in Figure 2, adapted from Scikit-learn (2021), it is a typical cross-validation workflow in model training. To develop and choose the machine learning model, the training was done on the training set, after which the evaluation was done on the cross-validation set. When the experiments seem to be successful, the final evaluation was done on the test set (Géron, 2019).

To avoid overfitting, a common practice when performing a supervised machine learning experiment is to hold out part of the available data as a test set. Here, 20% of the data (159 points) were used (Géron, 2019; Scikit-learn, 2021). The data were randomly divided into training and test set using a stratified shuffle split because most suction data are between 10^2 and 10^3 kPa. This creates divisions that preserve the same percentage in each interval of suctions defined in

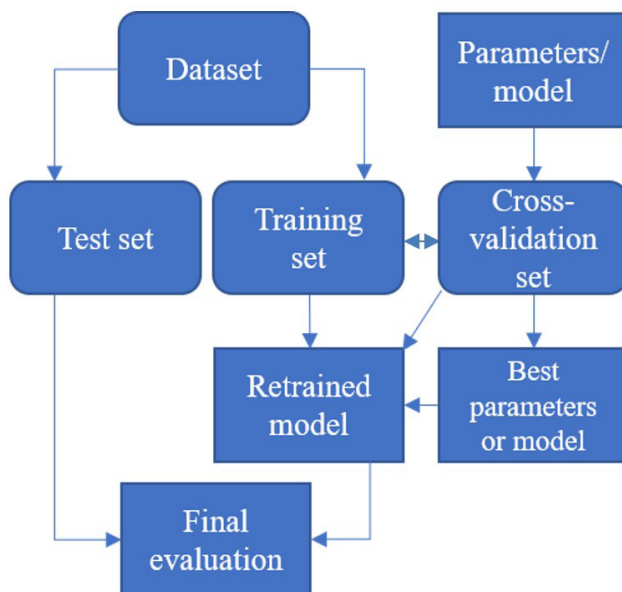


Figure 2. Flowchart of a typical cross-validation workflow.

Figure 3 and allows to assess generalization performance across the entire suction range.

It's necessary for most machine learning estimators that the data are scaled, and the missing ones are filled. They can misbehave if the individual features do not more or less resemble standard normally distributed data (Géron, 2019). So, the data were scaled by standardization, removing the mean and scaling to unit variance. Only 1.4% of the sand and silt percentages data and 4.6% of plasticity index data were missing, as it was a small amount, this was adjusted by filling in with the mean value of each feature.

When evaluating different estimators or different hyperparameter settings of estimators, there is still a risk of overfitting on the test set because the estimators can be chosen by the test performance or the parameters can be tweaked until the model performs optimally (Géron, 2019). This way, knowledge about the test can be part of the model, and evaluation metrics no longer report on generalization performance (Scikit-learn, 2021). To solve this problem, another part of the dataset was held out. This part is the cross-validation set, and the evaluation was done using the 5-fold cross-validation (Breiman & Spector, 1992). The hyperparameter space was searched to achieve the best cross-validation score, using the grid search cross-validation to exhaustively consider all parameter combinations provided and select the best combination.

As an example, some hyperparameters of a decision tree are: *min_samples_split*, minimum number of samples a node must have before it can be split; *min_samples_leaf*, minimum number of samples a leaf node must have; *max_features*, maximum number of features that are evaluated for splitting at each node. The *n_estimators*, *random_state*, *ccp_alpha* hyperparameters controls the number of decision trees ensemble, the randomness, and the pruning of the trees, respectively. Increasing *min_samples_split*, or *min_samples_leaf* or *ccp_alpha* hyperparameters or reducing

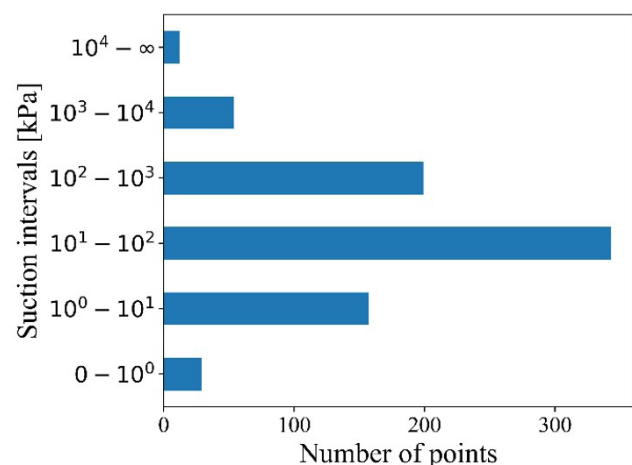


Figure 3. Suction intervals of all the datasets.

max_features hyperparameter will regularize the model (Scikit-learn, 2021; Géron, 2019).

The training was divided into two phases. Phase 1 was characterized by fitting various types of machine learning estimators, such as logistic regression, multilayer-perceptron with adam and different activation functions: ReLU, hyperbolic tangent, sigmoid and identity (Kingma & Ba, 2015), support vector machine with different kernels: linear, polynomial, radial basis function and sigmoid (Smola & Schölkopf, 2004), k-nearest-neighbors, decision tree, random forest (Kam, 1995), and extremely randomized trees (extra trees) (Geurts et al., 2006), with default settings of scikit-learn. More about each model can be seen in Géron's book (Géron, 2019) and Scikit-Learn user guide (Scikit-learn, 2021). In phase 2, the best algorithm obtained in phase 1 was selected to fine-tune the hyperparameters of this model with grid search cross-validation tool. Algorithm selection was based on the root mean squared error (*RMSE*) and the coefficient of determination R^2 measured in the evaluation of the model in the 5-fold cross-validation.

Toward facilitating output results application, predicted SWRC points were fitted using three different models. Cavalcante & Zornberg (2017) proposed a model considering one fitting parameter for the soil water retention curve. The study solved Richard's equation – which governs the unsaturated flow through porous media by a rigorous approach – analytically for a one-dimensional flow. The function was deduced as:

$$\theta(\psi) = \theta_r + (\theta_s - \theta_r) e^{(-\delta\psi)} \quad (1)$$

where δ = fitting hydraulic parameter [$M^{-1}LT^2$]; ψ = soil suction [$ML^{-1}T^2$]; θ_r = residual volumetric water content [L^3L^{-3}]; and θ_s = saturated volumetric water content [L^3L^{-3}].

van Genuchten's (1980) proposed a model considering three fitting parameters, and it is described as:

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{\left(1 + (a_{vg}\psi)^{n_{vg}}\right)^{m_{vg}}} \quad (2)$$

where a_{vg} = fitting parameter [$M^{-1}LT^2$]; n , and m are dimensionless fitting parameters [-].

For bimodal soils, Costa & Cavalcante (2021) proposed a model based on the linear superposition principle and on the Cavalcante's & Zornberg's (2017) model, described as:

$$\theta(\psi) = \theta_r + (\theta_s - \theta_r) \left[\lambda \psi e^{(-\delta_1\psi)} + (1 - \lambda) e^{(-\delta_2\psi)} \right] \quad (3)$$

where δ_1 and δ_2 = fitting hydraulic parameter corresponding to the microporous and macroporous regions, respectively [$M^{-1}LT^2$]; and λ = weight factor [-].

Costa & Cavalcante (2020) demonstrated that the δ parameter is inversely proportional to the air entry pressure and is given by:

$$\psi_{air} = \frac{\exp(1 - \exp(1))}{\delta} \quad (4)$$

where ψ_{air} is the air entry value [$ML^{-1}T^2$].

3. Analysis and results

The results from phase 1, characterized by the running of a diversity types of machine learning algorithms with scikit-learn default settings, are presented in Table 2. The extra trees and random forest were the best models, achieving cross validation R^2 greater than 0.85 and cross-validation *RMSE* less than 0.05. The decision tree algorithm has R^2

Table 2. Phase 1 - Evaluation of different machine learning models with scikit-learn default hyperparameters.

Model	Training <i>RMSE</i>	Cross-validation <i>RMSE</i>	Training R^2	Cross-validation R^2
Extra trees	0.001	0.040	0.99	0.90
Random forest	0.017	0.046	0.98	0.86
Decision tree	0.001	0.056	0.99	0.79
KNN	0.049	0.068	0.84	0.70
SVM (rbf)	0.081	0.088	0.57	0.49
MLP (relu)	0.087	0.092	0.50	0.44
MLP (logistic)	0.098	0.100	0.37	0.35
LR	0.099	0.101	0.36	0.34
MLP (identity)	0.100	0.101	0.35	0.33
MLP (tanh)	0.097	0.101	0.38	0.33
SVM (linear)	0.100	0.102	0.35	0.32
Dummy	0.124	0.124	0.00	0.00
SVM (poly)	0.090	0.183	0.47	-2.50
SVM (sigmoid)	7.328	7.143	-3516.33	-3383.65

Notes: Extra trees is the extremely randomized trees, KNN is the K-Nearest-Neighbors algorithm, SVM is the Support Vector Machine algorithm and in parenthesis is the kernel used, MLP is the Multi-Layer Perceptron algorithm and in parenthesis is the activation function used.

of 0.99 in training, but it presented an inferior performance in generalization phase. K-nearest neighbors (KNN) is the fourth best algorithm analyzed with 0.70 cross-validation R^2 .

Support vector machine (SVM), multi-layer perceptron (MLP) and logistic regression (LR) didn't perform satisfactorily with default settings. The dummy algorithm is a simple regressor that just uses the mean value of the volumetric water content of the training set, it was used to compare the $RMSE$ with the other estimators. Using extra trees algorithm, the cross-validation $RMSE$ was improved in 0.084 compared to that obtained with the dummy algorithm. The SVM with sigmoid kernel and scikit-learn default settings couldn't find a pattern in the training dataset.

The extra trees algorithm was the best model analyzed in phase 1, so it was selected to phase 2. This algorithm is an ensemble of decision trees with all the hyperparameters to control how trees are grown, plus all the parameters to control the ensemble itself. A decision tree is formed by roots nodes and leaf nodes. The root node is the question to make a decision and the leaf node is the answer after all questions (Géron, 2019). Figure 4 presents an example of decision tree, where the root nodes set thresholds towards the leaf node with the corresponding value of volumetric water content. The function used to measure the quality of the split was the mean squared error (mse). In Figure 5, the regression of this tree is shown. The extra trees algorithm fits a number of randomized decision trees on various subsamples of the dataset using random thresholds for each feature rather than searching for the best possible thresholds, and uses averaging to join the predictions from each decision tree (Géron, 2019).

The extra trees algorithm was refined by varying the hyperparameters of the forest and trees to generate a model with higher generalization performance. To find this model, the grid search cross-validation was evaluated varying the $n_estimators$ from 8 to 40, $random_state$ from 0 to 30 in multiples of 2, $min_samples_split$ from 2 to 5, $min_samples_leaf$ from 1 to 4, $max_features$ from 1 to 6, and ccp_alpha with 0, 0.0001 and 0.0002.

In total 152064 combinations were evaluated, and the hyperparameters of the best estimator were: $n_estimators = 24$,

$random_state = 10$, $min_samples_leaf = 1$, $min_samples_split = 2$, $max_features = 6$, $ccp_alpha = 0$. Figure 6 illustrates the learning curve of the best extra trees model, which tells how the performance of this estimator varies according to the amount of data. The R^2 in the training set reaches the top with $R^2 = 0.99$, and the R^2 in validation increases very fast with data reaching $R^2 = 0.90$. It displays that the amount of data leverages the generalization capacity of the algorithm, which could reach up until $R^2 = 0.99$.

Figure 7 shows the graph of the estimator's evaluation in the test set. High scattering occurs at θ ranges with few amounts of data. The coefficient of determination in the test set is $R^2 = 0.90$, and the θ error in a 95% confidence interval is between $0.029\text{ cm}^3/\text{cm}^3$ to $0.043\text{ cm}^3/\text{cm}^3$. This order of magnitude can negatively influence the results depending on the soil type, especially in low porosity soils, where a volumetric water content variation of this magnitude can be significant in the saturated zone of the curve. However, the algorithm will decrease this error as new data is computed.

An interesting aspect is that the decision tree-based estimators (such as extra trees) provide the relative importance of each parameter for the prediction of volumetric water content by the model. Feature importance is a weighted average of how much the trees nodes that use that feature reduce Gini impurity, then the results are scaled to adds up to 1 (Géron, 2019). Table 3 presents this result, indicating that suction is the most significant feature, followed by porosity, and percentage of sand.

Suction is the most relevant variable because it makes the volumetric water content vary from dry to saturated and this makes it out of scale. The other variables have similar importance to each other, despite the percentage of gravel and silt that have feature importance close to zero, probably because the granulometry are linearly dependent on the fourth-dimensional hyperplane where the sum of all the percentages results 100%.

Figure 8 illustrates some predictions of the model for a (a) sandy, (b) silty, and (c) clayey soil. Visually, there is a good adherence between predicted and experimental results. The model is able to predict the SWRC with the

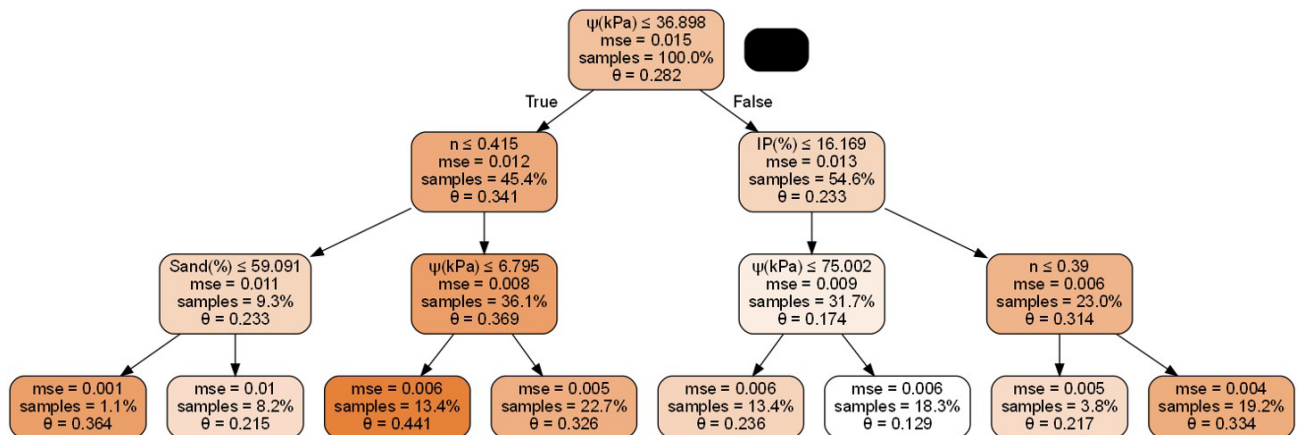


Figure 4. A decision tree example.

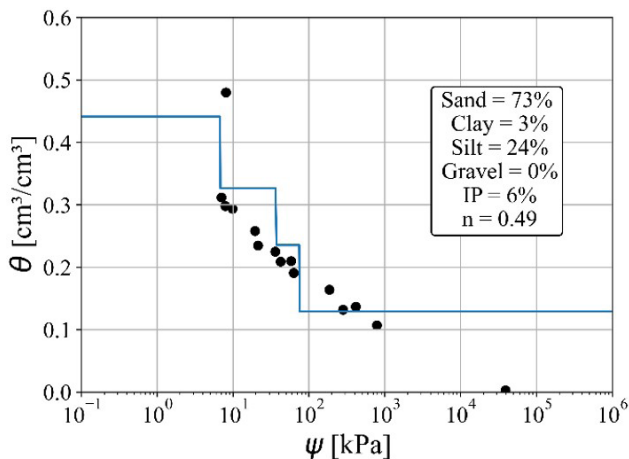


Figure 5. A decision tree regression example.

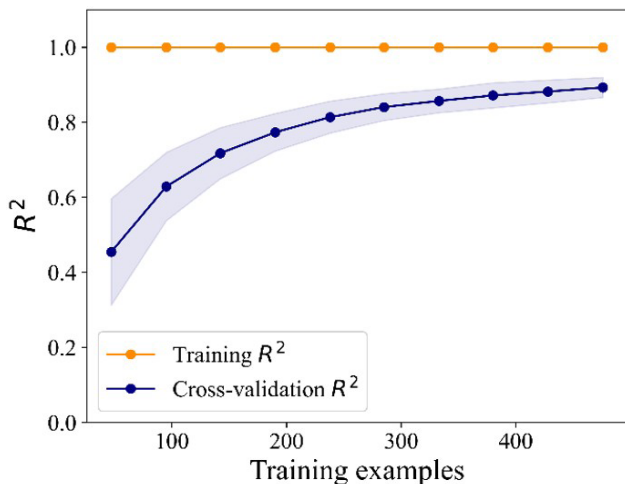


Figure 6. Learning curve of extremely randomized trees.

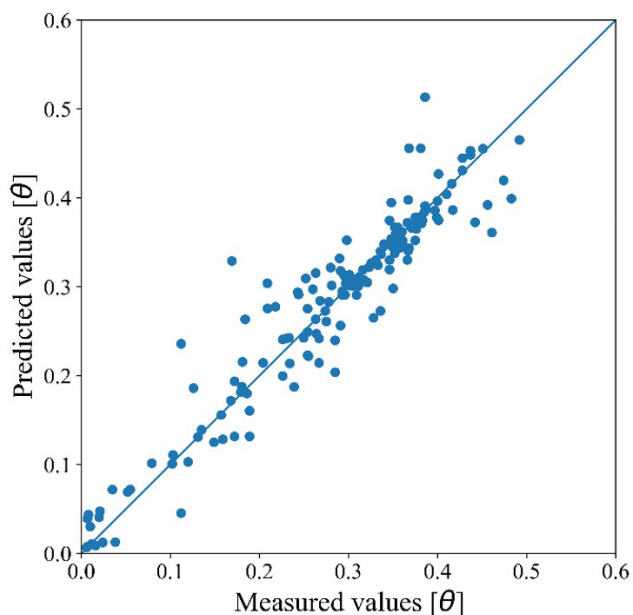


Figure 7. Evaluation in the test set of extremely randomized trees.

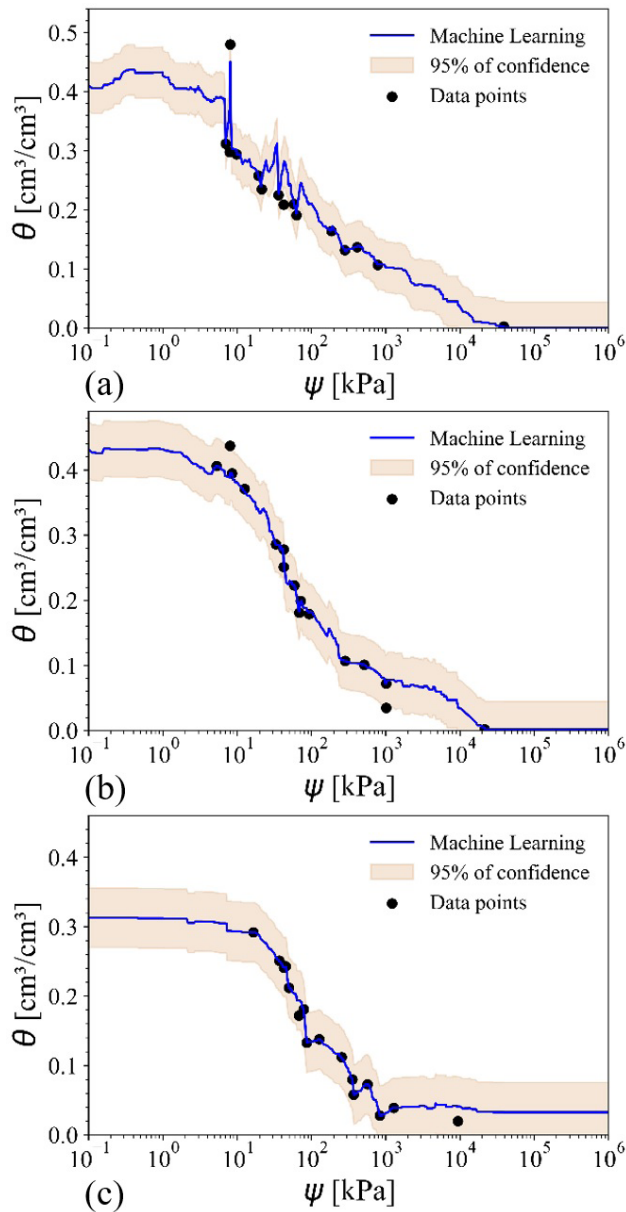


Figure 8. Prediction of the soil water retention curve with extremely randomized trees machine learning algorithm: (a) for a sandy soil with 73% of sand; (b) for a silty soil with 85% of silt; (c) for a clayey soil with 89% of clay.

Table 3. Feature importance for predicting the soil water retention with extremely randomized trees.

Feature	Feature importance
Suction ψ	0.427
Porosity n	0.155
% Sand	0.127
% Clay	0.103
Plasticity index PI	0.085
% Gravel	0.058
% Silt	0.045

three well-defined sections, presenting the saturated zone, the desaturation zone and the residual zone. Some points are far from the machine learning predictions because part of the data was used for testing the model. In Figure 8a, there are some valleys and picks, and probably the first point is an outlier. In some areas, increasing suction causes an increase in volumetric water content, which is not physically defensible. This occurs because the algorithm tries to better fit the data of different types of soils and due to measurement errors.

To overcome these deficiencies and obtain a continuous and smooth curve, a model of soil water retention can be

fitted to the predictions of the machine learning model. The saturated volumetric water content parameter was set equal to the porosity. The graphics on the left of the Figure 9 show the adjustment of Cavalcante's & Zornberg's (2017) function to machine learning predictions, and on the right are the adjustment of van Genuchten's (1980) function. Cavalcante's & Zornberg's (2017) function fits well the predicted silty soil presented, but got lower coefficient of determination R^2 for the sandy and clayey soils. van Genuchten's (1980) model present a visually pleasing performance to fit the sandy and silty soils analyzed; machine learning and the curves almost

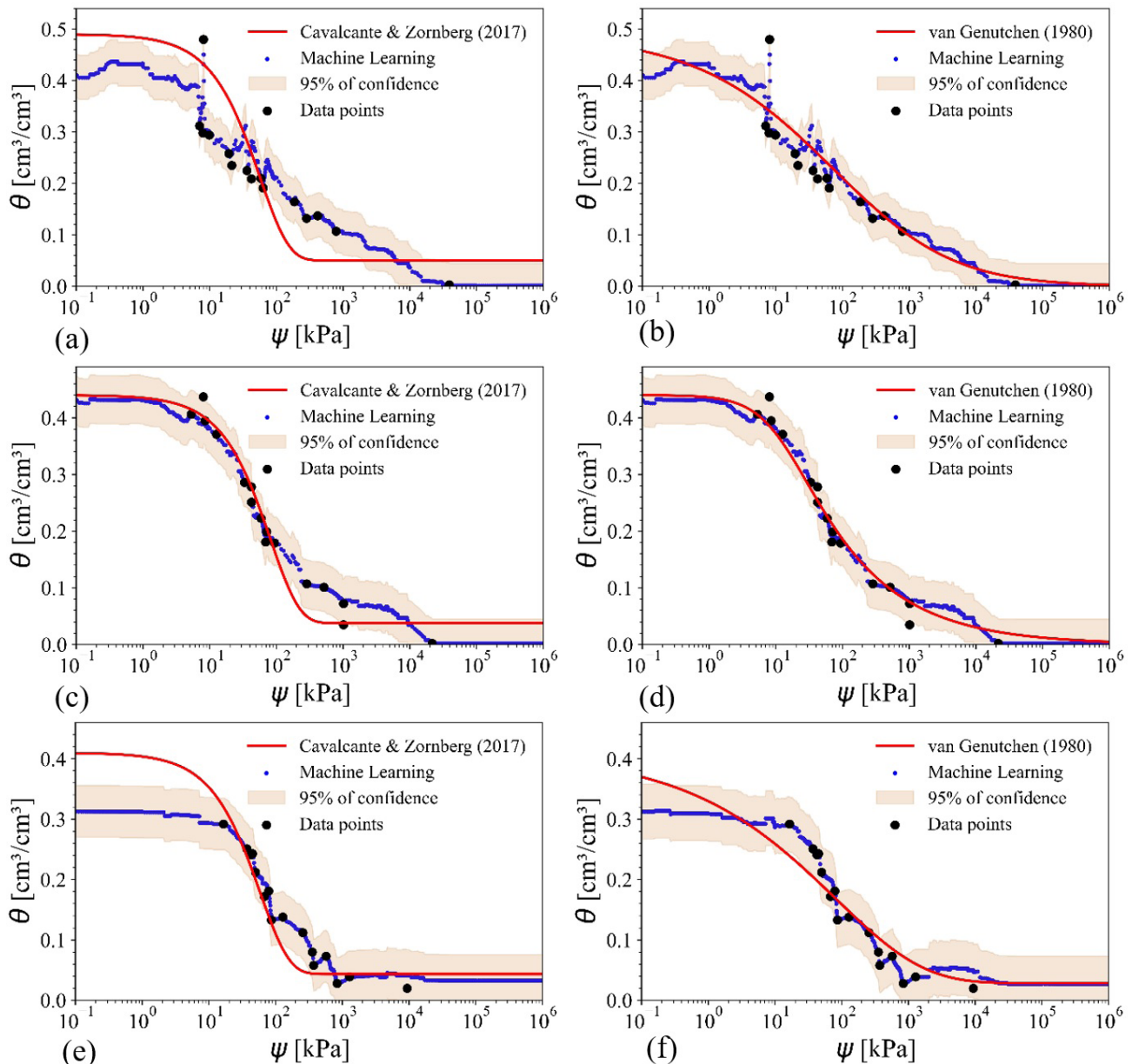


Figure 9. Prediction of the soil water retention curve with Cavalcante & Zornberg (2017) and with van Genuchten (1980) fitting extremely randomized trees prediction with saturated volumetric water content equal to porosity: (a) and (b) experimental points from dataset for a sandy soil with 73% of sand; (c) and (d) experimental points from dataset for a silty soil with 85% of silt; (e) and (f) experimental points from dataset for a clayey soil with 89% of clay.

Table 4. Cavalcante's & Zornberg's (2017) parameters and coefficient of determination of fitting to the machine learning prediction data.

Parameters	Sandy soil	Silty soil	Clayey soil
θ_s (cm ³ /cm ³)	0.49	0.44	0.41
θ_r (cm ³ /cm ³)	0.05	0.04	0.04
δ (kPa ⁻¹)	0.0176	0.0122	0.0170
R^2	0.86	0.97	0.84

Table 5. van Genuchten's (1980) parameters and coefficient of determination of fitting to the machine learning prediction data.

Parameters	Sandy soil	Silty soil	Clayey soil
θ_s (cm ³ /cm ³)	0.49	0.44	0.41
θ_r (cm ³ /cm ³)	0.00	0.00	0.03
a_{vg} (kPa ⁻¹)	0.005	0.084	1.1×10^{-7}
n_{vg} (-)	0.401	1.377	0.349
m_{vg} (-)	1.534	0.289	54.245
R^2	0.99	0.99	0.96

overlap in all regions. The porosity of the analyzed clayey soil is 0.41 and probably has some measurement error.

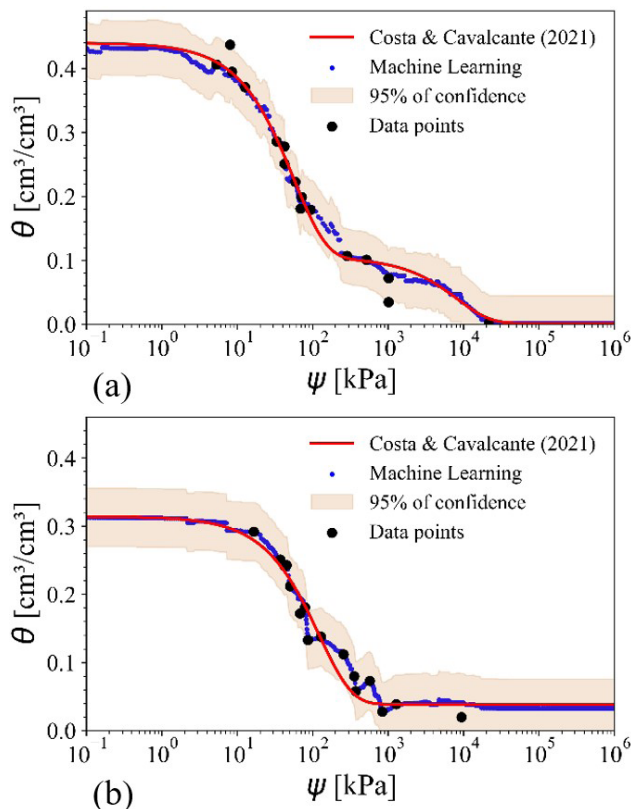
Table 4 and Table 5 presents, respectively, the adjusted parameters and coefficient of determination R^2 for Cavalcante's & Zornberg's (2017) function and van Genuchten's (1980) function. The δ parameter is inversely proportional to the air entry pressure as demonstrated by Costa & Cavalcante (2020). So, the sandy and clayey soil fitted with Cavalcante's & Zornberg's (2017) function have similar air entry pressure and the silty soil has a higher air entry pressure. The clayey soil fitted with van Genuchten's (1980) function resulted in unusual parameters of a_{vg} and m , this was due to the points predicted by the machine learning model in the saturated zone being far from measured porosity.

For bimodal soils, Costa's & Cavalcante's (2021) model can be used to fit the predictions. Figure 10 shows the fit of the silty and clayey soil with Costa & Cavalcante (2021) function. Comparing with Figure 9c, Costa's & Cavalcante's (2021) function fits better to the silty soil analyzed than Cavalcante's & Zornberg's (2017) function. Table 6 shows the adjusted parameters and coefficient of determination R^2 for Costa's & Cavalcante's (2021) function. The silty soil has microporous that results in an air entry pressure about 1800 kPa, and macroporous with an air entry pressure of 10 kPa. The clayey soil analyzed resulted 22 kPa and 10^{-4} kPa for microporous and macroporous, respectively. This low value of air entry for macroporous was due to the function trying to fit the machine learning data and the saturated volumetric water content of 0.41. The use of bimodal SWRC models is necessary for modeling the entire SWRC of soils with a bimodal pore size distribution, common in many Brazilian tropical formations (Kühn et al., 2021; Futai & Almeida, 2005; Cordão-Neto et al., 2018; Miguel & Bonder, 2012).

Although Cavalcante's & Zornberg's (2017) function didn't model as well as Costa's & Cavalcante's (2021)

Table 6. Costa's & Cavalcante's (2021) parameters and coefficient of determination of fitting to the machine learning prediction data.

Parameters	Silty soil	Clayey soil
θ_s (cm ³ /cm ³)	0.44	0.41
θ_r (cm ³ /cm ³)	0.00	0.04
δ_1 (kPa ⁻¹)	0.0001	0.0082
δ_2 (kPa ⁻¹)	0.0178	1758.0848
λ	0.235	0.742
R^2	0.99	0.99


Figure 10. Prediction of the soil water retention curve with Costa & Cavalcante (2021) fitting extremely randomized trees prediction with saturated volumetric water content equal to porosity: (a) experimental points from dataset for a silty soil with 85% of silt; (b) experimental points from dataset for a clayey soil with 89% of clay.

and van Genuchten's (1980) functions, its advantages are: analytical solutions for transient unsaturated flow problems and fewer adjusted parameters.

4. Conclusion

In this paper, the performance of machine learning models for predicting the soil water retention curve was evaluated using a dataset with 794 measured water retention and suctions points from 51 different soils with a wide range of soil properties. Several models have been trained, such as linear regression, logistic regression, multi-layer

perceptron (with identity, sigmoid, hyperbolic tangent activation functions), support vector machine (with radial basis function, linear and sigmoid kernels), k-nearest-neighbors, decision tree, random forest, and extremely randomized trees (extra trees). Extra trees regressor was the best model with scikit-learn default settings, it was selected by measurement of the performance on the cross-validation dataset referring to root mean squared error (*RMSE*) and coefficient of determination R^2 .

This model was then fine-tuned varying its training hyperparameters. For the training set, $RMSE = 0.001$ and $R^2 = 0.99$. For the cross-validation set, $RMSE = 0.039$ and $R^2 = 0.90$. And for the test set, $RMSE = 0.037$ and $R^2 = 0.90$. The most important features, in decreasing order, for prediction were: suction, porosity, percentage of sand, percentage of clay, plasticity index, percentage of gravel, and percentage of silt.

Alternatives to obtain a continuous and smooth curve from the machine learning model were presented by fitting soil water retention functions. Cavalcante's & Zornberg's (2017) function reached R^2 between 0.84 and 0.97 and has the advantage of having analytical solutions, and fewer parameters. van Genuchten's (1980) function reached R^2 between 0.96 and 0.99. Costa's & Cavalcante's (2021) function reached $R^2 = 0.99$ in the silty and clayey soils. Its advantages are being able to fit a bimodal soil model to the predictions of the machine learning estimator, and according to Costa & Cavalcante (2021), it's efficient in representing bimodal soils, and mathematically and physically consistent.

In the preliminary stage of design projects, where not much data is readily available, the model developed can be used to predict the engineering behavior of unsaturated soils. It may be utilized to guide the geotechnical engineers throughout the preliminary analyses and design procedures.

It's important to highlight that machine learning models can always be updated by presenting new training soil samples as new data with measured suction, volumetric water content, and corresponding characterization parameters become available.

This is a work in progress and the quality of predictions will be better the greater the collaboration. Instructions on how to collaborate with this project are provided in the Appendix A.

Acknowledgements

This study was financed in part by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Finance Code 001. The authors also acknowledge the support of the National Council for Scientific and Technological Development (CNPq Grant Nos. 435962/2018-3 and 305484/2020-6), the Foundation for Research Support of the Federal District (FAPDF) (Projects 0193.002014/2017-68 and 0193.001563/2017), the National Electric Energy

Agency (ANEEL) and its R&D partners Neoenergia/CEB Distribuição S.A. (Grant number PD-05160-1904/2019, contract CEBD782/2019), and the University of Brasília. The authors also acknowledge the support of GEOAMB for making the data available on its website.

Declaration of interest

The authors have no conflicts of interest to declare. All co-authors have observed and affirmed the contents of the paper and there is no financial interest to report.

Authors' contributions

Enzo Aldo Cunha Albuquerque: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing - original draft. Lucas Parreira de Faria Borges: software, visualization, writing - review and editing. André Luís Brasil Cavalcante: conceptualization, project administration, supervision, writing - review and editing. Sandro Lemos Machado: conceptualization, resources, supervision, writing - review and editing.

List of symbols

a_{vg}	Fitting parameter of van Genuchten's (1980) function [$M^{-1}LT^2$]
ccp_alpha	Complexity parameter used for minimal cost-complexity pruning
m_{vg}	Fitting parameter of van Genuchten's (1980) function [-]
$max_features$	Number of features to consider when looking for the best split
$min_samples_leaf$	Minimum number of samples required to be a leaf node
$min_samples_split$	Minimum number of samples required to split an internal node
mse	Mean squared error
n	Porosity [L^3L^{-3}]
n_{vg}	Fitting parameter of van Genuchten's (1980) function [-]
$n_estimators$	Number of trees in the forest
PI	Plasticity index
R^2	Coefficient of determination
$RMSE$	Root mean squared error
$random_state$	Random state of parameters initialization
δ	Fitting parameter of Cavalcante's & Zornberg's (2017) function [$M^{-1}LT^2$]
θ	Volumetric water content [L^3L^{-3}]
θ_s	Volumetric water content at saturation (L^3L^{-3})
θ_r	Residual volumetric water content [L^3L^{-3}]
ψ	Soil suction [$ML^{-1}T^{-2}$]
ψ_{air}	Air entry pressure [$ML^{-1}T^{-2}$]

References

- Achieng, K.O. (2019). Modelling of soil moisture retention curve using machine learning techniques: artificial and deep neural networks vs support vector regression models. *Computers & Geosciences*, 133, 104320. <http://dx.doi.org/10.1016/j.cageo.2019.104320>.
- Anaconda. (2016). *Anaconda software (No. 2-2.4.0)*. Austin.
- Arya, L.M., & Paris, J.F. (1981). A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. *Soil Science Society of America Journal*, 45(6), 1023-1030. <http://dx.doi.org/10.2136/sssaj1981.03615995004500060004x>.
- Belcher, W., Camp, T., & Krzhizhanovskaya, V. (2015). Detecting erosion events in earth dam and levee passive seismic data with clustering. In *2015 IEEE 14th International Conference on Machine Learning and Applications* (pp. 903-910), Miami, FL, USA. <https://doi.org/10.1109/ICMLA.2015.9>.
- Brasil. Ministério da Ciência, Tecnologia, Inovações e Comunicações – MCTIC. (March 24, 2020). Portaria nº 1.122, de 19 de março de 2020. *Diário Oficial [da] República Federativa do Brasil*.
- Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression: the X-random case. *International Statistical Review*, 60(3), 291. <http://dx.doi.org/10.2307/1403680>.
- Carvalho, L.O., & Ribeiro, D.B. (2019). Soil classification system from cone penetration test data applying distance-based machine learning algorithms. *Soils and Rocks*, 42(2), 167-178. <http://dx.doi.org/10.28927/SR.422167>.
- Carvalho, L.O., & Ribeiro, D.B. (2020). Application of kernel k-means and kernel x-means clustering to obtain soil classes from cone penetration test data. *Soils and Rocks*, 43(4), 607-618. <http://dx.doi.org/10.28927/SR.434607>.
- Cavalcante, A.L.B., & Zornberg, J.G. (2017). Efficient approach to solving transient unsaturated flow problems. I: analytical solutions. *International Journal of Geomechanics*, 17(7), 04017013. [http://dx.doi.org/10.1061/\(ASCE\)GM.1943-5622.0000875](http://dx.doi.org/10.1061/(ASCE)GM.1943-5622.0000875).
- Ching, J., & Phoon, K.-K. (2019). Constructing site-specific multivariate probability distribution model using bayesian machine learning. *Journal of Engineering Mechanics*, 145(1), 04018126. [http://dx.doi.org/10.1061/\(ASCE\)EM.1943-7889.0001537](http://dx.doi.org/10.1061/(ASCE)EM.1943-7889.0001537).
- Cordão-Neto, M.P., Hernández, O., Lorenzo Reinaldo, R., Borges, C., & Caicedo, B. (2018). Study of the relationship between hydro-mechanical soil behavior and microstructure of a structured soil. *Earth Sciences Research Journal*, 22(2), 91-101. <http://dx.doi.org/10.15446/esrj.v22n2.65640>.
- Costa, M.B.A. (2017). *Modelagem numérica do fluxo transiente em meio poroso não saturado sob a ação de centrifugação* [Master's dissertation, University of Brasília]. University of Brasília repository (in Portuguese). Retrieved in January 10, 2021, from <https://repositorio.unb.br/handle/10482/24849>
- Costa, M.B.A., & Cavalcante, A.L.B. (2020). Novel approach to determine soil-water retention surface. *International Journal of Geomechanics*, 20(6), 04020054. [http://dx.doi.org/10.1061/\(ASCE\)GM.1943-5622.0001684](http://dx.doi.org/10.1061/(ASCE)GM.1943-5622.0001684).
- Costa, M.B.A., & Cavalcante, A.L.B. (2021). Bimodal soil-water retention curve and k-function model using linear superposition. *International Journal of Geomechanics*, 21(7), 04021116. [http://dx.doi.org/10.1061/\(ASCE\)GM.1943-5622.0002083](http://dx.doi.org/10.1061/(ASCE)GM.1943-5622.0002083).
- Fisher, W.D., Camp, T.K., & Krzhizhanovskaya, V. (2016). Crack detection in earth dam and levee passive seismic data using support vector machines. *Procedia Computer Science*, 80, 577-586. <http://dx.doi.org/10.1016/j.procs.2016.05.339>.
- Fisher, W.D., Camp, T.K., & Krzhizhanovskaya, V. (2017). Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection. *Journal of Computational Science*, 20, 143-153. <http://dx.doi.org/10.1016/j.jocs.2016.11.016>.
- Fredlund, M.D., Wilson, G.W., & Fredlung, D.G. (2002). Use of the grain-size distribution for estimation of the soil-water characteristic curve. *Canadian Geotechnical Journal*, 39(5), 1103-1117. <https://doi.org/10.1139/t02-049>.
- Futai, M.M., & Almeida, M.S.S. (2005). An experimental investigation of the mechanical behaviour of an unsaturated gneiss residual soil. *Geotechnique*, 55(3), 201-213. <http://dx.doi.org/10.1680/geot.2005.55.3.201>.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and Tensorflow: concepts, tools, and techniques to build intelligent systems*. Beijing: O'Reilly Media.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3-42. <http://dx.doi.org/10.1007/s10994-006-6226-1>.
- Haghverdi, A., Leib, B.G., & Cornelis, W.M. (2015). A simple nearest-neighbor technique to predict the soil water retention curve. *Transactions of the ASABE*, 58(3), 697-705. <http://dx.doi.org/10.13031/trans.58.10990>.
- Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. <http://dx.doi.org/10.1109/MCSE.2007.55>.
- Kam, T.H. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (pp. 278-282), Montreal, QC, Canada.
- Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., & Cornelis, W.M. (2016). Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. *European Journal of Soil Science*, 67(3), 276-284. <http://dx.doi.org/10.1111/ejss.12345>.
- Kingma, D.P., & Ba, J.L. (2015). Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (pp. 1-15), San Diego, CA.

- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks: a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016* (pp. 87-90). Amsterdam: IOS Press. <https://doi.org/10.3233/978-1-61499-649-1-87>.
- Kühn, V.O., Lopes, B.C.F., Caicedo, B., & Cordão-Neto, M.P. (2021). Micro-structural and volumetric behaviour of bimodal artificial soils with aggregates. *Engineering Geology*, 288, 106139.
- Marjanović, M., Kovačević, M., Bajat, B., & Voženilek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology*, 123(3), 225-234. <http://dx.doi.org/10.1016/j.enggeo.2011.09.006>.
- McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56-61), Austin, Texas.
- Miguel, M.G., & Bonder, B.H. (2012). Soil–water characteristic curves obtained for a colluvial and lateritic soil profile considering the macro and micro porosity. *Geotechnical and Geological Engineering*, 30(6), 1405-1420. <http://dx.doi.org/10.1007/s10706-012-9545-y>.
- Oliphant, T. (2006). *A guide to NumPy*. USA: Trelgol Publishing.
- Oliveira Filho, A.G., Totola, L.B., Bicalho, K.V., & Hisatugu, W.H. (2020). Prediction of compression index of soft soils from the Brazilian coast using artificial neural networks and empirical correlations. *Soils and Rocks*, 43(1), 109-121. <https://doi.org/10.28927/sr.431109>.
- Ozelim, L.C., Borges, L.P.F., Cavalcante, A.L.B., Albuquerque, E.A.C., Diniz, M.S., Góis, M.S., Costa, K.R.C.B., Sousa, P.F., Dantas, A.P.N., Jorge, R.M., Moreira, G.R., Barros, M.L., & Aquino, F.R. (2022). Structural health monitoring of dams based on acoustic monitoring, deep neural networks, fuzzy logic and a CUSUM control algorithm. *Sensors*, 22(7), 1-25. <http://dx.doi.org/10.3390/s22072482>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michler, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <http://dx.doi.org/10.1289/EHP4713>.
- Prayogo, D., & Susanto, Y.T.T. (2018). Optimizing the prediction accuracy of friction capacity of driven piles in cohesive soil using a novel self-tuning least squares support vector machine. *Advances in Civil Engineering*, 2018, 6490169. <http://dx.doi.org/10.1155/2018/6490169>.
- Scikit-learn. (2021). *User guide*. Retrieved in January 10, 2021, from https://scikit-learn.org/stable/user_guide.html
- Smola, A.J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222. <http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88>.
- Tien Bui, D., Tuan, T.A., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13(2), 361-378. <http://dx.doi.org/10.1007/s10346-015-0557-6>.
- Universidade Federal da Bahia – UFBA. (2022). Retrieved in January 6, 2022, from <http://www.geoamb.eng.ufba.br/dados/buscaSolos.php>
- van Genuchten, M.T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5), 892-898.
- Vanapalli, S.K., & Catana, M.C. (2005). Estimation of the soil-water characteristic curve of coarse-grained soils using one point measurement and simple properties. In *International Symposium on Advanced Experimental Unsaturated Soil Mechanics*, Trento, Italy.

Appendix A. How to collaborate.

To add new data to the machine learning set, one should sign up at the website (<https://www.geofluxo.com/>) and go to the restricted area, then to applications, Figure A1.

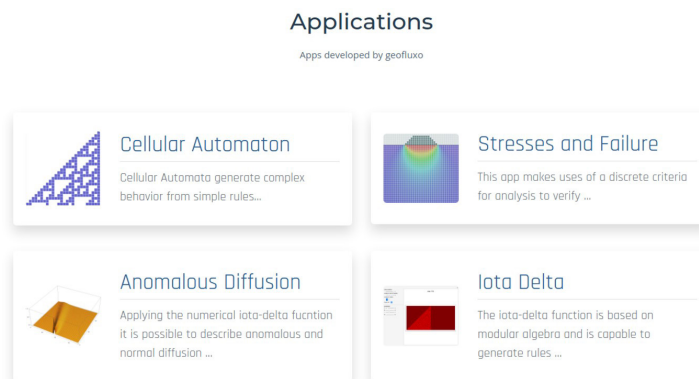


Figure A1. Applications page at <https://geofluxo.com/geoapps/>.

Then, one should select the SWRC AI application. On the application page, (<https://geofluxo.com/geoapps/swrc-ai/>) it is possible to insert the inputs and come up with the outputs values plus the chart with the artificial intelligence fit.

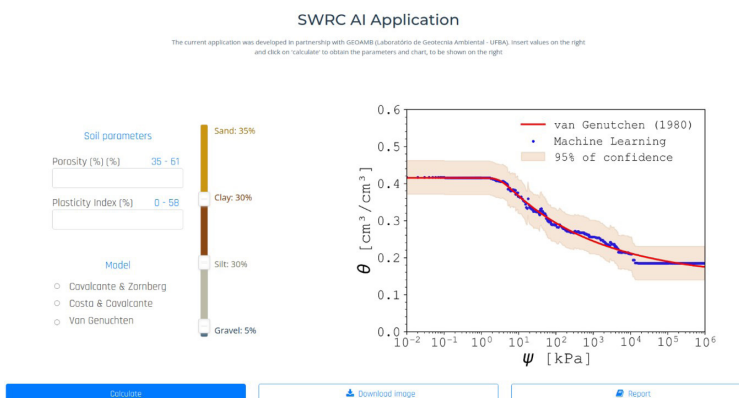


Figure A2. SWRC AI application page.

If the user scrolls down and hit the form button, Figure A3, he will be redirected to a webpage that does not host the application itself, but allows the user to collaborate with new data to the dataset.



Figure A3. SWRC AI form button.

Once the user access the form page at <https://geofluxo.com/geoapps/swrc-ai/form/>, it is possible to write an introduction about the data, the percentages of gravel, sand, clay and silt, porosity, plasticity index, and insert several points of suction and volumetric water content.

SWRC AI Form

Form

This form allows to insert data that describes the soil water retention curve, granulometry and physical indexes to calibrate the models based on Artificial Intelligence. The data may come from laboratory or field experimentation.

Circumstantial description of data

Granulometric Cruve

Gravel (%)	<input type="text" value="0"/>
Sand (%)	<input type="text" value="0"/>
Silt (%)	<input type="text" value="0"/>
Clay (%)	<input type="text" value="0"/>

Physical Indexes

Porosity (%)	<input type="text" value="0"/>
Plasticity Index (%)	<input type="text" value="0"/>

Soil Water retention curve

Moisture (%)	Suction (kPa)
<input type="text" value="0"/>	<input type="text" value="0"/>
<input type="text" value="0"/>	<input type="text" value="0"/>
<input type="text" value="0"/>	<input type="text" value="0"/>
<input type="text" value="0"/>	<input type="text" value="0"/>
<input type="text" value="0"/>	<input type="text" value="0"/>

Figure A4. SWRC AI form.

Once the form is submitted, the admins will accept or deny the new dataset. If accepted, the new dataset will integrate the next machine learning training session.