PANAMGEO CHILE 2024 — 17th PAN-AMERICAN CONFERENCE ON SOIL MECHANICS AND GEOTECHNICAL ENGINEERING — 2nd LATIN-AMERICAN REGIONAL CONFERENCE OF IAEG — LA SERENA • CHILE / NOVEMBER 12 - 16, 2024

Proceedings of the 17th Pan-American Conference on Soil Mechanics and Geotechnical Engineering (XVII PCSMGE), and 2nd Latin-American Regional Conference of the International Association for Engineering Geology and the Environment (IAEG), La Serena Chile, 2024.

Predicting the Soil-Water Characteristic Curve of Tropical Bimodal Soils Using Gradient Boosting

Previsão da Curva Característica Solo-Água de Solos Tropicais Bimodais Utilizando Gradient Boosting

Sávio A. dos Santos Pereira & Gilson de F. N. Gitirana Jr.
School of Civil and Environmental Engineering, Universidade Federal de Goiás, Brazil, savioaparecido1@gmail.com

Thiago A. Mendes & Arlam C. Silva Jr.
Department of Civil Engineering, Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Brazil

Juan F. Rodriguez Rebolledo
Department of Civil and Environmental Engineering, Universidade de Brasília, Brazil

ABSTRACT: The soil-water characteristic curve (SWCC) plays a crucial role in modeling unsaturated soil phenomena. However, obtaining SWCC measurements is often a labor-intensive and costly undertaking. To provide preliminary estimations, various models have been devised to predict SWCC through simpler parameters like the grain-size distribution and the Atterberg limits. Within this context, this study introduces a novel SWCC prediction model for bimodal tropical soils based on data from soils from the central-western region of Brazil. The model employs the gradient-boosting machine-learning technique. The input parameters encompass the percentages of sand, silt, and clay for the soil in the aggregated and disaggregated states, alongside Atterberg limits. As output, the model provides the degree of saturation corresponding to any user-defined suction value. The model's training dataset comprises 696 records, with 174 data points being used for model testing. The model yielded an R^2 value of 0.95 for training data and 0.90 for test data, surpassing the performance of widely employed literature models for SWCC prediction. However, it presented limitations in the predictions of high degrees of saturation.

KEYWORDS: water retention curve; lateritic soils; unsaturated soils; machine learning; artificial intelligence.

1 INTRODUCTION

The soil-water characteristic curve (SWCC) plays a crucial role in flow, stress, and strain analyses of unsaturated soils (Fredlund et al. 2012). The permeability, shear strength, and deformation parameters of unsaturated soils are commonly analyzed using the SWCC (Fredlund et al. 1994, Fredlund et al. 1996, Vanapalli et al. 1996, Fredlund 2000, Zhai et al. 2020, Amadi et al. 2023). In addition, various models for predicting unsaturated soil parameters using the SWCC have been developed.

The SWCC can be obtained through laboratory tests such as filter paper and pressure plate. These tests tend to be costly, time-consuming and are not always accessible in geotechnical laboratories, making them unfeasible for small projects or during preliminary feasibility studies. On the other hand, numerous SWCC prediction models using more accessible parameters, commonly known as pedotransfer functions (PTF), have been developed. The particle size distribution (PSD) has been widely used in SWCC prediction models since the SWCC is directly related to the pore-size distribution, which is affected by the PSD (Arya & Paris 1981, Alves et al. 2020, Silva et al. 2020, Campos-Guereta et al. 2021, Satyanaga et al. 2023, Zhai et al. 2023). Other parameters are also often incorporated into prediction models, such as liquid limit, plasticity index, and bulk density.

SWCC prediction methods can be divided into two categories: theoretical and empirical models. Theoretical models assume that the pore-size distribution is directly related to the soil's water retention capacity. Different forms of modeling are used in theoretical models, such as fractals or pore-scale analyses (Arya & Paris 1981, Aubertin et al. 2003, Likos & Jaafar 2013, Wang et al. 2017, Alves et al. 2020). Theoretical models are, therefore, models whose parameters have a clearly defined physical relationship with the SWCC. Empirical models, on the other hand, use statistical techniques to correlate more accessible geotechnical parameters with the SWCC. These statistical techniques can range from standard non-linear regression models to machine learning models (Minasny et al. 1999, Botula et al 2012, Achieng 2019, Wang et al. 2019, Chai & Khaimook 2020, Li & Vanapalli 2021, Albuquerque et al. 2022).

Machine learning techniques are robust tools that can establish more complex relationships, even when there is little correlation between the input and output parameters. Various machine learning models have been used to predict SWCC, such as artificial neural networks, ANN (Haghverdi et al. 2012, Pham et al. 2019, Rudiyanto et al. 2021, dos Santos Pereira et al. 2023), support vector machine, SVM (Achieng 2019), and gradient boosting, GB (Bakhshi et al. 2023, Pham et al. 2023).

The study of SWCC prediction models using machine learning methods has evolved rapidly, and several models with adequate performance have been presented in the literature. However, there are several gaps in the studies of SWCC prediction models. Few models have been dedicated to predicting the SWCC of special soils, other than those soils from temperate climate regions. Only a few models have been presented based on machine learning for the prediction of multimodal and tropical soils.

Dos Santos Pereira et al. (2023) presented a model using ANNs specifically developed for tropical bimodal soils and using a database of soils from the central-western region of Brazil. That model presented good results, offering predictions with R^2 values of 0.690 on average. This paper presents a new SWCC prediction model that uses the same modeling framework and soil database from dos Santos Pereira et al. (2023). However, this paper replaces ANNs, using the Gradient Boosting machine learning method.

2 MATERIAL AND METHODS

The database used in this paper comes from the work carried out by dos Santos Pereira et al. (2023). A data collection study of tropical bimodal soils in the central-western region of Brazil was carried out based on papers, dissertations, and theses developed at the Universidade Federal de Goiás (UFG) and the Universidade de Brasília (UnB). Table 1 summarizes the data collected in statistical terms.

Figure 1 shows the histogram of the soil suction and degree of saturation variables and Figure 3 shows all 870 experimental points present in the database. More information on the database can be found in the original paper (dos Santos Pereira et al. 2023). The histograms shown in Figure 1 indicate the importance of adopting a natural logarithm for the suction variable. With the logarithm, the data tends to show two bell-shaped distributions, with two peaks, characteristic of bimodal soil.

Table 1. Statistical summary of the database developed by dos Santos Pereira et al. (2023).

Variables	Min	Max	Mean	Std. Dev.	COV (%)
% of gravel	0	35	2	7	232
% of sand CD	0	77	49	20	41
% of silt CD	7	64	26	14	54
% of clay CD	0	91	22	25	113
% of sand SD	31	95	61	14	24
% of silt SD	0	68	31	14	46
% of clay SD	0	37	6	11	188
w_L, %	24	54	39	6	15
PI, %	5	36	14	4	33
ψ, kPa	0.10	248476	6734	18153	270
ln(ψ)	-2.30	12.4	5.6	3.3	59
S, %	2	100	46	26	55

where: CD is with disaggregation; SD is without disaggregation; w_L is the liquid limit; *PI* is the plastic index; ψ is soil suction; and *S* is the degree of saturation.

Table 1 also shows a reduction in the coefficient of variation when the natural log is used. Consistent with what was presented by Gitirana Jr. and Fredlund (2016) and dos Santos Pereira et al. (2023), in this work the suction variable was analyzed on a logarithmic scale. Regarding the degree of saturation data, the values are well distributed over the range of 0 and 1. This is an interesting fact, as the model was trained with data covering a large part of the SWCC, as shown in Figure 2.

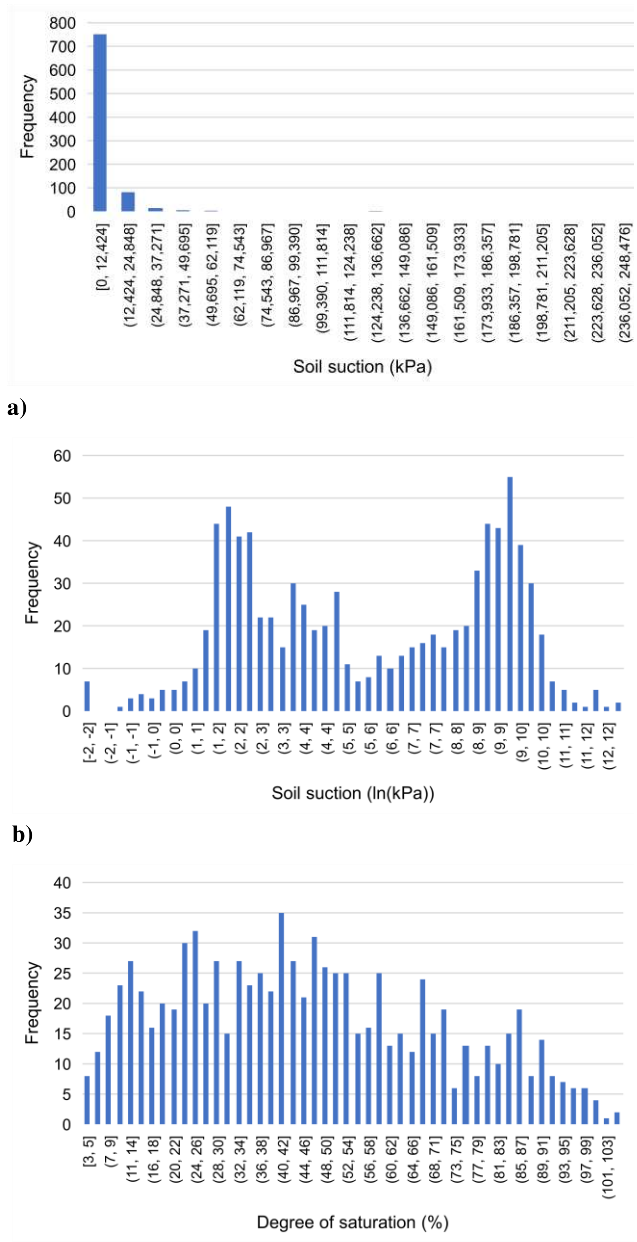

Figure 1. Histogram of: a) soil suction; b) ln(soil suction); and c) degree of saturation.

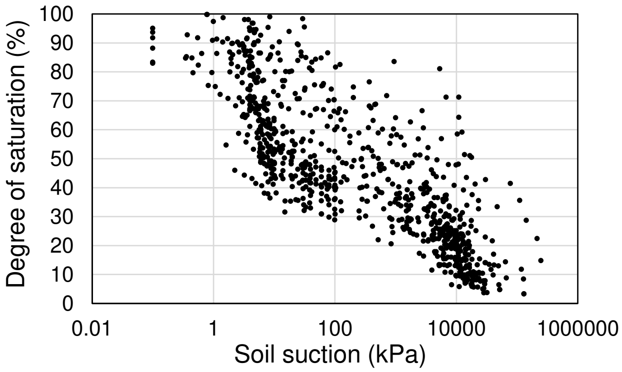

Figure 2. 870 experimental points from the database of dos Santos Pereira et al. (2023).

The input parameters for the model are identical to those used in dos Santos Pereira et al. (2023): percentage of sand, silt, and clay (both in the aggregated and disaggregated states), liquid limit (w_L), and plastic index (PI).

A pseudo-continuous approach was used, as presented by Haghverdi et al. (2012). In this approach, suction is also an input parameter and the selected water content variable is the model output. In the case of the present mode, the degree of saturation was selected as the variable describing the amount of water stored in the soil.

CatBoost, also known as Categorical Gradient Boosting, was adopted as the prediction model. In CatBoost, models are trained sequentially, and each model tries to correct the previous model. CatBoost is a highly predictive model that uses a decision tree structure and was developed by engineers and researchers at Yandex. More details about the model can be found at: https://catboost.ai/.

Table 2 shows the main variables that control the model and the values adopted. All the variables were established by trial and error. The data was partitioned into training and test data. The total number of records is 870, with 80% of the data used as training (696 records) data and 20% as test data (174 records).

3 RESULTS

Figure 3 shows the behavior of the RMSE during model training, indicating that it was a stable phase, despite having shown small gains in prediction. Since no overfitting was identified in the model, as shown later, the number of 800 epochs was considered an optimal value.

Figure 4 shows the relationship between the experimental values and the predicted values, considering the training and test data. For both training and testing, the gradient boosting model was superior to the ANN developed by dos Santos Pereira et al. (2023). While dos Santos Pereira et al. (2023) obtained R^2 values of 0.70 and 0.68 for training and testing, respectively, the gradient boosting model was able to reach R^2 values of 0.95 and 0.89 for training and testing, respectively. The results observed in this paper corroborate the performance increase caused by gradient boosting observed by Bakhshi et al. (2023) and Pham et al. (2023).

To test the quality of the prediction, a soil was selected from outside the database that had similar characteristics to the soils in the training dataset. The work selected was that of Calle (2013), who studied soils from Taguatinga Park Road (EPTG), in the city of Brasília, Brazil. The bimodal SWCC was obtained from undisturbed samples and using the axis translation method, an osmotic cell, a suction plate, and the filter paper technique. Table 3 shows the input parameters of the Calle's soil profile.

Figure 5 shows in black the results obtained experimentally and in red those obtained using gradient boosting. The data points are also represented by the best-fit curve based on the equation proposed by Gitirana Jr. and Fredlund (2004). The curve shown in blue was obtained using the ANN developed by dos Santos Pereira et al. (2023). Other models present in the literature were not employed for comparison since dos Santos Pereira et al. (2023) has presented analyses showing that models that were not developed for tropical bimodal soils offer poor predictions.

Analyzing Figure 5, it can be observed that the SWCC indirect prediction using the newly developed model, based on CatBoost, was able to predict the air-entry values of the macropores but failed to predict the corresponding degrees of saturation. This indicates that the input data available was unable to allow the prediction of the maximum water storage capacity of the soil.

Table 2. CatBoost model input variables.

Iterations	800
Learning rate	0.01
Depth	10
Loss function	RMSE
L2 leaf regularization	0.1
Bootstrap type	Bayesian

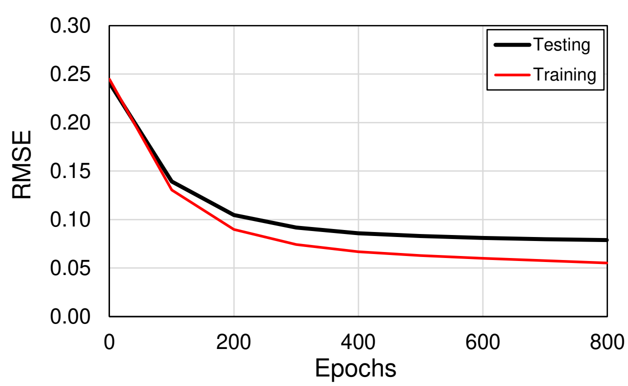

Figure 3. RMSE during CatBoost model training.

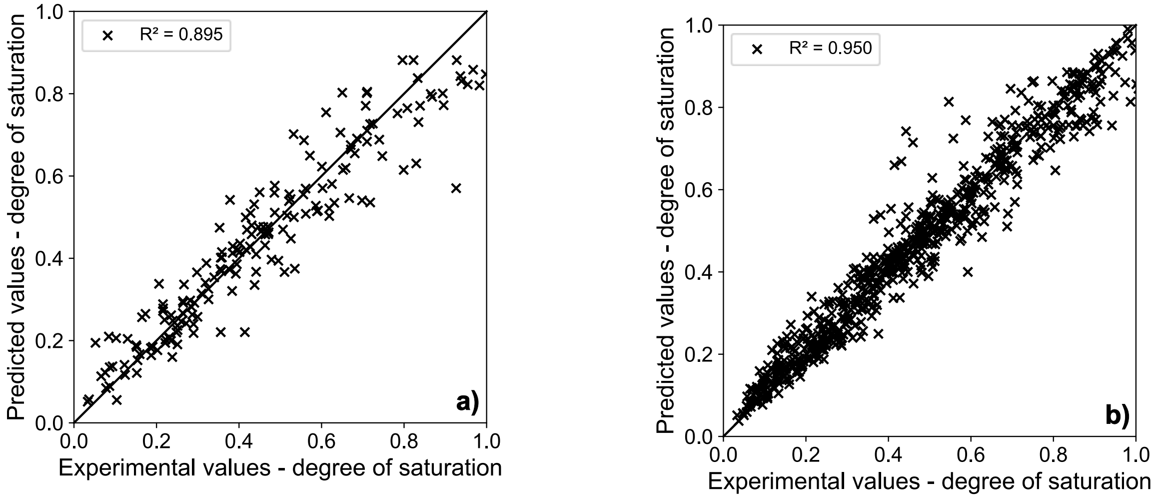

Figure 4. Relationship between experimental and predicted values: a) experimental versus predicted (CatBoost test); b) experimental versus predicted (CatBoost training).

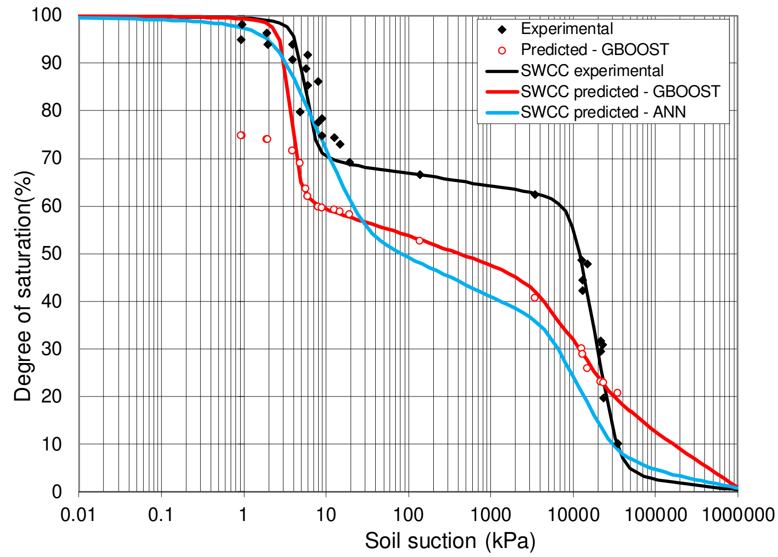

Figure 5. SWCC prediction exercise using the soil studied by Calle (2023).

Table 3. Input parameters of the Calle's soil profile.

with disaggregation (%)			without disaggregation (%)				
sand	silt	clay	sand	silt	clay	w_L (%)	PI (%)
48	16	35	81	12	1	40	12

where w_L is the liquid limit and PI is the plastic index.

Both models (CatBoost and ANN) underestimated the maximum degree of saturation of the micropores. This is shown by degree of saturation values at the air-entry value of micropores of 0.63 that are predicted by the CatBoost and ANN models as 0.46 and 0.39, respectively. The suction value corresponding to the air-entry value, however, was predicted with greater accuracy, with values of 4.7, 3.0, and 2.7 kPa obtained from direct testing, CatBoost, and ANN model, respectively.

The observed limitations of the model may be because the water present in lateritic aggregations of tropical soil, in its natural state, is not quantifiable using the particle size data. In addition, even adding the input data PI and w_L to the gradient boosting and ANN prediction models, the errors are still considerable, that is, around 30% on average (Table 4).

The result obtained by ANN was consistent in the regions up to the first residual zone and in the second residual zone. On the other hand, the gradient boosting model limited the maximum degree of saturation of the SWCC to 75%. The 75% limitation was only overcome by fitting the SWCC to the predicted points, which is an advantage of using a pseudo-continuous approach. For the soil in question, ANN proved to be better in terms of prediction.

Table 4. Comparison between the main curve fit parameters for the SWCC studied by Calle (2023).

	ψ_{b1}	ψ_{res1}	S_{res1}	ψ_{b2}	S_b	ψ_{res2}	S_{res2}
Experimental (Exp)	4.7	6.60	0.69	10,000.00	0.63	35,000.00	0.02
Gradient Boosting (GB)	3.0	4.50	0.60	5,000.00	0.46	15,000.00	0.20
ANN	2.7	25.2	0.50	5,991.08	0.39	25,075.12	0.04
Error % (Exp-GB)	-36	-32	-13	-50	-27	-57	900
Error % (Exp-ANN)	-43	282	-28	-40	-38	-28	100

where ψ_{b1} is the first air-entry suction, ψ_{res1} is the first residual suction; S_{res1} is the first residual degree of saturation; ψ_{b2} is the second air-entry suction; S_b is the air-entry degree of saturation; ψ_{res2} is the second residual suction; S_{res2} is the second residual degree of saturation.

4 CONCLUSIONS

This paper presented a new prediction model for bimodal tropical soils of the central-western region of Brazil, based on gradient boosting. The model presented an R^2 of 0.90 for the test data and 0.95 for the training data. In the prediction exercise using a soil that is not part of the database, but which has similar characteristics to those in the database, the gradient boost model has difficulty in predicting certain features of the SWCC. The model managed to predict the air-entry value, but the results were limited to a degree of saturation of 75%, unlike the 100% obtained using the previously developed ANN. Furthermore, the model was unable to reproduce accurately the SWCC in the first residual zone region. For the soil in question, the ANN model proved to be superior. However, the test was only carried out on one soil, and it should be noted that to be more certain about the limitations and capabilities of the gradient boosting model, it is necessary to apply the model to other soils.

Despite the R^2 of the model presented in this paper being higher than that of the ANN model, its use should be cautious, and whenever possible, tests should be carried out. Given the results obtained from the test data, the model may be superior to the ANN model, but it would need to be tested on different soils. A preliminary conclusion that the previously developed ANN model is superior to the gradient boosting model requires further verification.

The model developed here can be practically applied by using a source file containing the saved model, which can be accessed in Python via the CatBoost library. The practical application of the SWCC prediction model is crucial for feasibility analyses, less robust engineering projects, and short-term decision-making. Currently, using the model requires knowledge of Python, but in the future, an accessible and intuitive platform will be developed to enable any user to easily utilize the models.

For future research, it is necessary to create a larger database for Brazilian tropical bimodal soils to develop more accurate SWCC prediction models using machine learning. Additionally, we recommend exploring different machine-learning algorithms and output forms. Developing new models is essential for producing higher quality and more precise predictions.

5 ACKNOWLEDGEMENTS

The authors are thankful for the financial support provided by FAPEG (Fundação de Amparo à Pesquisa do Estado de Goiás, in English: Goias State Research Support Foundation) and Technological Development of Brazil (CNPq) – Chamada CNPq 25/2021 - Pós-Doutorado Júnior - PDJ 2021.

6 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 and Geosciences* 133. https://doi.org/10.1016/j.cageo.2019.104320

Amadi A.A., Osinubi K.J. and Okoro J.I. 2023. Hydraulic conductivity of unsaturated specimens of lateritic soil-bentonite mixtures. *Geotechnical and Geological Engineering* 1-14. https://doi.org/10.1007/s10706-023-02524-3

Albuquerque E.A.C., Borges L.P.F., Cavalcante A.L.B. and Machado S.B. 2022. Prediction of soil water retention curve based on physical characterization parameters using machine learning. *Soils and Rocks* 45. https://doi.org/10.28927/SR.2022.000222

Alves R.D., Gitirana Jr G.F.N. and Vanapalli S.K. 2020. Advances in the modeling of the soil–water characteristic curve using pore-scale analysis. *Computers and Geotechnics* 127:103766. https://doi.org/10.1016/j.compgeo.2020.103766

Arya L.M. and Paris J.F. 1981. A physicoempirical model to predict soil moisture characteristics from particle-size distribution and bulk density data. *Soil Science Society of America Journal* 45:1023-1030. https://doi.org/10.2136/sssaj1981.03615995004500060004x

Aubertin M., Mbonimpa M., Bussière B. and Chapuis R.P. 2003. A model to predict the water retention curve from basic geotechnical properties. *Canadian Geotechnical Journal*, 40(6): 1104–1122. https://doi.org/10.1139/t03-054

Bakhshi A., Alamdari P., Heidari A. and Mohammadi M. H. 2023. Estimating soil–water characteristic curve (SWCC) using machine learning and soil micro-porosity analysis. *Earth Science Informatics* 16(4), 3839-3860. https://doi.org/10.1007/s12145-023-01131-3

Botula Y.D., Cornelis W.M., Baert G. and van Ranst E. 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). *Agricultural Water Management* 111: 1–10. https://doi.org/10.1016/j.agwat.2012.04.006

Calle I.F.O. 2013. *Comportamento Hidromecânico De Um Solo Tropical Compactado (in portuguese).* Doctoral thesis. Programa de Pós-Graduação em Geotecnia, Departamento de Engenharia Civil e Ambiental, Faculdade de Tecnologia. Universidade de Brasília, Brasília, DF. 143p.

Campos-Guereta I., Dawson A. and Thom N. 2021. An alternative continuous form of Arya and Paris model to predict the soil water retention curve of a soil. *Advances in Water Resources* 154:103968. https://doi.org/10.1016/j.advwatres.2021.103968

Chai J. and Khaimook P. 2020. Prediction of soil-water characteristic curves using basic soil properties. *Transportation Geotechnics* 22. https://doi.org/10.1016/j.trgeo.2019.100295

Fredlund D.G., Xing A. and Huang S. 1994. Predicting the permeability function for unsaturated soils using the soil-water characteristic curve. *Canadian Geotechnical Journal* 31(4): 533-546. https://doi.org/10.1139/t94-062

Fredlund D.G., Xing A., Fredlund M.D. and Barbour S.L. 1996. The relationship of the unsaturated soil shear strength to the soil-water characteristic curve. *Canadian Geotechnical Journal* 33(3): 440-448. https://doi.org/10.1139/t96-065

Fredlund D.G. 2000. The 1999 RM Hardy Lecture: The implementation of unsaturated soil mechanics into geotechnical engineering. *Canadian Geotechnical Journal* 37(5): 963-986. https://doi.org/10.1139/t00-026

Fredlund D.G., Rahardjo H. and Fredlund M.D. 2012. *Unsaturated soil mechanics in engineering practice*. John Wiley & Sons, Inc.

Gitirana Jr G.F.N. and Fredlund D.G. 2004. Soil-water characteristic curve equation with independent properties. *Journal of Geotechnical and Geoenvironmental Engineering* 130(2):209-212. https://doi.org/10.1061/(ASCE)1090-0241(2004)130:2(209)

Gitirana Jr G.F.N. and Fredlund D.G. 2016. Statistical assessment of hydraulic properties of unsaturated soils. *Soils and Rocks* 39(1):81-95. https://doi.org/10.28927/SR.391081

Haghverdi A., Cornelis W.M. and Ghahraman B. 2012. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology* 442:46-54. https://doi.org/10.1016/j.jhydrol.2012.03.036

Likos W.J. and Jaafar R. 2013. Pore-Scale Model for Water Retention and Fluid Partitioning of Partially Saturated Granular Soil. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(5): 724–737. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000811

Li Y. and Vanapalli S.K. 2021. Prediction of soil-water characteristic curves using two artificial intelligence (AI) models and AI aid design method for sands. *Canadian Geotechnical Journal* 59(1):129-143. https://doi.org/10.1139/cgj-2020-0562

Minasny B., Mcbratney A.B. and Bristow K.L. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma* 93(3–4): 225–253. https://doi.org/10.1016/S0016-7061(99)00061-0

Pham K., Kim D., Yoon Y. and Choi H. 2019. Analysis of neural network based pedotransfer function for predicting soil water characteristic curve. *Geoderma* 351:92-102. https://doi.org/10.1016/j.geoderma.2019.05.013

Pham K., Kim D., Le C.V. and Won J. 2023. Machine learning-based pedotransfer functions to predict soil water characteristics curves. *Transportation Geotechnics*, p. 101052.

Rudiyanto, Minasny B., Chaney N.W., Maggi F., Giap S.G.E., Shah R.M., Fiantis D. and Setiawan B.I. 2021. Pedotransfer functions for estimating soil hydraulic properties from saturation to dryness. *Geoderma* 403:115194. https://doi.org/10.1016/j.geoderma.2021.115194

dos Santos Pereira S.A., Silva Junior A.C., Mendes T.A., Gitirana Junior G.D.F.N. and Alves R.D. 2023. Prediction of Soil–Water Characteristic Curves in Bimodal Tropical Soils Using Artificial Neural Networks. *Geotechnical and Geological Engineering*, 1-20. https://doi.org/10.1007/s10706-023-02716-x

Satyanaga A., Rahardjo H., Zhai Q., Moon S.-W. and Kim J. 2023. Modelling Particle-Size Distribution and Estimation of Soil–water Characteristic Curve utilizing Modified Lognormal Distribution function. *Geotechnical and Geological Engineering* 1-19. https://doi.org/10.1007/s10706-023-02638-8

Silva A.C., Armindo R.A. and Prevedello C.L. 2020. Splintex 2.0: A physically-based model to estimate water retention and hydraulic conductivity parameters from soil physical data. *Computers and Electronics in Agriculture* 169:105157. https://doi.org/10.1016/j.compag.2019.105157

Silva, F.C.D. 2022. *Estimativa de curvas características de solos tropicais brasileiros via métodos indiretos (in portuguese)*. Masters dissertation. Programa de Pós-Graduação em Geotecnia, Departamento de Engenharia Civil e Ambiental, Faculdade de Tecnologia. Universidade de Brasília, Brasília, DF. 236p.

Vanapalli S.K., Fredlund D.G., Pufahl D.E. and Clifton A.W. 1996. Model for the prediction of shear strength with respect to soil suction. *Canadian Geotechnical Journal* 33(3): 379-392. https://doi.org/10.1139/t96-060

Wang J.P., Hu N. François B. and Lambert P. 2017. Estimating water retention curves and strength properties of unsaturated sandy soils from basic soil gradation parameters. *Water Resources Research* 53(7): 6069–6088. https://doi.org/10.1002/2017WR020411

Wang L., Zhang W. and Chen F. 2019. Bayesian approach for predicting soil-water characteristic curve from particle-size distribution data. *Energies* 14(15). https://doi.org/10.3390/en12152992

Zhai Q., Rahardjo H., Satyanaga A. and Dai G. 2020. Estimation of tensile strength of sandy soil from soil–water characteristic curve. *Acta Geotechnica* 15:3371-3381. https://doi.org/10.1007/s11440-020-01013-8

Zhai Q., Zhu Y., Rahardjo H., Satyanaga A., Dai G., Gong W., Zhao X. and Ou Y. 2023. Prediction of the soil–water characteristic curves for the fine-grained soils with different initial void ratios. *Acta Geotechnica* 1-10. https://doi.org/10.1007/s11440-023-01833-4