Machine Learning-Based Estimation of Soil's True Air-Entry Value from GSD Curves

Es-haghi, M. S., Rezania, M. & Bagheri, M

DOI 10.1016/j.gr.2022.06.012
ISSN 1342-937X

Publisher: Elsevier

Machine Learning-Based Estimation of Soil's True Air-Entry Value from GSD Curves

Mohammad Sadegh Es-haghi[1,2], Mohammad Rezania[3,*], Meghdad Bagheri[4]

[1] Escuela Técnica Superior de Ingenieros Navales (ETSIN), Universidad Politécnica de Madrid (UPM), Av. de la Memoria, 4, 28040 Madrid, Spain

[2] Centre Internacional de Mètodes Numèrics a l'Enginyeria (CIMNE), Edifici C1, Campus Norte, UPC, Gran Capitán s/n, 08034 Barcelona, Spain

[3] School of Engineering, University of Warwick, Coventry, CV4 7AL, UK

[4] School of Energy, Construction and Environment, Coventry University, Coventry, UK

[*]Corresponding author; Email: m.rezania@warwick.ac.uk

Abstract

The application of machine learning (ML) methods has proven to be promising in dealing with a wide range of geotechnical engineering problems in recent years. ML methods have already been used for the prediction of soil water retention curves (SWRC) and estimation of air-entry values (AEV). However, the reported works in the literature are generally based on limited data and conventional, less accurate approaches for AEV estimation. In this paper, a large database, known as UNsaturated SOil hydraulic DAtabase (UNSODA), is studied and the conventional and true AEVs of 790 soil samples are estimated based on determination methods reported in the literature. A ML approach is then employed for the development of a predictive model for the estimation of true AEV from water content-based SWRCs of a wide range of soil types taking into account the impact of bulk density and grain size distribution parameters. The obtained results reveal an enhanced accuracy in AEV determination, featuring R^2 values of 0.964, 0.901 and 0.851 for

training, validation, and testing data, respectively, which confirm the marked performance of the developed ML model. Based on the results of a sensitivity analysis, the particle sizes of 50 and 250 μm are found to have the highest impact on the AEV estimation.

Keywords: Air-entry value, Soil water retention, Grain size distribution, Machine learning

List of Notations

P	particle size
p'_c	preconsolidation pressure
R^2	coefficient of determination
s	soil suction
s_{ae}	suction at air-entry
s_{rs}	residual soil suction
S_r	degree of saturation
w	gravimetric water content
θ	volumetric water content
AEV	air-entry value
AI	artificial intelligence
ANN	artificial neural network
EPR	evolutionary polynomial regression
GSD	grain size distribution
GP	genetic programming
HCT	high-capacity tensiometer
MEP	multi-expression programming
ML	machine learning
MLP	multilayer perceptron
RF	random forest
RMSE	root mean square error
SVM	support vector machine
SWRC	soil water retention curve
UNSODA	UNsaturated SOil hydraulic DAtabase
USDA	United States Department of Agriculture

Introduction

In unsaturated soil mechanics, the relationship between the amount of water held by the soil and the pore-water tension (suction) is generally presented in the form of water content (w) versus suction (s), or degree of saturation (S_r) versus suction in a semi-logarithmic graph known as soil water retention curve (SWRC). Generally, three distinct parts are recognized on an SWRC (Figure 1). Initially, with the soil undergoing drying, the curve remains a relatively flat line until the air-entry value (AEV); during this stage, with an increase of suction, the pores remain saturated (S_r = 100%) with pore-water being under tension (saturated state or boundary effect zone). As suction exceeds the AEV, air breaks into the larger pores, and a continuous network of air-filled and water-filled pores is formed. During this stage, the degree of saturation (or water content) is progressively decreased as more water evaporates or leaves the pores (partially saturated state or transition effect zone). Finally, with an increase in suction, a state is reached where water only remains at the particle contacts and is no longer continuous with the pore space. At this stage, the SWRC almost flattens meaning that much less water will be expelled with an increase in suction, and the soil dries without significant volume changes (residual state). The two inflation points, namely AEV and residual suction (s_{rs}), are considered as fundamental parameters for the determination of soils' water retention properties, with the former being a key input parameter in several unsaturated constitutive models (e.g., Alonso et al. 1990; Russell and Khalili 2006), hence, highlighting the importance of its accurate estimation from SWRCs.

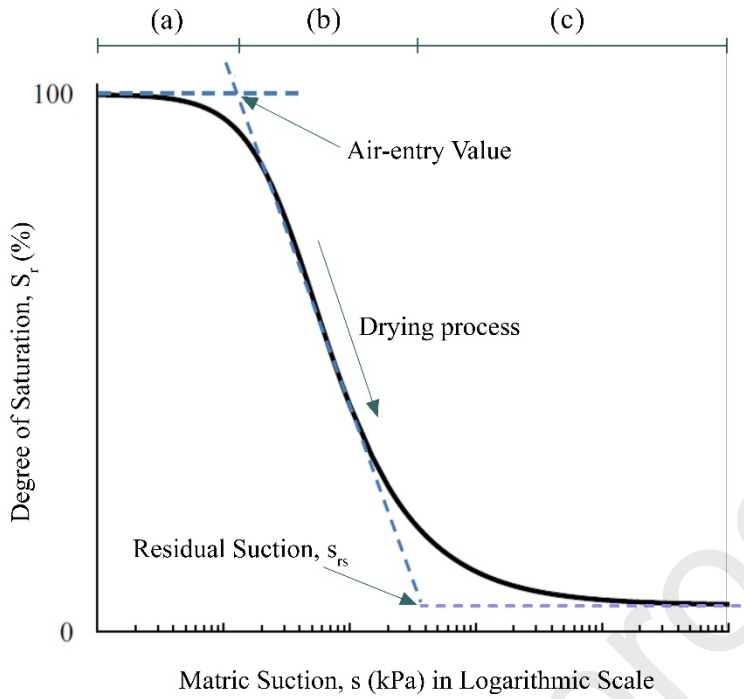


Figure 1 Typical SWRC in $S_r - s$ plane: (a) boundary effect zone; (b) transition zone; (c) residual zone

On a degree of saturation-based SWRC, the AEV is graphically determined as the intersection point of a horizontal tangent line drawn to the saturated portion, with a tangent line drawn to the transition portion of the curve (see Figure 1). Pasha et al. (2016) showed that using the same method to derive AEV from water content-based SWRCs can produce highly erroneous results. They proposed a simple method for estimation of true AEV from water content-based SWRC. This method was based on plotting the data on both semi–log and log–log scales on the same graph. In this way, the saturated part of the curve is identified by a linear or bilinear behavior in a semi–log plot, and the unsaturated part of the curve is represented by a straight line (linear behavior) on the log–log plot. Therefore, by drawing these two complementary curves on a single graph, the AEV as the boundary between saturated and transition (unsaturated) zones can be readily identified. This graphical technique can be used for the evaluation of true AEV from SWRC data where the information on the volume change of the sample during testing is not available.

SWRCs are generally developed based on experimental methods such as axis translation (Bagheri et al. 2019a), negative water column (Pagano et al. 2016) and pressure plate (Tarantino et al. 2011). Such methods are generally time-consuming and expensive, and in most cases produce discontinuous measurements, bringing difficulties in the accurate determination of the AEV. For such measurements, the method proposed by Pasha et al. (2016) appears promising. Bagheri (2018), Bagheri et al. (2018), and Bagheri and Rezania (2022) later showed that direct measurement of soil suction changes using high-capacity tensiometers (HCT) can rectify the need for such approximation methods, as an accurate estimation of AEV can be readily obtained from continuous measurements of suction variations with water content. However, this method is also limited to the maximum capacity of HCTs which is typically in the range of 1.5 – 2.0 MPa. Other methods including estimation of SWRC based on the soils' grain size distribution (GSD) curve (Alves et al. 2020; Zhai et al. 2020), statistical methods (Saxton et al. 1986; Chiu et al. 2012), and artificial intelligence (AI) -based methods (Schaap and Leij 1998) are also available. However, it is apparent that accurate estimation of AEV is profoundly subordinate to the accuracy of predicted SWRCs. Furthermore, the reported works on statistical and AI methods generally utilized databases that were not large enough to include a variety of soil types. It is therefore imperative to consider a relatively large database to directly estimate true AEV from common soil parameters.

Recently, machine learning (ML) approaches have proven to be promising in solving nonlinear and complicated problems using large databases (Rezania 2008; Javadi and Rezania 2009a; Jin and Yin 2020; Zhang et al. 2020; Zhang et al. 2021; Wang et al. 2020 Aug; Zhang et al. 2022; Wang et al. 2020 Nov). In geotechnical engineering, the ML algorithms have been favorably employed for predicting various phenomena such as the settlement of shallow foundations on cohesionless soils (Rezania and Javadi 2007), thermo-hydro-mechanical behavior of hydrate reservoirs (Zhou et al. 2020), non-stationary and non-Gaussian geotechnical properties (Shi and Wang 2021), soil constitutive modeling (Javadi and Rezania 2009b; Rezania and Ma 2019) and

suction distribution in shallow soil layers (Cheng et al. 2020). In these studies, various ML algorithms including support vector machine (SVM), multi-expression programming (MEP), genetic programming (GP), evolutionary polynomial regression (EPR), and random forest (RF) have been used. The effectiveness of ML methods has led to their employment in the estimation of SWRCs (Jain et al. 2004; Moreira de Melo and Pedrollo 2015). A few works have been also reported in the literature on the estimation of the SWRC from GSD curves employing ML methods (D'Emilio et al. 2018; Amanabadi et al. 2019; Li and Vanapalli 2021). However, estimation of the AEV by ML algorithms has been rarely studied. Recently, Wang et al. (2020) utilized ML algorithms for the prediction of AEV of compacted soils based on some physical properties. The authors have used the parameters of sand content, fines content, plasticity index, initial water content and initial void ratio as input variables. In addition, they consider the conventional less accurate method of estimating AEV from water content-based SWRC which can be considered as the shortcoming of this study. With these explanations, a thorough understanding of ML algorithms in the prediction of the true AEV is imperative and worth investigating.

In this paper, the UNsaturated SOil hydraulic DAtabase (UNSODA) (Leij et al. 1996) has been thoroughly reviewed and for a considerable number of soil samples, true AEVs are estimated from water content-based SWRCs using the method proposed by Pasha et al. (2016) and compared to the conventional method for AEV evaluation. Furthermore, the influence of physical soil parameters (e.g., bulk density, grain size, etc.) on AEV is investigated. Finally, a neural network model is employed to estimate true AEV. A sensitivity analysis is also performed in order to indicate which parameters are essential for true AEV prediction.

UNSODA

UNSODA (UNsaturated SOil hydraulic DAtabase) is a collection of data for unsaturated hydraulic soil properties in Microsoft Access-97 format. Data for 790 soil samples have been

deposited in 36 tables, which have stored data in rational groups of fields including relevant information. A 4-digit code has been allocated to each sample. Figure 2 shows the frequency histogram of soil classification of 790 samples according to USDA (United States Department of Agriculture) system and the summary of their geographical distribution represented in UNSODA. Most of the samples of the database are from Europe and North America. Most of the samples are coarse-textured soils although there is an acceptable number of soil samples with fine textures.

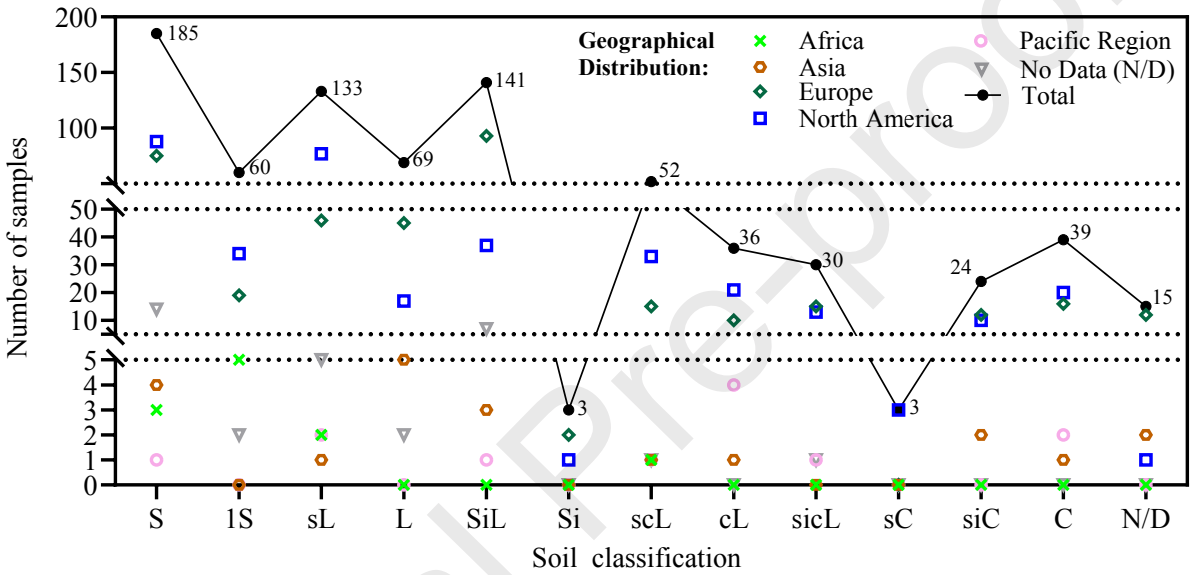


Figure 2 Geographical and textural distribution according to the USDA classification of soils in UNSODA (S: Sand, lS: loamy Sand, sL: sandy Loam, L: Loam, SiL: silty Loam, Si: Silt, scL: sandy clay Loam, cL: clay Loam, sicL: silty clay Loam, sC: sandy Clay, siC: silty Clay, C: Clay, N/D: Not Determined)

UNSODA provides a wide range of soil properties as shown in Table 1. It must be noted here that bulk density, particle density, porosity, saturated conductivity, and saturated water content are the only parameters required for the estimation of true AEV in the present study. In addition, UNSODA provides appropriate information about the grain size distribution of samples; however, unfortunately, there is no uniform set of particle sizes in UNSODA, and the reported values of particle fraction are presented for different sizes in different samples due to the diversity in

experimental methods. For instance, in the soil sample 1090, particle fractions for the particle sizes of 2, 50, 125, 250, 500, 1000, and 2000 μm have been determined; however, in the soil sample 1087 only the particle size of 2, 50, and 2000 μm have been considered.

In this study, the particle size fractions of 2, 20, 50, 250, 500, 1000, and 2000 μm, which have the most frequency are used. Figure 3 presents the frequency of particle sizes in the UNSODA database.

Table 1. Soil properties provided in UNSODA database

Parameter	Definition	Available Values	Missing Values	Considered in this study?
bulk density	Bulk density as a mass of solids per bulk volume	762	28	Yes
particle density	Mass of solids per volume of solids	339	391	Yes
porosity	Volume of voids per bulk volume	270	420	Yes
OM content	**Organic Matter Content**. The mass of organic matter content as a percentage of the total solid mass.	388	402	No
k_{sat}	**Saturated Conductivity.** The measured saturated hydraulic conductivity	429	361	Yes
$theta_{sat}$	**Saturated Water Content.** The experimental water content of a water-saturated sample	305	485	Yes
CEC	**Cation Exchange Capacity**. in cmol of charge per kg of dry soil (i.e., meq/100 g soil)	150	640	No
pH	Measured soil pH	300	490	No
electrolyte level	The approximate total solute concentration of the soil solution during the experiments	26	764	No
SAR	Sodium Adsorption Ratio	80	710	No
ESP	Exchangeable Sodium Percentage	19	771	No
EC	Electrical conductivity of the saturation extract	62	728	No
free Fe Al oxide	The mass fraction of the Fe and Al oxides as a percentage of the total solid phase	14	776	No

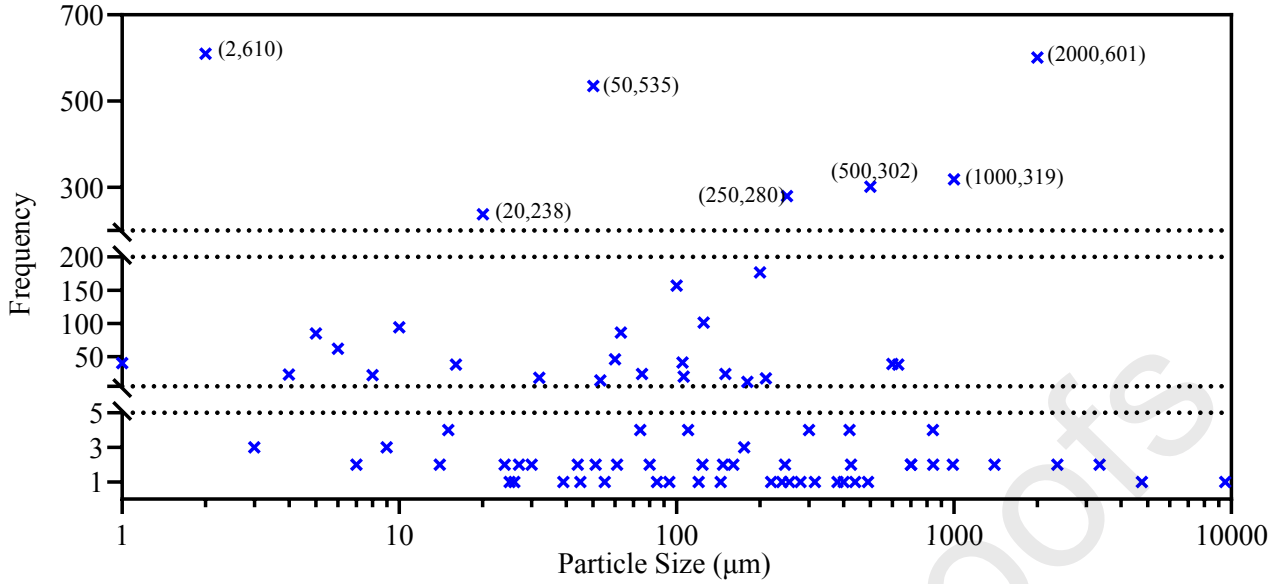


Figure 3 Frequency of the particle size in UNSODA database

Table 2 presents the number of hydraulic curves and the number of data pairs partitioned into drying and wetting parts, as well as curves obtained from laboratory or field measurements. In this study, the water retention curves (a total of 902 curves) which have the most data were used. From the 902 water retention (h-θ) curves, the drying curves, having a total number of 867 curves obtained from both laboratory and field measurements, were deemed sufficient and considered for the analysis.

Table 2. Summary of the number of hydraulic curves

Hydraulic Curves		Number of Curves		Total Number of Data Pairs		Average Number of Data Pairs in each Curve	
		Field	Lab	Field	Lab	Field	Lab
Water Retention (h-θ)	Drying	137	730	2621	8066	19.1	11.0
	Wetting	2	33	8	528	4.0	16.0
	Total	902		11223		12.4	
Hydr. Conductivity (h-K)	Drying	133	730	2826	6187	21.2	8.5
	Wetting	0	8	0	71	-	8.9
	Total	871		9084		10.4	
Hydr. Conductivity (θ-K)	Drying	294	293	5391	5177	18.3	17.7
	Wetting	0	20	0	216	-	10.8
	Total	607		10784		17.8	
Soil Water Diffusivity (θ-D)	Drying	56	92	1282	1456	22.9	15.8

	Wetting	0	2	0	13	-	6.5
	Total	150		2751		18.3	

Figure 4 presents the scatter plot of the main drying branch of SWRCs reported in the UNSODA database for both field and lab measurements. In this Figure, the range of changes and how suction and water content data are distributed for the lab and field data are well described. As it turns out, in lab-based methods, the results involve a wider range of suction and water content values.

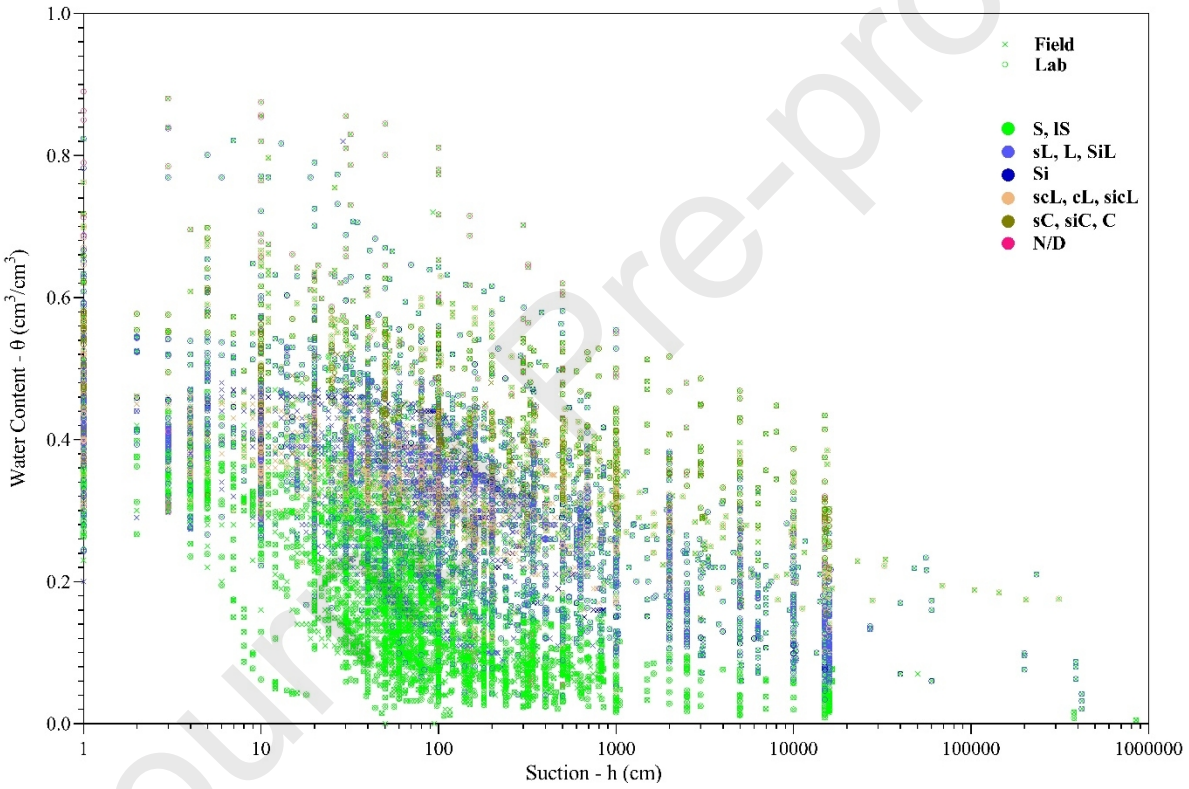


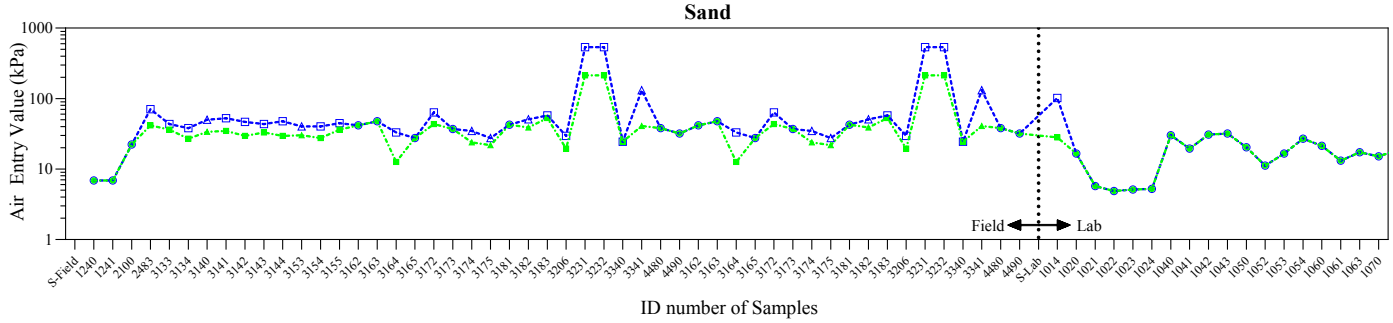
Figure 4 Main drying branch of SWRCs in UNSODA database (S: Sand, lS: loamy Sand, sL: sandy Loam, L: Loam, SiL: silty Loam, Si: Silt, scL: sandy clay Loam, cL: clay Loam, sicL: silty clay Loam, sC: sandy Clay, siC: silty Clay, C: Clay, N/D: Not Determined)

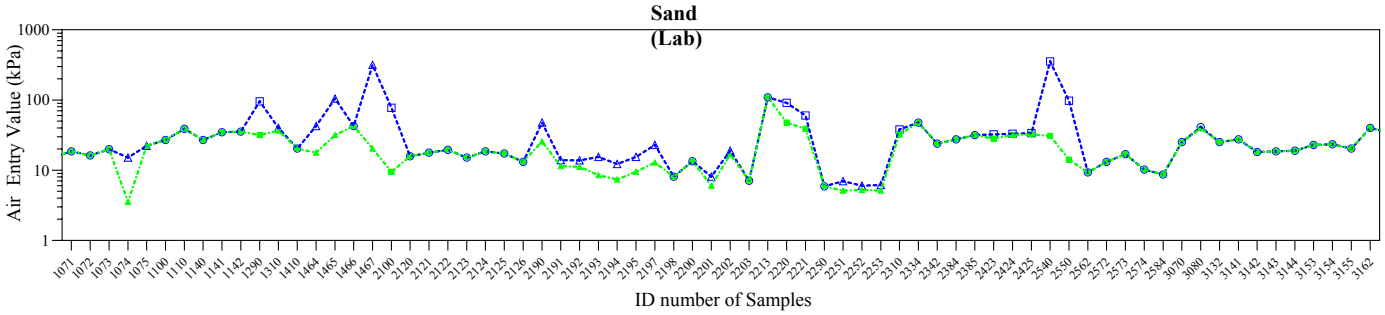
True air entry value

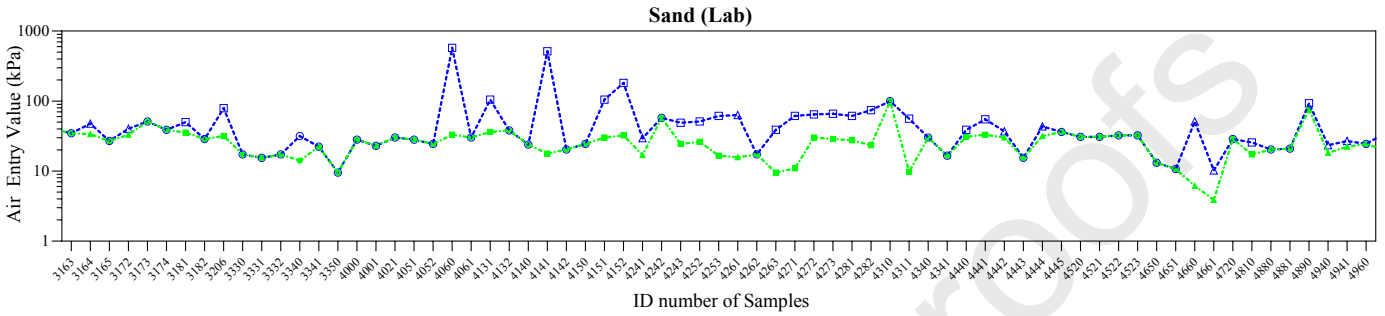
Although numerous researchers recognize the first break in the drying branch of the water content-based SWRC as evidence of air-entry point (Fredlund and Xing 1994; Zhai and Rahardjo

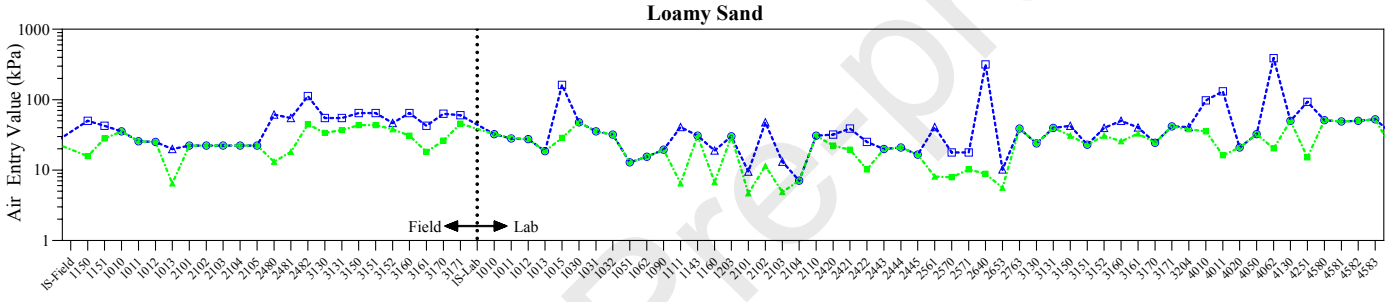
2012; Bagheri et al., 2019b, Rezania et al., 2020), this may not be true for some deformable soils, depending on factors such as the mechanical characteristics of the soil, pore-size distribution index, and the stress history of the soil. In the graphical method presented by Pasha et al. (2016), estimation of the true AEV was carried out based on the stress history of the soil, and consideration of three stress states, namely normally consolidated, overconsolidated with $s_{ae} > p'_c$, and overconsolidated with $s_{ae} < p'_c$, where s_{ae} is the suction at air-entry and p'_c is the preconsolidation pressure of the soil.

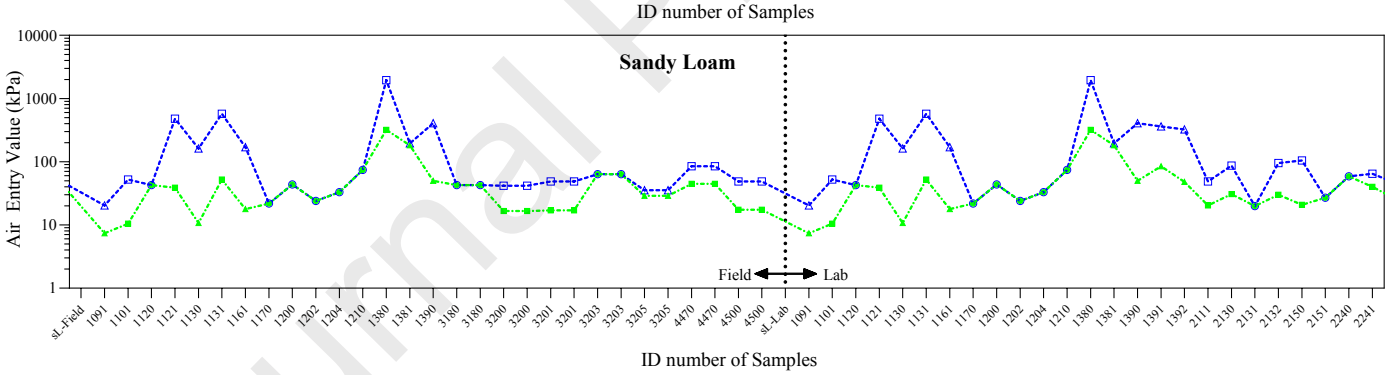
UNSODA database contains 867 SWRCs, 221 of which could not be used for various reasons, such as an insufficient number of data points or soil not entering the transition zone Therefore, only SWRCs that would allow for appropriate graphical estimation of AEV were selected and considered in his study. Of the remaining 644 curves, 212 curves are related to normally consolidated state, 182 related to overconsolidated state with $s_{ae} > p'_c$, and 250 related to overconsolidated state with $s_{ae} < p'_c$. For all these 644 curves, one by one, the AEVs were obtained following the method proposed by Pasha et al. (denoted true AEV) and also following the conventional method of intersection of the tangent lines to the boundary effect zone and transition zone (denoted AEV) as shown in Figure 5. The horizontal axis of the graphs of Figure 5 represents the codes allocated to the samples, which can be traced in the UNSODA database. As some examples including the soils of 4440, 4612, 2670 and 1135, the way to obtain the true AEV value is shown in Figure 6. In addition, Figure 7 presents the difference between the true AEV and conventional AEV which can be in a range of 30 to 300 kPa.

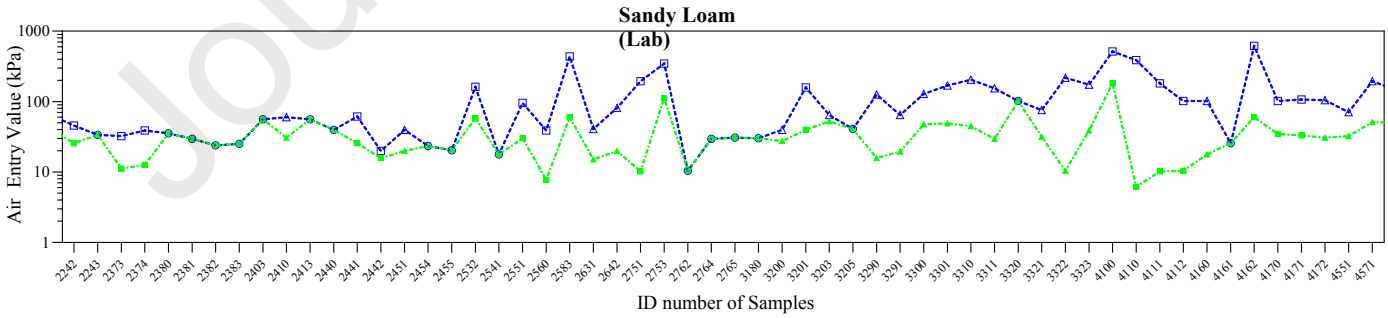


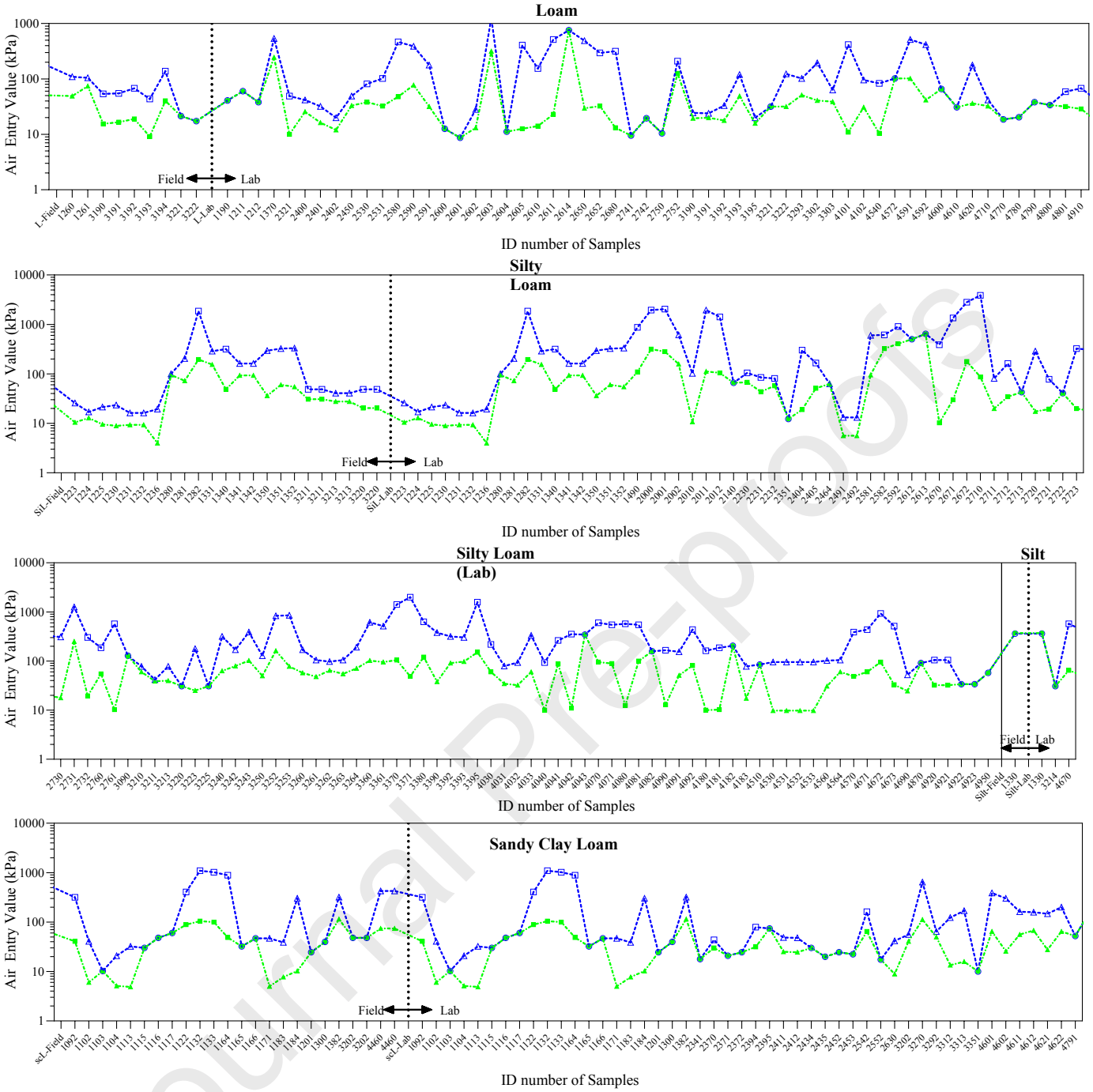












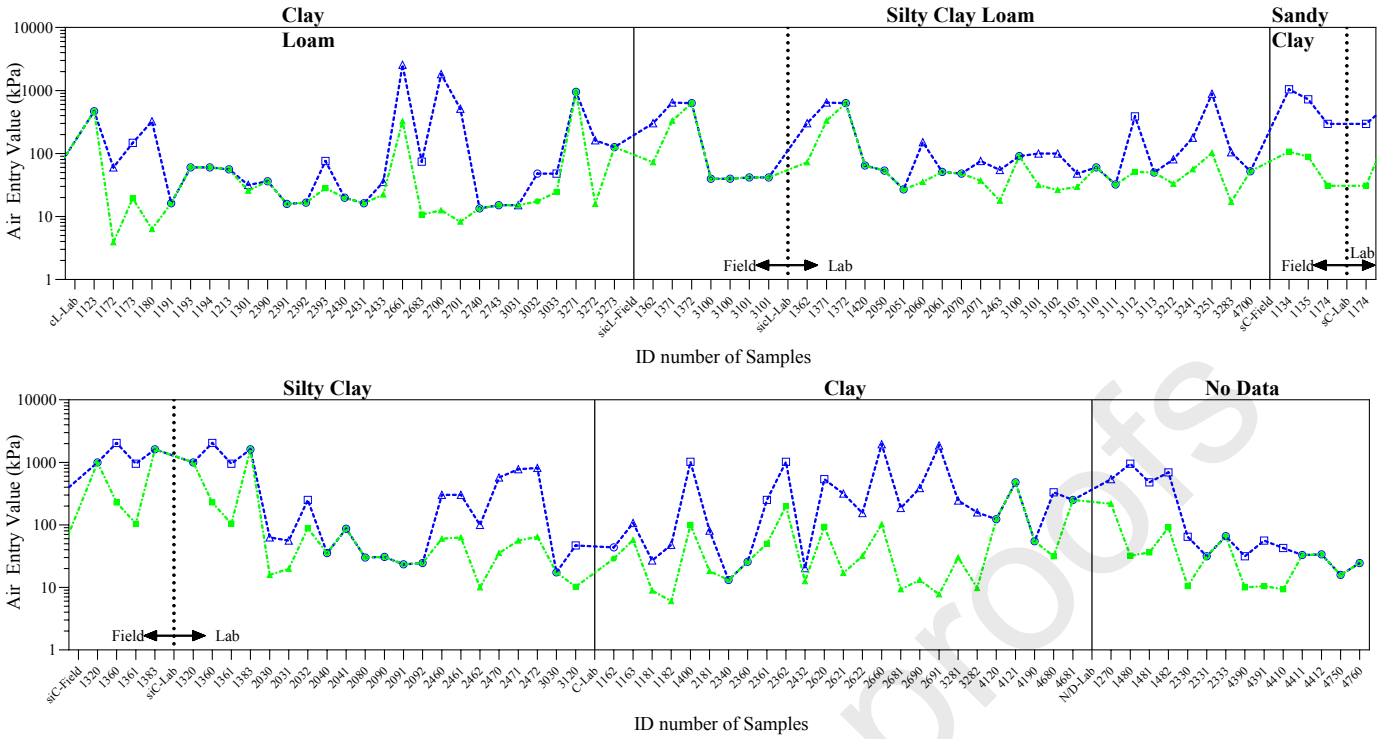


Figure 5. AEV and true AEV for each sample of UNSODA (Blue: true AEV, Green: miscalculated AEV, Triangle: normally consolidated, Square: overconsolidated with $s_{ae} > p_c'$, and Circle: overconsolidated with $s_{ae} < p_c'$)

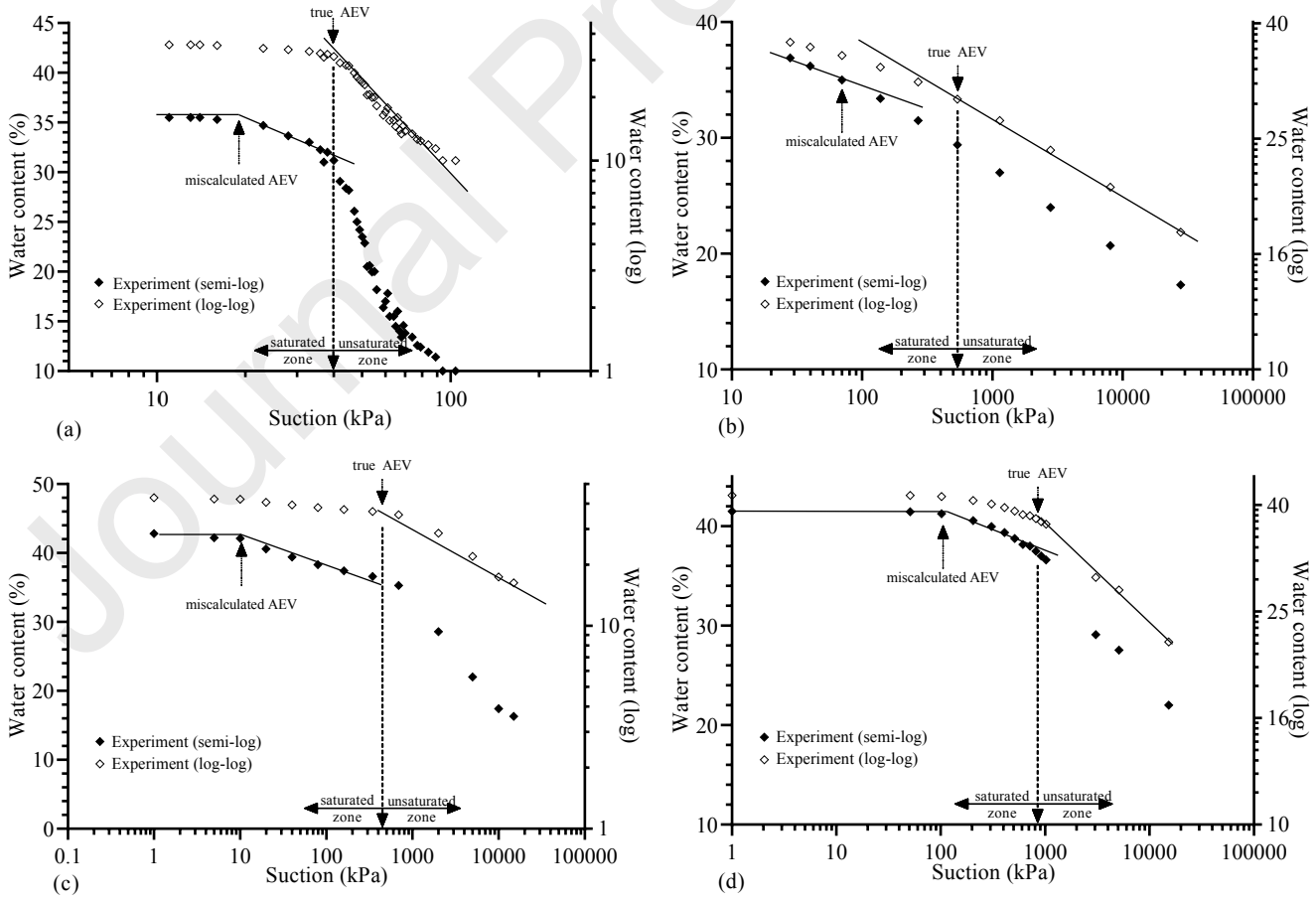


Figure 6. Graphical determination of True and Miscalculated AEVs for some soil samples in the UNSODA database (a: sample 4440, b: sample 4612, c: sample 2670, d: sample 1135)

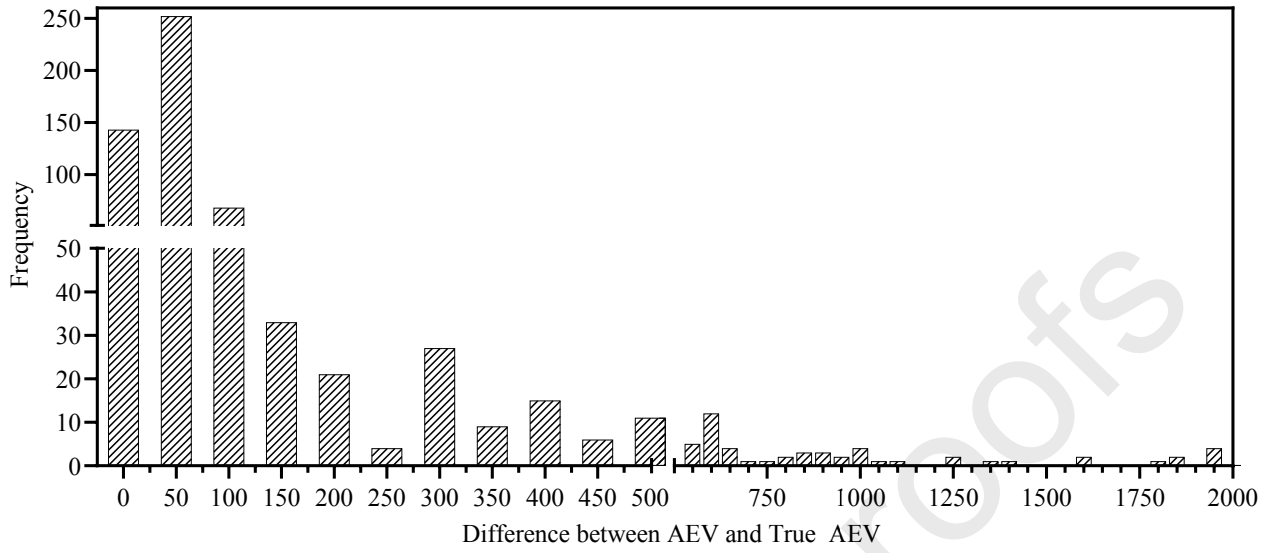


Figure 7. Histogram of the difference between the conventional and true AEVs

Preparing the database

The extracted data from UNSODA database is presented in a way that is not convenient for training the ML models. On the other hand, as it is shown in Table 1, some of the parameters are not provided for all samples. For instance, the data points of GSD curves are presented in different particle size values (see Figure 3). Therefore, at first, to extract information from UNSODA and arrange it so that it could be used to train neural networks, coding was done in MATLAB software. The developed code allowed also for extrapolation of the missing data and formation of the complete database required for ML.

As shown in Figure 3, the most common particle size values in UNSODA are $P_{2\mu}$, $P_{20\mu}$, $P_{50\mu}$, $P_{250\mu}$, $P_{500\mu}$, $P_{1000\mu}$, and $P_{2000\mu}$. As a result, the fraction corresponding to these values were selected from the GSD curve as part of the ML inputs, and for samples that did not have these particle size distribution values, the corresponding values were estimated using the method proposed in Vaz et al. (2020). To overcome the common limitation of GSD curves, which is the lack of standardization

for granulometric fractions collected from various soil analysis methods, Vaz et al. (2020) evaluated and compared the performance of several GSD equations and showed that the three-parameter Equation 1 has an acceptable performance with root mean square error (RMSE) of 0.463, 0.205, and 0.013, respectively for sand, silt, and clay. Therefore, this equation was used in this study:

$$F(d) = \left[1 + \left(\frac{a}{d}\right)^{b}\right]^{-c} \quad (1)$$

In the above equation, a, b, and c are equation parameters, d is the particle size and $F(d)$ is the fraction corresponding to the particle size of d. Thus, for each sample, the parameters of Equation 1 were extracted and the fraction values in demanded particle size (if not present) were calculated. For instance, in Figure 8, the fitted curve for sample 3340 in UNSODA is presented and the value of $P_{20\mu}$ and $P_{50\mu}$ for this sample were estimated. This process was done for all samples to complete the database.

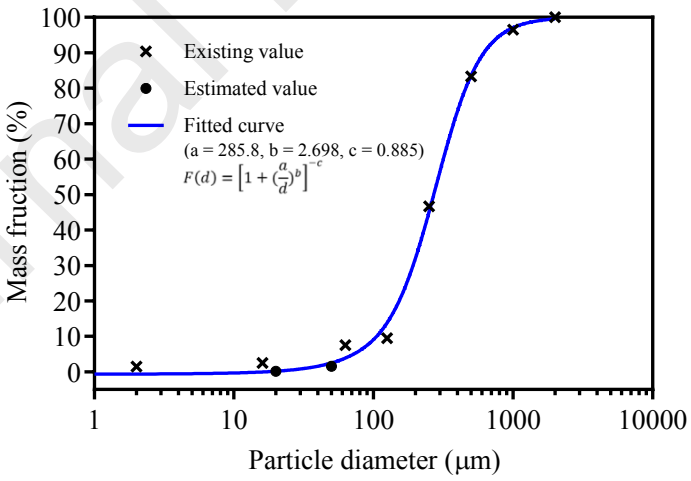


Figure 8. Fitted GSD curve for sample 3340

The detailed statistics of all variables in the database, with the values of minimum, maximum, median, mean, and standard deviation are summarized in Table 3 and their descriptive frequency histograms are shown in Figure 9. Bulk density and porosity have the highest frequency

at 1.4 to 1.6 and 0.4 to 0.5 respectively and follow an approximately normal distribution (Figures 9a and 9b). Regarding mass fractions (Figures 9c to 9j), it is observed that the highest frequency of $P_{2\mu}$ is between 0.0 and 0.1 for 279 soils (Figure 9c), while the highest frequency of $P_{2000\mu}$ is between 0.9 and 1.0 for 641 soils (Figure 9j). In terms of the true and conventional AEVs, most of the values are located in the range between 0 to 50 kPa (Figures 9k and 9l), while the true AEV distribution is slightly more uniform than the conventional AEV distribution. Since the parameters of $P_{2000\mu}$ and $P_{1000\mu}$ do not differ in various samples, they are not considered in subsequent analyses.

Table 3. Descriptive statistics of all variables in the database

	Variable	Min	Max	Median	Mean	Standard deviation
Inputs	Bulk Density (g/cm^3)	0.459	1.970	1.500	1.472	0.212
	Porosity	0.264	0.915	0.444	0.461	0.093
	$P_{2\mu}$	0.000	0.697	0.116	0.153	0.132
	$P_{20\mu}$	0.000	0.920	0.277	0.282	0.226
	$P_{50\mu}$	0.000	1.000	0.393	0.416	0.303
	$P_{250\mu}$	0.000	1.000	0.853	0.783	0.225
	$P_{500\mu}$	0.200	1.000	0.976	0.909	0.145
	$P_{1000\mu}$	0.651	1.000	0.997	0.971	0.061
	$P_{2000\mu}$	0.776	1.000	1.000	0.998	0.016
Outputs	AEV (kPa)	3.548	1621.800	30.199	51.232	102.882
	True AEV (kPa)	4.898	3890.400	50.119	185.427	371.372

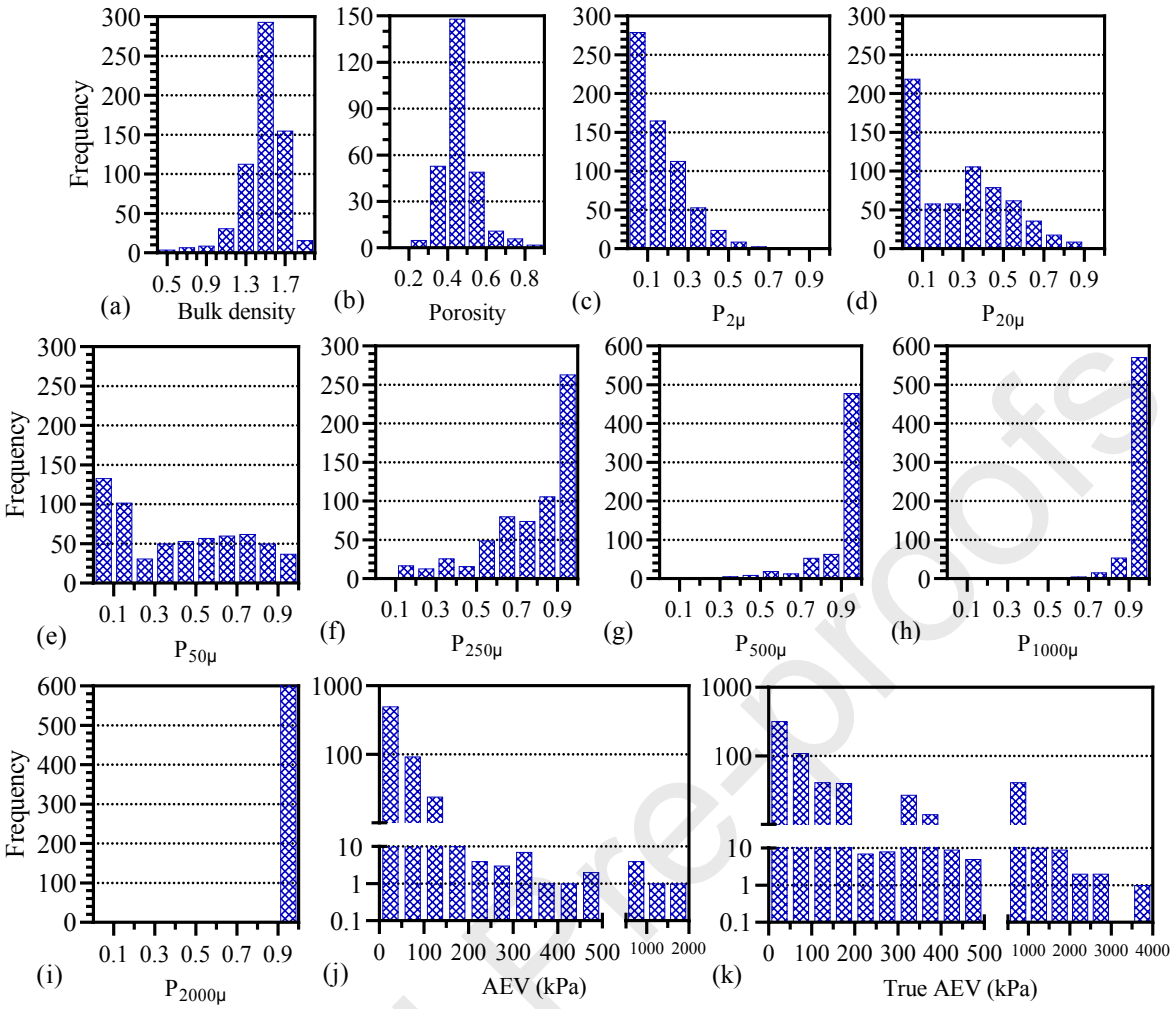


Figure 9. Frequency histograms of the variables: (a) Bulk Density, (b) Porosity, (c) $P_{2\mu}$, (d) $P_{20\mu}$, (e) $P_{50\mu}$, (f) $P_{250\mu}$, (g) $P_{500\mu}$, (h) $P_{1000\mu}$, (i) $P_{2000\mu}$, (j) AEV, (k) True AEV

Figure 10 presents a comparison of the AEV and true AEV obtained for different soil classifications. It is obvious that the range of true AEV has shifted to larger values than conventional AEV.

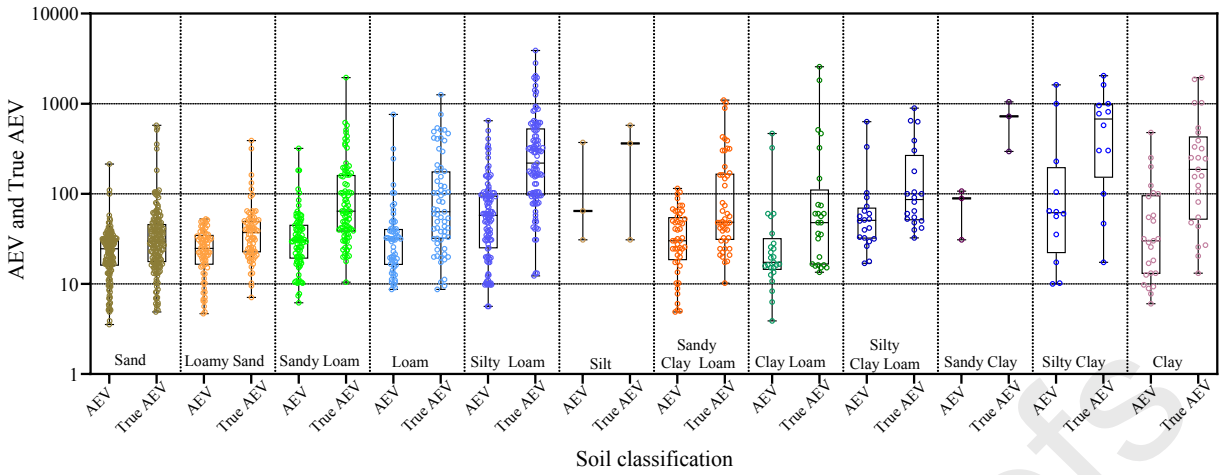


Figure 10. Comparison of AEV and true AEV of soils with different classifications in the UNSODA database

Variations of AEV and true AEV with each input soil property are shown in graphs of Figure 11. Also shown in the figure is the linear trend lines fitted to the data along with the coefficient of determination (COD), R^2. The COD allows for assessment of the fitting accuracy and is mathematically expressed as:

$$R^2 = 1 - \left[\Sigma_{i=1}^{N}(Y_i - \hat{Y}_i)^2 / \Sigma_{i=1}^{N}(Y_i - \overline{Y}_i)^2\right] \quad (2)$$

$$\overline{Y}_i = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad (3)$$

where Y_i and $\hat{Y}_i$ are real and predicted output values for the i^{th} dataset, N is the number of outputs, and $\overline{Y}_i$ is the average value of the real outputs. Overall, a comparatively low fitting accuracy ($R^2 <$ 0.177) was obtained for variations of both AEV and true AEV with input soil properties. The parameters $P_{50\mu}$, $P_{250\mu}$, $P_{500\mu}$, $P_{20\mu}$, $P_{2\mu}$, $P_{1000\mu}$, and bulk density, having the highest R^2 values were selected for developing the ML model and the parameters $P_{2000\mu}$ and porosity, having the lowest R^2 values, were omitted. This was done to avoid complexity associated with developing the ML model.

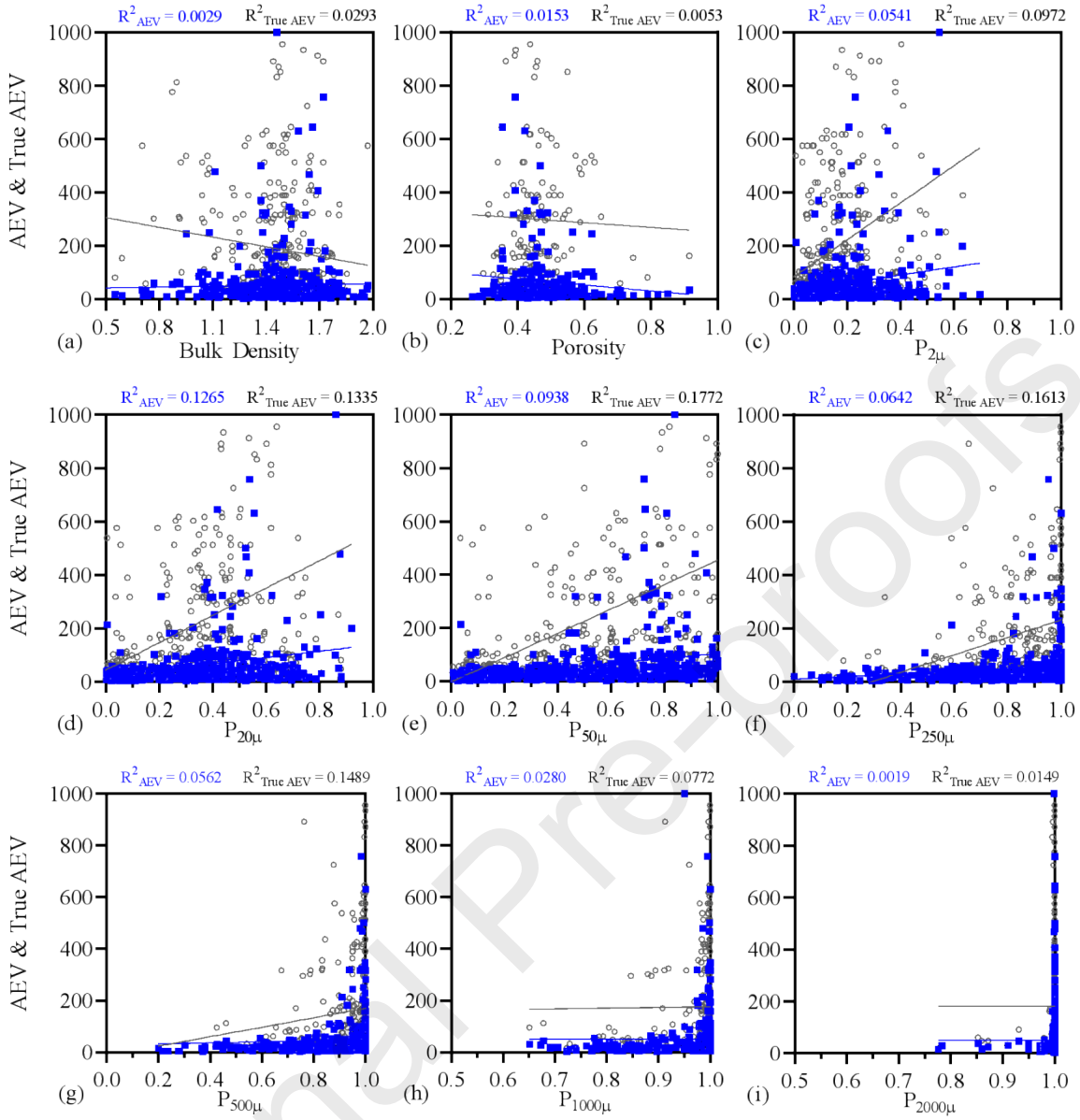


Figure 11. Basic linear fittings between AEV and true AEV and each input soil property: (a) Bulk Density, (b) Porosity, (c) $P_{2\mu}$, (d) $P_{20\mu}$, (e) $P_{50\mu}$, (f) $P_{250\mu}$, (g) $P_{500\mu}$, (h) $P_{1000\mu}$, (j) $P_{2000\mu}$

Results of analysis

According to Figure 11, the database is described by seven input parameters namely, bulk density, $P_{2\mu}$, $P_{20\mu}$, $P_{50\mu}$, $P_{250\mu}$, $P_{500\mu}$, and $P_{1000\mu}$, and one output parameter namely, true AEV. In order to accurately generate the prediction model, the ML method of multilayer perceptron (MLP) was selected because of its simplicity and availability (comparison between different methods of

machine learning is not the purpose of the current study). MLP, developed in the 1960s, is a class of artificial neural networks (ANNs).

A basic building block in many machine learning applications is an artificial neural network (ANN), which is a method based on consideration of the nervous system of living things. These networks prepare a technique for dealing with complex pattern-oriented problems. The nonparametric nature of ANNs permits models to be established without including any previous awareness of the distribution of the input population or conceivable interaction influences between variables as required by frequently used parametric statistical methods.

For instance, in multiple regression it is imperative that the error term of the regression equation be distributed normally (in other words, $\mu = 0$) and also be nonheteroscedastic. Another statistical technique is discriminant analysis which is usually used for doing classification, however, discriminant analysis needs that the predictor variables be multivariate normally distributed. The easiness of developing a domain problem solution is expanded with ANNs because such presumptions are eliminated from ANN models. In addition, ANNs' ability to constitute nonlinear models as well as conventional linear models is another point contributing to the success of them and, therefore, artificial neural network solutions are useful across a more varied range of problem varieties (both linear and nonlinear) (Walczak, 2019).

ANN can be described as a multivariate and multi-dimensional function $f{:}\mathbb{R}^n \longrightarrow \mathbb{R}^m$. It consists of an input layer of n neurons (input values), an output layer of m neurons (output values) and an arbitrary number of interior layers with a variable number of neurons, called hidden layers. The neurons are storage cells for scalar values obtained by an activation function applied to the neuron values in the previous layer (Figure 12). In particular, to each neuron in the output and the hidden layers, a vector of weights and a scalar bias are associated, and the value u_k^l stored by the k^{th} neuron in the l^{th} layer can be written in the form (Krenker et al. 2011):

$$u_k^l = \sigma^l\left(\sum_j \omega_{kj}^l u_j^{l-1} + b_k^l\right) \tag{4}$$

where σ^l is the activation function for the l^{th} layer, and ω_{kj}^l and b_k^l are the weights and biases respectively. Typical activation functions used can be linear (i.e. $\sigma(x) = x$), or non-linear e.g. $\sigma(x) = \tanh(x)$. Under mild assumptions on the activation functions, it can be shown that a neural network with even a single hidden layer is a universal function approximator, as given enough neurons, any continuous function on a compact domain can be approximated with arbitrary precision (Csáji, 2001). In case where more than one hidden layer is used (there is a so-called deep neural network), very efficient approximations can be achieved with a relatively small number of network parameters i.e. weights and biases (Huang, 2003).

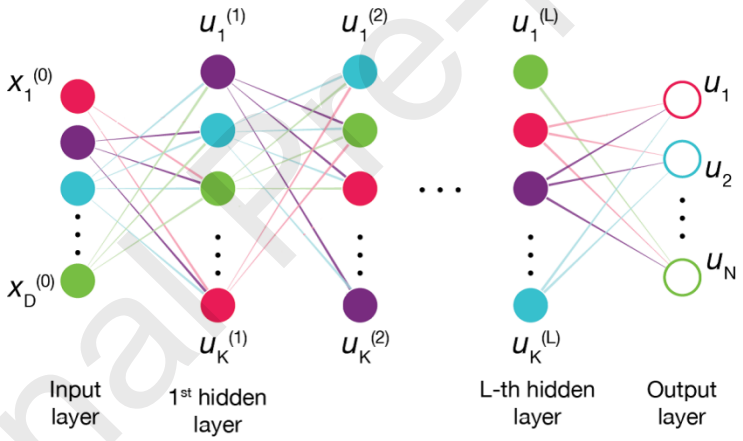


Figure 12: A Sample Artificial Neural Network

To recognize the best structure of the MLP, an initial analysis was carried out to determine the optimal number of layers and neurons in order to perfectly represent the relationship between the input and output of the database. To speed up the optimization procedure, only 25% of the database was considered. Figure 13 presents a comparison of the performance of all the examined structures for one-hidden-layer and two-hidden-layer ANNs, at the end of the training process. The various architecture of ANNs were compared in terms of the RMSE given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \quad (5)$$

In Figure 13, the continuous line describes the result of the one-hidden-layer ANN, which is obviously uniquely determined by the neurons number in the hidden layer itself, and the discontinuous lines show the solutions for the two-hidden-layer ANNs. Each point of this graph has been calculated as the average of ten repetitions of the training process, to guarantee robustness against stochastic impacts. As shown in Figure 13, the two-hidden-layer ANN with 5 and 23 neurons in the first and second layers provides the best performance. Moreover, for a total number of neurons larger than that value, there is no considerable change in the RMSE. Therefore, the 7-5-23-1 ANN architecture was selected.

To prevent overfitting, 70% of the dataset was just used for the training process, 15% was kept for validation, and 15% was used for testing. Therefore, out of the total 644 data, 450 were used for training, 96 for testing, and 96 for validation.

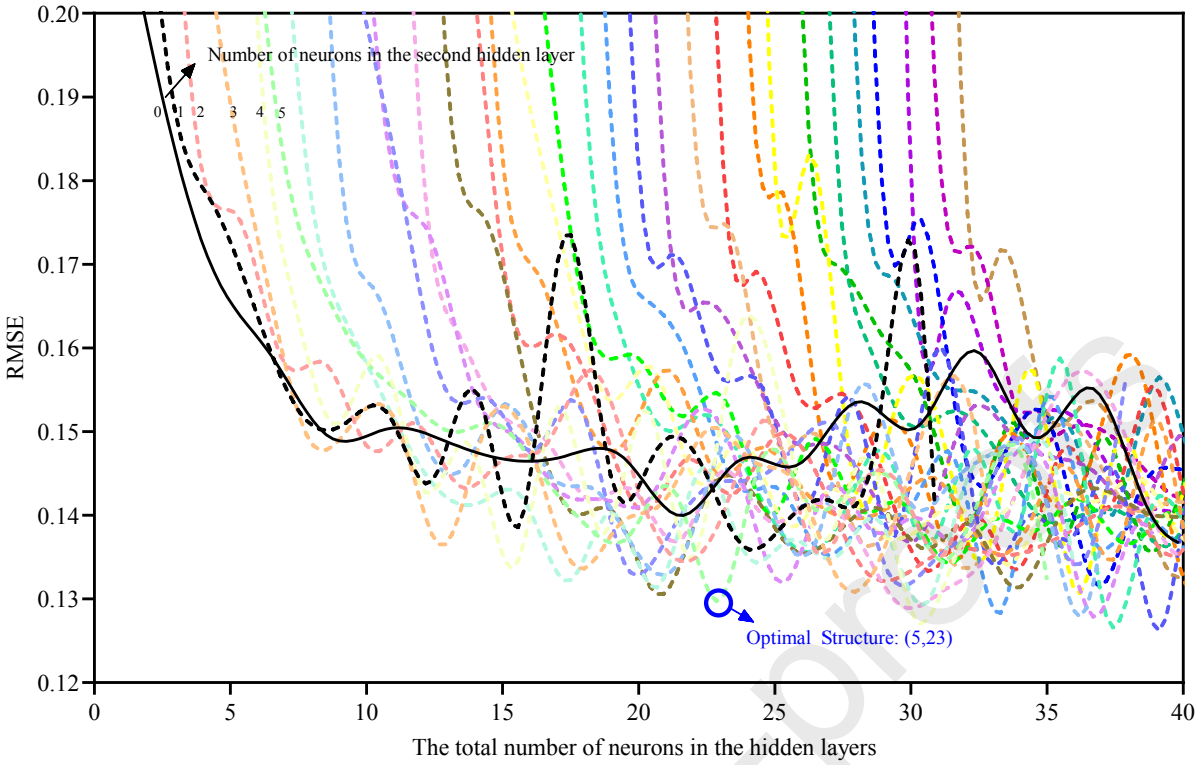


Figure 13. Dependence of the RMSE at the end of training on the total number of neurons in the ANN, featuring either one or two hidden layers.

Figure 14 presents a comparison between the predicted and the existing true AEV of the dataset. It is recognized that although the model exhibits a higher prediction accuracy for the training data, similar accuracy can be recognized for testing and validation data, with the R^2 diverging from 0.851 to 0.964. Overall, satisfying prediction efficiency was proved by the MLP model for the true AEV of soils.

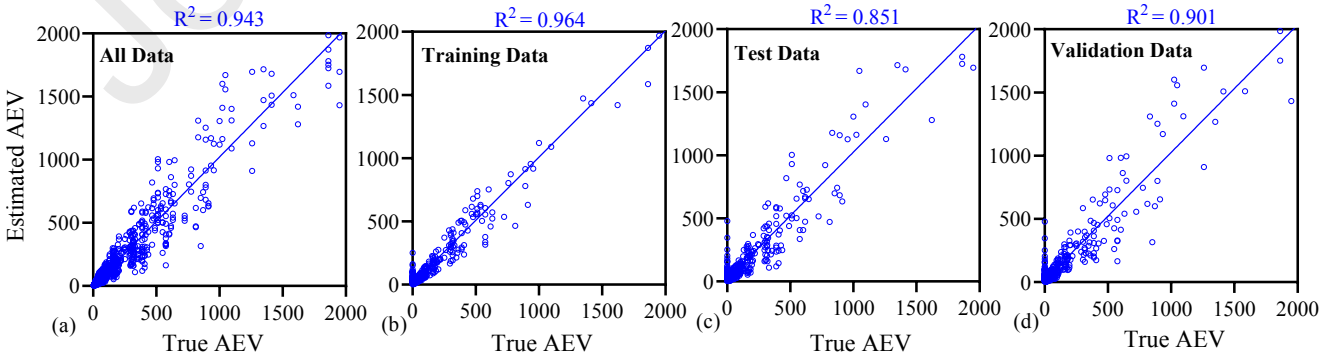


Figure 14. Performances of the ML algorithms: parity plots showing the ML output against the corresponding ground-truth numerical values for (a) all the data; (b) training data; (c) validation data; (d) test data.

Figure 15 presents a comparison between the data obtained from UNSODA and the predicted true AEVs. It is seen that the true AEV estimated by the ML technique matches relatively well with their UNSODA counterparts; the insignificant scattering around them is clearly in accordance with the results described in Figure 14.

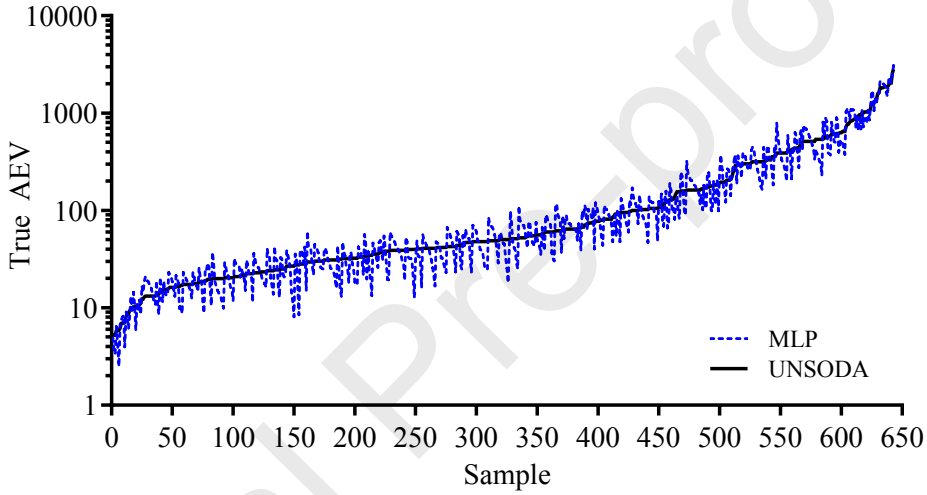


Figure 15. Performance of the MLP algorithm: comparison between the UNSODA data and the MPL algorithm predictions in terms of the true AEV.

Sensitivity analysis

To check the contribution of the input soil property on the model predictions, a sensitivity analysis was performed using the method proposed by Vu-Bac et al. (2016). An uncertainty or sensitivity analysis quantifies the impact of all uncertain input parameters with respect to the specific outputs of interest, which are, in our case, the AEV and true AEV. Therefore, the results from UNSODA have been assessed for this purpose. The sensitivities of the AEV and true AEV for the different independent input variables are presented in Figure 16 confirming that the AEV is most sensitive

respectively to $P_{50\mu}$ and $P_{250\mu}$, and the true AEV is most sensitive respectively to $P_{250\mu}$ and $P_{50\mu}$. The bulk density and $P_{1000\mu}$ are less important input parameters.

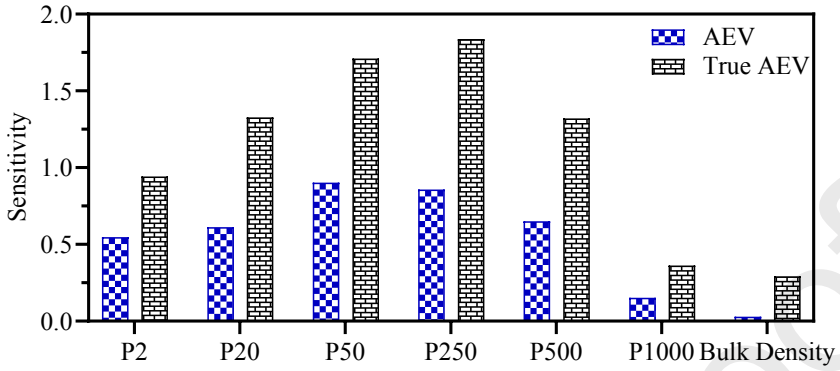


Figure 16. Sensitivity analysis about the relevance of the input variables on the predicted AEVs and true AEVs.

Conclusions

In this paper, pitfalls in the interpretation of water content-based SWRC and estimation of AEV for a significant number of soil samples were investigated. UNSODA database was used and thoroughly examined for this purpose. For the 644 SWRCs, the AEVs were obtained using the conventional method and compared with the corresponding true AEVs obtained using the method proposed by Pasha et al. (2016) and considering the stress history of the samples. The differences between AEVs and true AEVs were found to be generally in the range of 30 – 300 kPa, although this could reach as high as 1500 kPa. A machine learning approach was considered to predict the true AEVs based on the bulk density and grain size distribution as input parameters. The obtained results revealed that the developed ML model operates reasonably well and provides an accurate estimation of the true AEVs with R^2 values of 0.964, 0.901 and 0.851 for the training, validation and testing data, respectively. Furthermore, the sensitivity analysis showed that $P_{250\mu}$ and $P_{50\mu}$ are the most important parameters for AEV estimation. The study shows that methodically trained ML

approaches can be readily used for derivation of true AEVs of a wide range of soils, provided appropriate information regarding the grain size distributions are available.

References

Alonso, E.E., Gens, A., and Josa, A. (1990). A constitutive model for partially saturated soils. Géotechnique, 40(3), 405-430.

Alves, R. D., de FN Gitirana Jr, G., and Vanapalli, S. K. (2020). Advances in the modeling of the soil–water characteristic curve using pore-scale analysis. Computers and Geotechnics, 127, 103766.

Amanabadi, S., Vazirinia, M., Vereecken, H., Vakilian, K.A., and Mohammadi, M.H. (2019). Comparative study of statistical, numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data. Eurasian Soil Science, 52(12), 1555-1571.

Bagheri, M. (2018). Experimental investigation of the time- and rate-dependent behaviour of unsaturated clays. Ph.D. thesis, University of Nottingham.

Bagheri, M., and Rezania, M. (2022). Effect of soil moisture evaporation rate on dynamic measurement of water retention curve with high-capacity tensiometer. International Journal of Geomechanics, 22(3), 04021301.

Bagheri, M., Mousavi Nezhad, M. and Rezania, M. (2019a). A CRS oedometer cell for unsaturated and non-isothermal tests. Geotechnical Testing Journal 43(1): 20-37.

Bagheri, M., Rezania, M. and Mousavi Nezhad, M. (2018). Cavitation in High-Capacity Tensiometers: Effect of Water Reservoir Surface Roughness. Geotechnical Research 5(2): 81-95.

Bagheri, M., Rezania, M. and Mousavi Nezhad, M. (2019b). Rate dependency and stress relaxation of unsaturated clays. International Journal of Geomechanics 19(12): 04019128.

Cheng, Z.L., Zhou, W.H., and Garg, A. (2020). Genetic programming model for estimating soil suction in shallow soil layers in the vicinity of a tree. Engineering Geology, 268, 105506.

Cheng, Z.L., Zhou, W.H., Ding, Z., and Guo, Y.X. (2020). Estimation of spatiotemporal response of rooted soil using a machine learning approach. Journal of Zhejiang University-SCIENCE A, 21(6), 462-477.

Chiu, C.F., Yan, W.M., and Yuen, K.V. (2012). Estimation of water retention curve of granular soils from particle-size distribution—a Bayesian probabilistic approach. Canadian Geotechnical Journal, 49(9), 1024-1035.

Csáji, B.C. (2001). Approximation with artificial neural networks. Faculty of Sciences, Etvs Lornd University, Hungary. 24(48):7.

D'Emilio, A., Aiello, R., Consoli, S., Vanella, D. and Iovino, M., 2018. Artificial neural networks for predicting the water retention curve of Sicilian agricultural soils. Water, 10(10), p.1431.

Fredlund, D.G., and Xing, A. (1994). Equations for the soil-water characteristic curve. Canadian geotechnical journal, 31(4), 521-532.

Huang G.B. (2003). Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Transactions on Neural Networks. 14(2):274-81.

Jain, S.K., Singh, V.P., and Van Genuchten, M.T. (2004). Analysis of soil water retention data using artificial neural networks. Journal of Hydrologic Engineering, 9(5), 415-420.

Javadi, A.A. and Rezania, M. (2009a). Applications of artificial intelligence and data mining techniques in soil modeling. Geomechanics and Engineering, 1(1): 53-74.

Javadi, A.A. and Rezania, M. (2009b). Intelligent finite element method: An evolutionary approach to constitutive modeling. Advanced Engineering Informatics, 23(4): 442-451.

Krenker, A., Bešter, J., and Kos, A. (2011). Introduction to the artificial neural networks. Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech, 1-8.

Leij, F. J., Alves, W. J., van Genuchten, M. Th., Williams, J. R. (1996). The UNSODA unsaturated hydraulic database. USEPA, Cincinnati, Ohio. EPA/600/R-96/095.

Li, Y., and Vanapalli, S.K. (2022). 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.

Moreira de Melo, T., and Pedrollo, O.C. (2015). Artificial neural networks for estimating soil water retention curve using fitted and measured data. Applied and Environmental Soil Science, 2015.

Pagano, A., Tarantino, A., Bagheri, M. et al (2016). An experimental investigation of the independent effect of suction and degree of saturation on very small-strain stiffness of unsaturated sand. E3S Web of Conferences 9: 14015.

Pasha, A.Y., Khoshghalb, A., and Khalili, N. (2016). Pitfalls in interpretation of gravimetric water content–based soil-water characteristic curve for deformable porous media. International Journal of Geomechanics, 16(6), D4015004.

Rezania, M. (2008). Evolutionary polynomial regression based constitutive modelling and incorporation in finite element analysis. Ph.D. thesis, University of Exeter.

Rezania, M. and Javadi, A.A. (2007). A new genetic programming model for predicting the settlement of shallow foundations. Canadian Geotechnical Journal, 44(12): 1462-1473.

Rezania, M., and Ma, G. (2019). Stress-strain modelling of soils in drained and undrained conditions using a multi-model intelligent approach. In International Conference on Inforatmion technology in Geo-Engineering. Springer, pp. 419-428.

Rezania, M., Bagheri, M. and Mousavi Nezhad, M. (2020). Creep and consolidation of a stiff clay under saturated and unsaturated conditions. Canadian Geotechnical Journal 57(5): 728-741.

Russell, A.R., and Khalili, N. (2006). A unified bounding surface plasticity model for unsaturated soils. International Journal for Numerical and Analytical Methods in Geomechanics, 30(3), 181-212.

Saxton, K.E., Rawls, W., Romberger, J.S., and Papendick, R.I. (1986). Estimating generalized soil-water characteristics from texture. Soil science society of America Journal, 50(4), 1031-1036.

Schaap, M.G., and Leij, F.J. (1998). Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil and Tillage Research, 47(1-2), 37-42.

Shi, C., and Wang, Y. (2021). Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. Geoscience Frontiers, 12(1), 339-350.

Tarantino, A., Gallipoli, D., Augarde, C.E., De Gennaro, V., Gomez , R., Laloui, L., Mancuso, C., El Mountassir, G., Munoz, J.J., Pereira, J.-M., Peron, H., Pisoni, G., Romero, E., Raveendiraraj, A., Rojas, J.C., Toll, D.G., Tombolato, S. and Wheeler, S.J. (2011). Benchmark of experimental techniques for measuring and controlling suction. Géotechnique, 61, 303–312.

Vaz, C.M.P., Ferreira, E.J., and Posadas, A.D. (2020). Evaluation of models for fitting soil particle-size distribution using UNSODA and a Brazilian dataset. Geoderma Regional, 21, e00273.

Vu-Bac, N., Lahmer, T., Zhuang, X., Nguyen-Thoi, T., and Rabczuk, T. (2016). A software framework for probabilistic sensitivity analysis for computationally expensive models. Advances in Engineering Software, 100, 19-31.

Walczak S. Artificial neural networks. In Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction 2019 (pp. 40-53). IGI Global.

Wang, H.L., Yin, Z.Y., Zhang, P., and Jin, Y.F. (2020). Straightforward prediction for air-entry value of compacted soils using machine learning algorithms. Engineering Geology, 279, 105911.

Wang, L., Wu, C., Gu, X., Liu, H., Mei, G., and Zhang, W. (2020). Probabilistic stability analysis of earth dam slope under transient seepage using multivariate adaptive regression splines. Bulletin of Engineering Geology and the Environment. 2020 Aug;, 79(6),: 2763-75.

Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., and Gao, L. (2020). Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. Acta Geotechnica,. 2020 Nov;15(11), :3135-50.

Zhai, Q., and Rahardjo, H. (2012). Determination of soil–water characteristic curve variables. Computers and Geotechnics, 42, 37-43

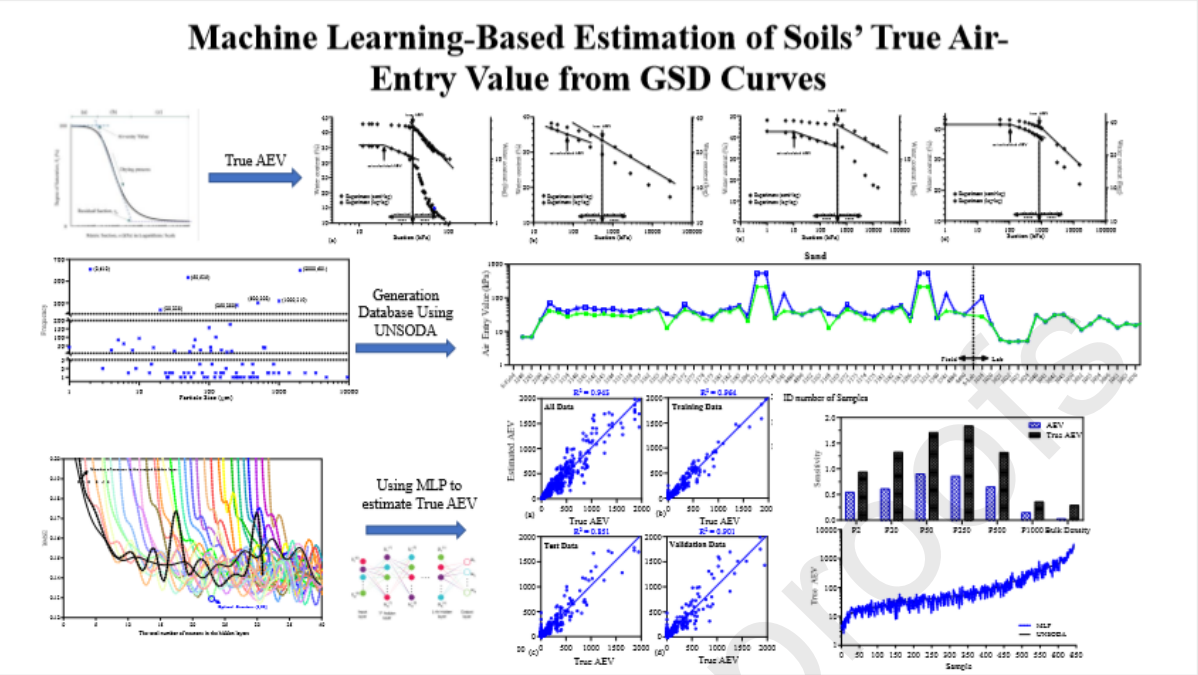
Zhai, Q., Rahardjo, H., Satyanaga, A., and Dai, G. (2020). Estimation of the soil-water characteristic curve from the grain size distribution of coarse-grained soils. Engineering Geology, 267, 105502.

Zhang, P., Wu, H.N., Chen, R. P., and Chan, T.H. (2020). Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study. Tunnelling and Underground Space Technology, 99, 103383.

Zhang, W., Li, H., Han, L., Chen, L., and Wang, L. (2022). Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China. Journal of Rock Mechanics and Geotechnical Engineering,. 2022 Jan 20.

Zhang, W., Wu, C., Zhong, H., Li, Y., and Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geoscience Frontiers,. 2021 Jan 1;12(1), :469-77.

Zhou, M., Shadabfar, M., Huang, H., Leung, Y. F., and Uchida, S. (2020). Meta-modelling of coupled thermo-hydro-mechanical behaviour of hydrate reservoir. Computers and Geotechnics, 128, 103848.



Highlights

- Pitfalls in the interpretation of water content-based SWRC and estimation of AEV are investigated.
- Notable differences between AEVs and true AEVs are obtained for a significant number of soil samples in the large UNSODA database.
- A machine learning model is developed to predict the true AEVs based on the bulk density and grain size distributions (GSDs).
- The developed ML model operates very well and provides an accurate estimation of the true AEVs from GSDs.

CRediT authorship contribution statement

Mohammad Sadegh Es-haghi: Conceptualization, Software, Validation, Investigation, Writing. **Mohammad Rezania:** Conceptualization, Methodology, Validation, Supervision, Writing, Editing. **Meghdad Bagheri:** Conceptualization, Methodology, Validation, Writing, Editing.