



Performance of traditional and machine learning-based transformation models for undrained shear strength

Tim Tapani Länsivaara¹ · Mohammad Sadegh Farhadi¹ · Pijush Samui²

Received: 12 September 2022 / Accepted: 30 December 2022 / Published online: 16 February 2023
© The Author(s) 2023

Abstract

In geotechnical engineering, transformation models are often used as first estimates of parameters and to verify the order of magnitude of field and laboratory tests, which reliability might be constrained by many uncertainties. The undrained shear strength has been for long of particular interest for such models. The traditional transformation models for undrained shear strength are often rather simple. Still, the geotechnical community does not seem to have agreed upon which models to use. In particular, the question of including index properties to the models seems to be open. In the paper, the performance of traditional transformation models is compared to that of machine learning (ML)-based models. In addition, the influence of data coherence is studied by using two datasets of different quality. The ML-based transformation models proved to perform better than traditional ones for both datasets. Clearly, most dominant variables in the transformation model are the preconsolidation pressure and the effective vertical stress. Although including additional variable often may well improve the performance of the training set, the prediction of the testing sets generally tends to worsen, indicating overtraining. The risks for overtraining increase with incoherent data.

Keywords Transformation model · Machine learning · Soft clays · Multivariate database · Undrained shear strength

Introduction

Geotechnical data obtained by field and laboratory investigations is often both complex and incoherent. This can be seen, for example, when trying to find correlations and transformation models between various parameters. Different researchers suggest different correlations where the key regressor varies. Many times, none of them can give very accurate description, while the original data seems rather scattered. Part of the problem is the variation and uncertainty related to geotechnical properties. This is often divided into an inherent (natural) variability and an epistemic uncertainty. The inherent variability represents the natural variation of the property. The epistemic uncertainty is more complex. It includes statistical uncertainty, measurement error, and

transformation model uncertainty. It is then difficult to unveil the true complexity of geotechnical data.

The use of machine learning (ML) has engaged popularity in many fields of application, also in geotechnical engineering (Zhang and Phoon 2022, Zhang et al. 2022). It is tempting to develop ML-based models that describe complicated geotechnical phenomena. The learning ability of ML models makes it an attractive tool to reveal the hidden connections for examples between various soil parameters. However, if the data is incoherent and there exists a lot of variation due to errors in measurements or some other reasons, this becomes a great challenge to the learning process. This might well cause overfitting of the ML model, and even though the statistical parameters might indicate improved correlation, for some combinations the model might give predictions that are completely out of plausible range. Such models might thus give an improved solution for a vast majority of cases, but for the other ones, the result might be quite misleading.

Transformation models for estimating undrained shear strength have been of particular interest in geotechnical engineering for long. Traditional transformation models can be roughly divided into two types, as shown by Eqs. (1) and (2).

Responsible Editor: Zeynal Abiddin Erguler

✉ Tim Tapani Länsivaara
tim.lansivaara@tuni.fi

¹ Research Centre Terra, Tampere University, Tampere, Finland

² National Institute of Technology Patna, Patna, India

$$S_u = \alpha \sigma'_p \quad (1)$$

$$\frac{S_u}{\sigma'_v} = SOCR^m \quad (2)$$

where α , S , and m are either constants or functions of some soil properties. Their values depend on the material and the test type. For example, Skempton (1954) suggested a category (1) equation for the field vane shear strength s_u^{FV} of normally consolidated clays, where the factor α depends on the plasticity index (PI). For Scandinavian clays, Hansbo (Hansbo 1957) suggested the use of liquid limit (LL) instead of PI. Later, Chandler (Chandler 1988) extended Skempton's and Larsson (Larsson 1980) Hansbo's ideas to overconsolidated (OC) clays. Equation (2) was first suggested by Ladd and Foot (Ladd and Foot 1974) and is often referred as the SHANSEP approach.

It has been recognized for long, that the s_u^{FV} should be corrected with respect to plasticity to account for anisotropy and rate effects (Bjerrum 1972; Taylor 2001). This corrected value is often referred to as $s_u(\text{mob})$. It is often assumed that the $s_u(\text{mob})$ value is approximately equal to the undrained shear strength determined by the direct simple shear (DSS) tests, i.e., the s_u^{DSS} value. However, it should be acknowledged that rate effects are present in the DSS as well.

Mesri (Mesri 1975) suggested the use of Eq. (1) for $s_u(\text{mob})$ using a constant value of $\alpha = 0.22$. Jamiolkowski et al. (Jamiolkowski et al. 1985) suggested category (2) equation with $S = 0.23 \pm 0.04$ and $m = 0.8$. However, for s_u^{DSS} , Larsson et al. (Larsson et al. 2007) suggested that S depends linearly on the LL, while Karlsrud and Hernandez-Martinez (Karlsrud and Hernandez-Martinez 2013) proposed the use of the water content (w). Ching and Phoon (Ching and Phoon 2012) also suggested the use of category (2) equation but proposed that S would be dependent on LI, sensitivity (S_t), or both. D'Ignazio et al. (D'Ignazio et al. 2016) concluded that while the s_u^{FV} correlates with the index properties (PI, LL, w , and LI), the $s_u(\text{mob})$ values seem to be unaffected by them.

Based on the above, it seems that while majority of moe that equations of category (re recent research seems to prefer the use of category (2) equations, it is unclear whether index properties should be included when predicting the design value $su(\text{mob})$. It should also be acknowledged that there is now evidence that equations of category (1) or (2) would be the best one to correlate the preconsolidation pressure and in case of (2) also the effective vertical stress, to the undrained shear strength. Machine learning (ML) offers a valuable tool to test the dependencies between undrained shear strength and other soil properties. Recently, Zhang et al. (Zhang and Phoon 2022) studied the use of extreme gradient Boosting (XGBoost) and random forest (RF) learning methods to

study these relationships for s_u^{FV} for a database consisting of Scandinavian clays. In short, they obtained the best performance with the XGBoost and RF models in comparison to three other ML models and two transformation models of category (2) where the value of S was made dependent on either w or LL. They also concluded that the preconsolidation pressure is the most dominant factor, with a relative importance of 54.3% for their XGBoost model, while the value was 20.3% for effective vertical stress, 12.5% for PL, 6.8% for w , and 6.1% for LL. Based on a fivefold cross-validation of the ML models, they concluded that the models are influenced by the data quality.

While the study of Zhang et al. (Zhang and Phoon 2022) gives some important insights on the possibilities of ML models and the importance of various soil properties, it remains unclear which properties to include in predicting the $s_u(\text{mob})$ used in design. In addition, the differences of various proposed models could be partly explained by the quality of data used. This paper attempts to bring more clarification to these questions. The scope of this paper is thus not on finding the best transformation model for engineering purposes but rather to (1) study which combination of soil properties gives the best estimate of $s_u(\text{mob})$, (2) study what is the influence of dataset quality on the models, and (3) study and exemplify the risks of overfitting. It is believed that ML models provide a good tool for such study providing much more freedom in the intervariable correlations than normally achieved by traditional transformation models. Artificial neural networks (ANN) will be used as the main ML method while multivariate adaptive regression spline (MARS) and genetic programming (GP) will be used as comparative ML methods and category (1) and (2) equations will be used for comparisons.

Datasets used in study

Two different datasets are used in the study. Ching and Phoon (Ching and Phoon 2014) proposed labeling a multivariate database as "soil type"/ "number of parameters of interest"/ "number of data points." The first dataset taken from D'Ignazio et al. (D'Ignazio et al. 2016) and labelled as F-CLAY/7/216 contains thus 216 data points each containing 7 parameters for Finnish clays. The dataset can be found on the webpage of ISSMGE technical committee 304 at <http://140.112.12.21/issmge/tc304.htm>. However, the preconsolidation pressures used in this study are all normalized with respect to strain rate to correspond to a typical strain rate applied in CRS oedometer tests ($1 \times 10^{-6} \text{ s}^{-1}$ – $4 \times 10^{-6} \text{ s}^{-1}$), using the same first-order correction as D'Ignazio et al. (D'Ignazio et al. 2016). The clay parameters in the dataset consist of undrained shear strength determined by the field vane (s_u^{FV}), in situ effective vertical

stress (σ'_v), preconsolidation pressure (σ'_p), liquid limit (LL), plastic limit (PL), water content (w), and sensitivity (S_t). In addition, the location and the depth of the data points are known. In this study, the focus is on the design value of undrained shear strength, i.e., on $s_u(\text{mob})$. The correction of s_u^{FV} is done according to Finnish guidelines (Ratahalintokeskus 2005) following the ideas of Bjerrum (Bjerrum 1972) and Helenelund (Helenelund 1977) applying a correction factor dependent of the liquid limit. The basic statistics, including mean, median, minimum, and maximum values, and coefficient of variation (COV) of the datasets are summarized in Table 1. In addition to the parameters listed above, also the statistics for calculated over consolidation ratio ($\text{OCR} = \sigma'_p / \sigma'_v$) and the ratio $s_u(\text{mob}) / \sigma'_p$ is provided.

The second dataset labelled as F-CLAY/7/135 is a subset of the previous dataset. Therein an attempt is made, to remove outliers. It is well known that the undrained shear strength is closely related to the preconsolidation pressure. The changes of these parameters with respect to depth should thus be consistent. As already discussed, Mesri (Mesri 1975) suggested the ratio $s_u(\text{mob}) / \sigma'_p$ to be 0.22 for clays in general. There should thus not be too high variation of the ratio within an otherwise homogenous clay layer, even if the expected value of this ratio is not restricted. Any clear deviations found in F-CLAY/7/216 with respect to the above will be considered as outliers and removed. Some sites also include data from the dry crust layers. As

this data is very limited, it is not considered in the second dataset. The criteria described above includes the use of engineering judgement and involves thus subjectivity. With respect to statistical criteria removing data outside the limits $s_u(\text{mob}) / \sigma'_p = \text{mean} \pm 2\text{COV}$, corresponding to a 95% confidence interval for a normally distributed data, the criteria applied removed all except one data point on the high side. The aim herein is not to produce a “perfect” dataset, but rather to study the influence of data coherence on the performance of various models. The basic statistics for the second datasets are summarized in Table 1. The performance of the criteria is further discussed in the “Discussion” section.

As can be seen from Table 1, the removal of outliers has not considerably narrowed the range of values for the variables, comparing datasets F-CLAY/7/216 and F-CLAY/7/135. For the OCR, the maximum value has reduced to almost half, but the mean and median values are close for the two datasets. With respect to data coherence for the present study, the most interesting variable is the ratio between the undrained shear strength and the preconsolidation pressure. As discussed above, this is sometimes considered a constant. In Fig. 1, the statistics for the ratio $s_u(\text{mob}) / \sigma'_p$ is illustrated for the two datasets.

As can be seen, the dataset F-CLAY/7/135 is much more coherent with respect to the ratio $s_u(\text{mob}) / \sigma'_p$ than F-CLAY/7/216. It is believed that the development of ML-based transformation models in the next sections will reveal

Table 1 Basic statistic for the two datasets

	Variable	Mean	Median	Min	Max	COV
F-CLAY/7/216	Depth	6.72	6.00	0.50	24.0	0.59
	s_u^{FV} (kPa)	21.44	20.00	5.00	75.00	0.50
	σ'_v (kPa)	46.99	42.62	7.50	163.00	0.48
	σ'_p (kPa)	97.33	78.37	25.40	292.10	0.54
	LL (%)	66.28	65.00	22.00	125.00	0.30
	PL (%)	27.74	25.00	10.00	50.00	0.20
	$w(\%)$	76.34	75.00	25.00	150.00	0.27
	S_t	17.45	11.00	2.00	64.00	0.79
	s_{imob} (kPa)	19.31	17.30	3.70	66.20	0.51
	OCR	2.22	1.87	1.00	8.02	0.52
	$s_{\text{imob}} / \sigma'_p$	0.21	0.20	0.06	0.47	0.28
F-CLAY/7/135	Depth	6.92	6.00	1.50	24.0	0.55
	s_u^{FV} (kPa)	21.25	20.00	5.00	55.00	0.43
	σ'_v (kPa)	47.83	43.50	10.10	163.00	0.47
	σ'_p (kPa)	93.91	80.00	25.40	279.40	0.50
	LL (%)	66.83	66.90	22.00	125.00	0.29
	PL (%)	28.01	27.00	10.00	50.00	0.21
	$w(\%)$	78.57	80.00	25.00	150.00	0.26
	S_t	18.15	12.50	2.00	52.30	0.74
	s_{imob} (kPa)	19.16	18.00	3.70	47.80	0.44
	OCR	2.03	1.79	1.11	4.23	0.35
	$s_{\text{imob}} / \sigma'_p$	0.21	0.21	0.15	0.34	0.18

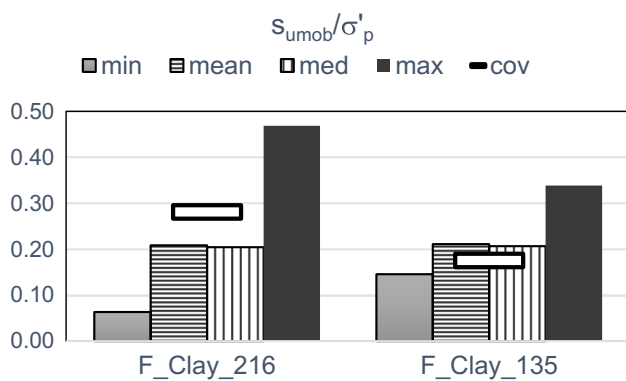


Fig. 1 Ratio between undrained shear strength and preconsolidation pressure for the two datasets

if true outliers were removed, or if the data F-CLAY/7/216 will better show hidden connections between the variables.

Transformation models and ML methods used in study

Statistical parameters used in the development and evaluation of the models

In the development and evaluation of the different models, the following statistical parameters are utilized: bias factor (b), coefficient of variation (COV), standard deviation σ (STD), the Pearson correlation coefficient (R) or its squared value (R^2), and the mean square error (MSE).

The bias b is the mean of the ratio between the actual target value (x) and the predicted value (y), i.e.,

$$b = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{y_i} \tag{3}$$

So, if the bias $b = 1$, the model provides on average an unbiased prediction.

The amount of uncertainty a parameter/property has can be described with the coefficient of variation (COV), which is a dimensionless ratio between the standard deviation σ and the mean value μ of the property, i.e.,

$$COV = \frac{\sigma}{\mu} \tag{4}$$

It shows thus the extent of variability in relation to the mean of the population, the closer the value of COV is to zero, the smaller is the scatter of the data, with $COV = 0$ indicating that the prediction is deterministic.

The standard deviation σ of the entire population of variable (x) is given by

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu_x)^2}{n}} \tag{5}$$

The Pearson correlation coefficient (R) is a measure of the linear correlation between two variables. It is defined as the covariance of the two variables (x, y) divided by the product of their standard deviations, i.e.,

$$R = \frac{COV(x, y)}{\sigma_x \sigma_y} \tag{6}$$

R has values in between $+1$ and -1 , where $+1$ indicates total positive linear correlation, 0 no linear correlation, and -1 total negative linear correlation.

MSE is a risk function that measures the average squared difference between the predicted (y) and the values that are estimated (x). It can be calculated as

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 \tag{7}$$

MSE obtains positive values, the closer to zero the better is the prediction.

Traditional transformation models

As discussed in “Datasets used in study,” the traditional transformation models will be based on categories (1) and (2) types of equations. Equation (1) will be utilized in two different ways. Firstly, α will be fitted to give the best average fit with bias equal to one, and the performance of this model is then evaluated by its COV value. Secondly, α will be evaluated to give perfect match for each data point to study if the resulting α correlates to the classification data. Equation (2) will be firstly evaluated by using the values $S = 0.244$ and $m = 0.763$ proposed by D’Ignazio et al. (D’Ignazio et al. 2016). Secondly, specific correlations for S and m will be studied.

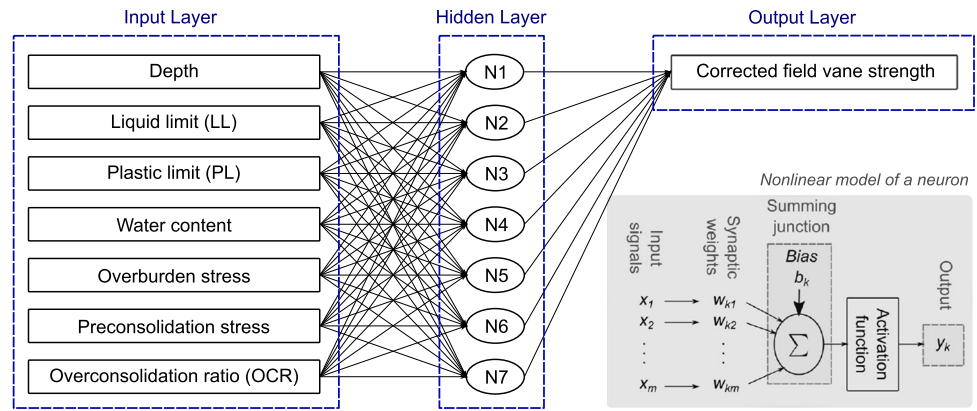
ANN-based transformation models

General

ANN, as a form of artificial intelligence, imitates the behavior of the brain and nervous system (Shahin et al. 2001). The brain is a highly complex, nonlinear, and parallel information-processing system. To perform the necessary computations, like pattern recognition, perception, and body control, it organizes some constituents named as neurons.

A neuron is a fundamental information-processing unit in the operation of a neural network (Haykin 2005). Figure 2 shows a neuron model and the architecture of the ANN model used in this study.

Fig. 2 Architecture of a typical 3-layer feedforward network of ANN (Haykin 2005) for simulation of corrected field vane strength



In the mathematical modeling, the three considered basic elements of the neural model are as below:

1. A set of synapses or connecting links which are modeled by a parameter considered as weight. An input signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj} .
2. A function for summing the weighted input signals for each neuron.
3. An activation function for limiting the amplitude of the output of a neuron. Different activation functions might be utilized, such as sigmoid which is used in this study, and the normalized amplitude of the output of a neuron often ranges in the closed interval $[0, 1]$ or $[-1, 1]$.

An externally applied bias, denoted by b_k , is also included in the neural model. The bias increases or decreases the input of the activation function, depending on whether it is positive or negative, respectively.

In other words, a neuron k is modeled mathematically by the following equations:

$$u_k = \sum_{j=1}^m w_{kj}x_j \tag{8}$$

and

$$y_k = \varphi(u_k + b_k) \tag{9}$$

where x_1, x_2, \dots, x_m are the input signals; $w_{k1}, w_{k2}, \dots, w_{km}$ are the synaptic weights of neuron k ; u_k is the linear summation function; b_k is the bias; φ is the activation function; and y_k is the output signal of the neuron. The use of bias b_k has the effect of applying an affine transformation to the output u_k of the linear combiner in the model of Fig. 2 (Haykin 2005).

Input–output mapping

A popular paradigm of learning called learning with a teacher or supervised learning involves modification of the synaptic weights (and biases) of a neural network by applying a set of training samples. Each example set of samples consists of a unique input signal and a corresponding estimated response (output signal). The network training is performed using some examples picked at random (or selected by user) from the samples sets, and the synaptic weights of the network are modified to minimize the difference between the estimated and the actual values of the desired output parameter(s). The difference between the estimated and actual outputs is computed by means of an appropriate statistical error criterion such as MSE. The training of the network is repeated for many examples in the input–output datasets until the network reaches a steady state, where there are no further significant changes in the synaptic weights.

Multilayer feedforward networks

The arrangement of the neurons of a neural network is intimately linked with the learning algorithm used to train the network and should be structured. As a commonly utilized architecture which is employed in this study also, the multilayer feedforward network, known also as Multilayer Perceptron (MLP), consists of one or more hidden layers, whose computation nodes are called hidden neurons or hidden units. The hidden layer(s)/neurons intervene the network input and output layers.

The nodes in the input layer (the input vector) constitute the input signals applied to the neurons (computation nodes) in the second layer (i.e., the first hidden layer). The output signals of the second layer are used as inputs to the third layer and so on for the rest of the network. The set of

output signals of the neurons in the output (final) layer of the network constitutes the response of the network to the source nodes in the input (first) layer. The three-layer feed-forward ANN architecture used in this study is shown in Fig. 2 (Haykin 2005).

As the learning algorithm, back propagation approach was used so that the variables of the model, i.e., weights and biases, would be optimized. In this algorithm, MSE is propagated back from the output to the connection weights and updates the weights to minimize the prediction error (Foroozesh et al. 2014). The Levenberg–Marquardt (LM) method was used to minimize the error criterion because it is fast, accurate, and reliable (Adamowski and Chan 2011; Adamowski and Karapataki 2010). The LM algorithm is a modification of the classic Newton algorithm for finding an optimum solution to a minimization problem (Adamowski and Karapataki 2010; Daliakopoulos et al. 2005).

The mathematical computations of the ANN modeling are implemented by means of the MATLAB ANN toolbox. The primary objective of the ANN modeling has been to perform the input–output mapping, i.e., to capture the relationships/correlations between the input and output variables with the least error criterion.

GP-based transformation models

Genetic programming (GP) is developed based on the concept of genetic algorithm (Koza 1992). In GP, transformation models are encoded as a set of genes that are then modified using an evolutionary algorithm. Firstly, many equations are created based on the input variables. The fitness of each equation is then determined in the second step. Thereafter, the best equation is selected. New equations are created through the procedure of reproduction, crossover, and mutation. In Fig. 3, the flowchart for the GP procedure is presented.

MARS-based transformation models

Multivariate adaptive regression splines (MARS) is developed based on the concept of nonparametric regression (Friedman 1991) and can be seen as an extension of linear models that automatically model nonlinearities and interactions between variables. It adopts the following relation between input (x) and output (y).

$$y = c_0 + \sum_{m=1}^M c_m B_m(x) \tag{10}$$

where c_0 is constant, $B_m(x)$ is m -th basis function, c_m is the coefficient of m -th basis function, and M is the number of basis functions.

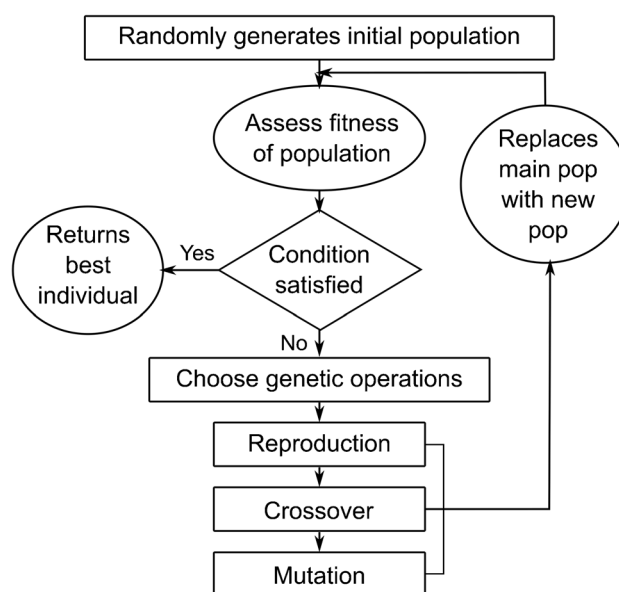


Fig. 3 Flowchart for simple genetic programming (GP) model

The expression of basis function is given below (Sekulic and Kowalski 1992):

$$b_q^-(x - t) = [- (x - t)]_+^q = \begin{cases} (t - x)^q & \text{if } x > t \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Otherwise,

$$b_q^+(x - t) = [+ (x - t)]_+^q = \begin{cases} (x - t)^q & \text{if } x > t \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where q is power, t is the knot location, and $b_q^-(x - t)$ and $b_q^+(x - t)$ are the spline functions. The development of MARS model involves two steps, namely, a forward and a backward step as follows.

Forward step: this step is used to introduce basis function in Eq. (10). MARS can suffer overtraining due to large number of basis functions.

Backward step: for preventing overfitting, extra basis functions have been deleted from Eq. (10). Generalized cross-validation (GCV) value is used to delete the extra basis functions. The expression of GCV is given below:

$$GCV = \frac{\frac{1}{N} \sum_{i=1}^N [y_i - f(x_i)]^2}{\left[1 - \frac{C(B)}{N}\right]^2} \tag{13}$$

where N is the number of data and $C(B)$ is a complexity penalty that increases with the number of BF in the model and which is defined as

$$C(B) = (B + 1) + dB \tag{14}$$

where d is a penalty for each basis function (BF) included into the model. The details about d are given by Friedman (Friedman 1991).

The flowchart for MARS is given in Fig. 4.

Performance of the models

Performance of traditional models

Firstly, the performance of Eq. (1) was evaluated by fitting the value of α to give an unbiased prediction of the undrained shear strength. The obtained values for α and corresponding b between the measured and the predicted COV for the ratio between measured and predicted and R^2 and MSE between the measured and the predicted mobilized undrained shear strength values are presented in Table 2.

In the second round, the value of α was fitted to each data point to give an exact fit. The obtained α values did not show any notable correlation with most of the soil data. For the dataset F-CLAY/7/216, no correlation was found. For the dataset F-CLAY/7/135, some correlation was found for the OCR. The best fit was found using Eq. (15).

$$\alpha = n_1 OCR^{n_2} \tag{15}$$

The obtained values for n_1 and n_2 are 0.2444 and -0.247 , respectively, giving a $COV = 0.446$ and $R^2 = 0.23$.

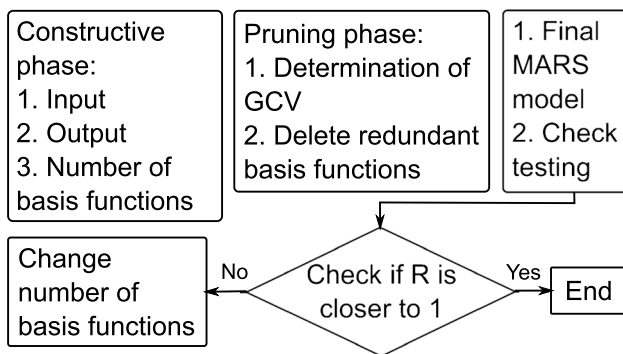


Fig. 4 Flowchart for the MARS model

Introducing Eq. (15) into Eq. (1), gives the following modified version of Eq. (1):

$$S_u = n_1 \sigma_p^i OCR^{n_2} \tag{16}$$

It is worth noticing that this equation is the same equation as (2) and can be rewritten into

$$S_u = n_1 \sigma_v' OCR^{1+n_2} \tag{2b}$$

The performance of Eq. (2b) with the obtained n_1 and n_2 values and corresponding b , COV , R^2 , and MSE values is presented in Table 2 for dataset F-CLAY/7/135.

The performance of Eq. (2) was first evaluated by using the values $S = 0.244$ and $m = 0.763$ obtained by D'Ignazio et al. (D'Ignazio et al. 2016). As can be expected, these values give a very good performance for dataset F-CLAY/7/135, as the data is almost the same and Eq. (2)_135 is almost identical to the previous Eq. (2b)_135; see Table 2. On the other hand, for dataset F-CLAY/7/216, the performance was much poorer. It is interesting to note that for this dataset the inclusion of OCR did not improve the prediction at all. It was also observed that if the preconsolidation pressures of dataset F-CLAY/7/216 had not been normalized, the prediction of Eq. (2) would have been poorer than the one from Eq. (1).

The performance of Eq. (2) was further evaluated by fitting both the value of S and m for each data point to yield the perfect match. However, the obtained S and m values did not show any correlation to any of the data. A special effort was tried out to correlate S to the plasticity index. The reason behind is that for the active undrained shear strength the value of S is likely to increase with the friction angle, as the yield surface gets more inclined. There is no data of the friction angles, but generally, the friction angle is found to correlate with the plasticity index. However, even if the value of S for active undrained strength might increase with the friction angle due to higher inclination of the yield surface, for the same reason, the S value for the passive undrained shear strength should decrease. The field vane value, representing more the average of active, direct shear and passive, is thus likely to be uninfluenced by the friction angle.

Table 2 Performance of various traditional transformation models, b between the measured and the predicted, COV for the ratio between measured and predicted, and R^2 and MSE between the measured and the predicted mobilized undrained shear strength values

Dataset	Equation		b	COV	R^2	MSE
F-CLAY/7/216	(1)_216	$s_u = 0.208 \cdot \sigma_p'$	1.000	0.2803	0.642	45.34
	(2)_216	$s_u = 0.244 \cdot \sigma_v' OCR^{0.763}$	1.010	0.2995	0.606	43.13
	(2)*_216	$s_u = 0.213 \cdot \sigma_v' OCR^{0.97}$	1.000	0.2797	0.640	44.60
F-CLAY/7/135	(1)_135	$s_u = 0.211 \cdot \sigma_p'$	1.000	0.1748	0.865	14.05
	(2b)_135	$s_u = 0.244 \cdot \sigma_v' OCR^{0.753}$	1.013	0.1531	0.879	9.78
	(2)_135	$s_u = 0.244 \cdot \sigma_v' OCR^{0.763}$	1.007	0.1532	0.880	9.96
	(2)*_135	$s_u = 0.25 \cdot \sigma_v' OCR^{0.735}$	1.000	0.1533	0.880	10.15

As any attempt to further improve the performance of Eq. (2) did not yield any notable improvement, an unbiased parameter set was finally introduced as set (2)* in Table 2. The COV values are shown with 4 digits just to show the infinite small differences between models (2b)_135, (2)_135, and (2)*_135. In practice, their performance is equally good. It is worth noting that for dataset F-CLAY/7/216, the best fit of Eqs. (1) and (2)* gives practically the same result. Their similarities are also shown in the value of parameter m for (2)*_216 being close to one. The predictions of Eqs. (1) and (2)* are shown in Fig. 5 together with a range of ± 2 COV.

Performance of ANN models

Based on the available databases, different scenarios of input–output mappings were performed, and their parameter estimation precisions were calculated. In all scenarios, the mobilized vane shear strength was considered as the output, but the input parameters, in situ vertical effective stress (σ'_v), preconsolidation pressure (σ'_p), overconsolidation ratio (OCR), liquid limit (LL), plastic limit (PL), water content (w), depth, and sensitivity (S_t) were arranged in different

scenarios such that the estimation accuracy of the ANN models and GP, MARS, and some traditional empirical equations could be compared. As another target of organizing different scenarios, a sensitivity analysis was performed among the input variables to identify the least number of soil properties that might lead to the best estimation of the modified vane shear strength of soil. Altogether 16 different scenarios were verified. The considered input parameter(s) in each scenario and the statistics of their performance are presented in Table 3.

Based on the observations from the previous studies, regarding the complexity of the models, usually 3 to 10 neurons lead to the best results (Foroozesh et al. 2014; Karaagac et al. 2009; Khosrojerdi et al. 2016). So, in this study, 7 neurons were considered in all ANN models so that the results of different models might be compared. For the training and verification, the data was divided into training (70%), validation (15%), and testing (15%) datasets. The same 70% training dataset was later used for training of the GP and MARS models, while testing of them was done for the rest of the data (30%). The validation dataset is used for the generalization of the ANN model in the training process.

Fig. 5 Performance of Eq. (1) for dataset F-CLAY/7/216 (a) and F-CLAY/7/135 (c) and Eq. (2) for F-CLAY/7/216 (b) and F-CLAY/7/135 (d). The dotted lines represent the range of predictions corresponding to ± 2 COV

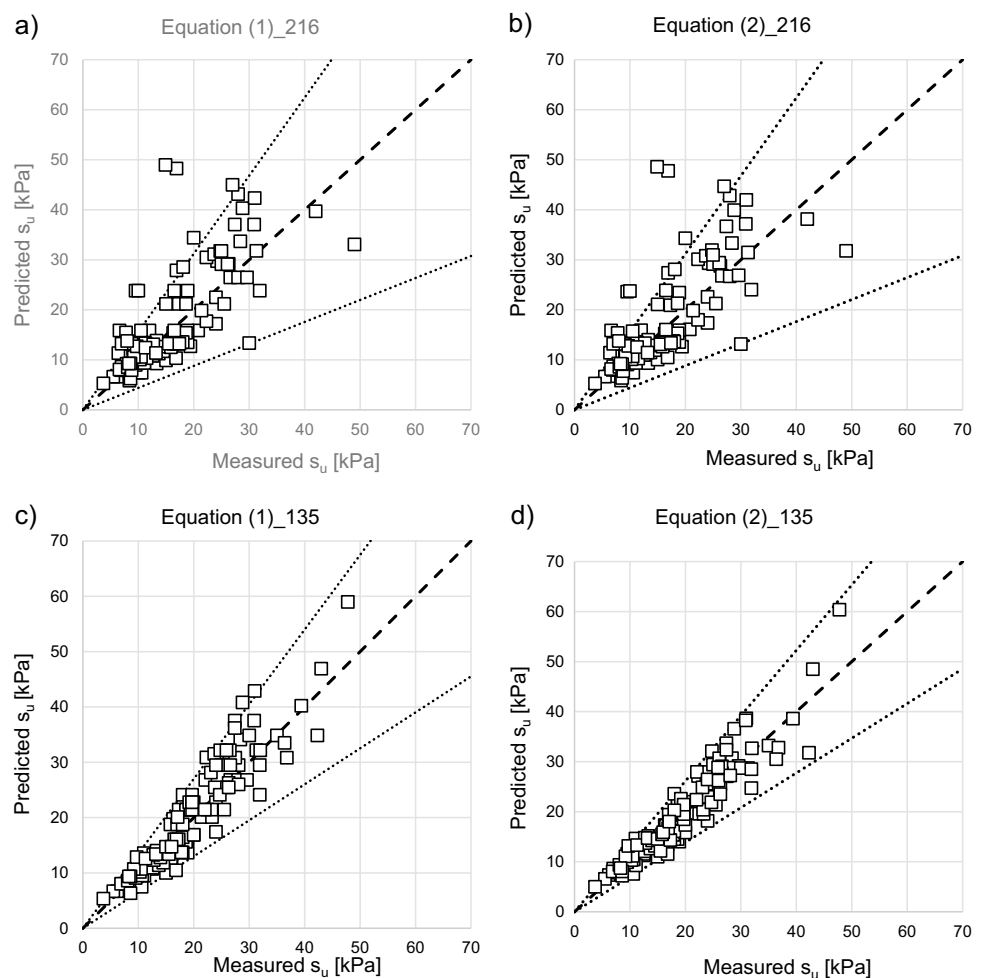


Table 3 Results of ANN analysis with various trials for databases of F-CLAY/7/216 and F-CLAY/7/135

Trial	Parameters to train ANN	Data	F-CLAY/216 database				F-CLAY/135 database			
			<i>b</i>	COV	<i>R</i> ²	MSE	<i>b</i>	COV	<i>R</i> ²	MSE
1	σ'_v	All	1.003	0.475	0.352	63.289	0.996	0.279	0.664	23.932
		Train (70%)	0.998	0.507	0.314	69.152	1.022	0.277	0.664	25.611
		Validation (15%)	0.933	0.467	0.268	54.101	0.920	0.255	0.789	14.232
		Test (15%)	1.104	0.304	0.532	45.094	0.953	0.286	0.491	26.228
2	σ'_p	All	1.012	0.248	0.764	23.196	1.005	0.153	0.892	7.714
		Train (70%)	1.000	0.255	0.755	24.638	1.000	0.142	0.891	8.233
		Validation (15%)	0.986	0.210	0.824	12.686	1.003	0.155	0.902	6.720
		Test (15%)	1.097	0.233	0.750	27.231	1.032	0.193	0.876	6.321
3	σ'_v, σ'_p	All	1.012	0.234	0.841	15.520	0.995	0.139	0.924	5.441
		Train (70%)	1.002	0.233	0.859	14.162	1.000	0.128	0.930	5.313
		Validation (15%)	0.997	0.206	0.839	11.876	0.963	0.150	0.938	3.979
		Test (15%)	1.077	0.250	0.729	25.686	1.004	0.166	0.848	7.583
4	σ'_v, OCR	All	1.009	0.235	0.831	16.585	0.995	0.142	0.927	5.169
		Train (70%)	0.996	0.235	0.845	15.597	1.002	0.128	0.934	5.012
		Validation (15%)	0.975	0.198	0.859	11.747	0.964	0.170	0.907	5.899
		Test (15%)	1.108	0.243	0.750	26.233	0.991	0.169	0.901	5.138
5	$\sigma'_v, \sigma'_p, \text{OCR}$	All	1.011	0.233	0.839	15.798	1.004	0.142	0.932	4.841
		Train (70%)	1.001	0.232	0.852	14.881	1.001	0.118	0.938	4.713
		Validation (15%)	0.984	0.197	0.878	9.358	1.012	0.203	0.938	3.976
		Test (15%)	1.090	0.253	0.731	26.767	1.012	0.163	0.874	6.353
6	$\sigma'_v, \text{OCR, LL}$	All	1.009	0.218	0.881	11.641	1.002	0.131	0.931	4.914
		Train (70%)	1.002	0.199	0.914	8.702	1.000	0.121	0.938	4.685
		Validation (15%)	0.971	0.228	0.846	14.692	0.980	0.129	0.932	4.298
		Test (15%)	1.081	0.264	0.767	22.364	1.033	0.165	0.870	6.634
7	$\sigma'_v, \text{OCR, LL, PL}$	All	1.009	0.244	0.844	15.220	1.010	0.154	0.891	7.771
		Train (70%)	0.999	0.221	0.872	12.853	1.002	0.143	0.898	7.728
		Validation (15%)	0.969	0.210	0.846	12.294	1.005	0.183	0.890	7.022
		Test (15%)	1.099	0.324	0.689	29.406	1.054	0.163	0.836	8.758
8	$\sigma'_v, \sigma'_p, \text{OCR, LL, PL}$	All	1.012	0.253	0.861	13.545	1.055	0.215	0.900	7.609
		Train (70%)	1.006	0.238	0.906	9.514	1.048	0.134	0.920	6.887
		Validation (15%)	0.971	0.267	0.795	17.946	1.077	0.401	0.883	7.393
		Test (15%)	1.082	0.289	0.700	28.024	1.068	0.224	0.789	11.233
9	$\sigma'_v, \text{OCR, LL, PL, } \omega, \text{ depth}$	All	1.052	0.215	0.885	11.680	1.231	1.929	0.912	6.409
		Train (70%)	1.044	0.184	0.922	8.216	1.004	0.116	0.935	4.946
		Validation (15%)	1.001	0.259	0.845	13.958	2.363	2.490	0.870	9.979
		Test (15%)	1.142	0.266	0.768	25.674	1.111	0.222	0.839	9.537
10	$\sigma'_v, \sigma'_p, \text{OCR, LL, PL, } \omega, \text{ depth}$	All	1.031	0.220	0.861	13.889	0.987	0.331	0.899	7.265
		Train (70%)	1.024	0.208	0.887	11.601	0.992	0.122	0.932	5.166
		Validation (15%)	0.985	0.214	0.842	12.566	0.867	0.842	0.830	12.491
		Test (15%)	1.110	0.251	0.758	26.049	1.088	0.239	0.795	11.641
11	$\sigma'_v, S_r, \text{OCR}$	All	1.010	0.227	0.843	15.383	0.998	0.141	0.920	5.687
		Train (70%)	1.000	0.216	0.869	13.237	1.003	0.126	0.930	5.322
		Validation (15%)	0.967	0.232	0.814	15.165	0.976	0.169	0.917	5.314
		Test (15%)	1.101	0.242	0.753	25.730	1.002	0.168	0.844	7.794
12	$\sigma'_v, \omega, \text{OCR}$	All	1.004	0.213	0.871	12.608	0.996	0.144	0.925	5.363
		Train (70%)	1.002	0.202	0.886	11.436	0.997	0.126	0.933	5.093
		Validation (15%)	0.954	0.206	0.883	11.595	1.007	0.194	0.902	6.095
		Test (15%)	1.066	0.243	0.796	19.178	0.980	0.157	0.885	5.863

Table 3 (continued)

Trial	Parameters to train ANN	Data	F-CLAY/216 database				F-CLAY/135 database			
			<i>b</i>	COV	R^2	MSE	<i>b</i>	COV	R^2	MSE
13	$\sigma'_v, LL, \omega, S_p, OCR$	All	1.026	0.206	0.866	13.359	0.997	0.110	0.952	3.424
		Train (70%)	1.023	0.188	0.901	10.230	1.005	0.071	0.977	1.754
		Validation (15%)	0.971	0.230	0.860	11.847	0.966	0.196	0.903	7.034
		Test (15%)	1.093	0.237	0.708	29.682	0.993	0.138	0.851	7.484
14	Depth, σ'_v, OCR	All	1.017	0.232	0.839	15.974	0.998	0.124	0.934	4.695
		Train (70%)	1.005	0.213	0.869	13.656	1.003	0.117	0.941	4.524
		Validation (15%)	0.981	0.224	0.833	12.963	1.000	0.129	0.940	3.954
		Test (15%)	1.112	0.284	0.698	30.021	0.973	0.145	0.880	6.275
15	Depth, $\sigma'_v, LL, \omega, S_p, OCR$	All	1.038	0.220	0.893	10.631	1.022	0.129	0.931	5.142
		Train (70%)	1.012	0.161	0.941	5.979	1.023	0.127	0.942	4.695
		Validation (15%)	1.086	0.269	0.838	13.590	1.001	0.111	0.940	3.913
		Test (15%)	1.111	0.322	0.700	29.529	1.040	0.148	0.845	8.530
16	Depth, σ'_v, ω, OCR	All	1.009	0.205	0.887	11.105	1.019	0.136	0.944	4.023
		Train (70%)	1.011	0.198	0.910	9.147	1.011	0.115	0.954	3.537
		Validation (15%)	0.941	0.228	0.850	13.554	1.040	0.200	0.932	5.534
		Test (15%)	1.069	0.194	0.816	17.817	1.037	0.134	0.912	4.722

However, the validation and test dataset error criteria of the ANN model are compared with the test dataset for the GP and MARS models, mainly because of limited amount of data, especially for F-CLAY/7/135. Traditionally, a single training is performed for the ANN models, accepting the results as such. However, due to random initial guess of the optimization model in training of the ANN, repeating training with the same data will yield different results. To study more this aspect, training was done 1000 times for each trial. The differences between the trials will be discussed later. For Table 3, the training sets giving the lowest weighted sum for MSE_{train}, MSE_{validation}, and MSE_{test}, with equal weights for each, were chosen. In Fig. 6, the COV, R^2 , and MSE values are presented for the different trials for the training and test data.

Studying the data in Table 3 and Fig. 6, several interesting conclusions can be made. Trial 2 seems to give a relatively good prediction using only the preconsolidation pressure as input. All statistical indicators for trial 2 are better than for the traditional transformation model using Eq. (1), indicating that the relationship is not linear. Adding information about the vertical effective stress, i.e., comparing trials 3, 4, and 5 to trial 2, improves the overall performance for both data. However, for the more scattered data F-CLAY/7/216, no improvement is found for the testing data. The different setups using preconsolidation stress and effective vertical stress, i.e., trials 3, 4, and 5, give in practice equal performance. It should though be noted that the results were received from different ANN training sets and relying on a single training would have

indicated differences in between the trials. For trials 3, 4, and 5, the statistical measures are again much better than for the traditional transformation model, indicating that the relationship is not as given by Eq. (2).

Adding information about plasticity, i.e., trials 6, 7, and 8, clearly improves the performance of the training dataset for F-CLAY/7/216. However, for the testing data, the performance drops, except for trial 6, which shows a slight improvement. For the more coherent dataset F-CLAY/7/135, no improvement can be found. However, also for this dataset, some of the training sets gave a clear improvement for the training sets of trials 6, 7, and 8, but then, the testing data gave very poor results. Adding information about water content and depth to previous (trials 9 and 10) does not improve the performance.

Comparing trials 4, 11, 12, 14, and 16 reveals that adding sensitivity data, i.e., trial 11, does not give any improvement. However, the inclusion of water content in trial 12 improves the performance for dataset F-CLAY/7/216, while no improvement is found for F-CLAY/7/135. The inclusion of information about depth (trial 14) improves the performance of training and validation, while the testing performance gets worse. The inclusion of both water content and depth information (trial 16) improves the overall performance of both datasets and seems to give the best performance of all trials. However, in practice, the improvement to trial 4 is rather insignificant. It is also interesting to note that the inclusion of information about plasticity, sensitivity, water content, and depth (trial 15) gives the best performance of the training

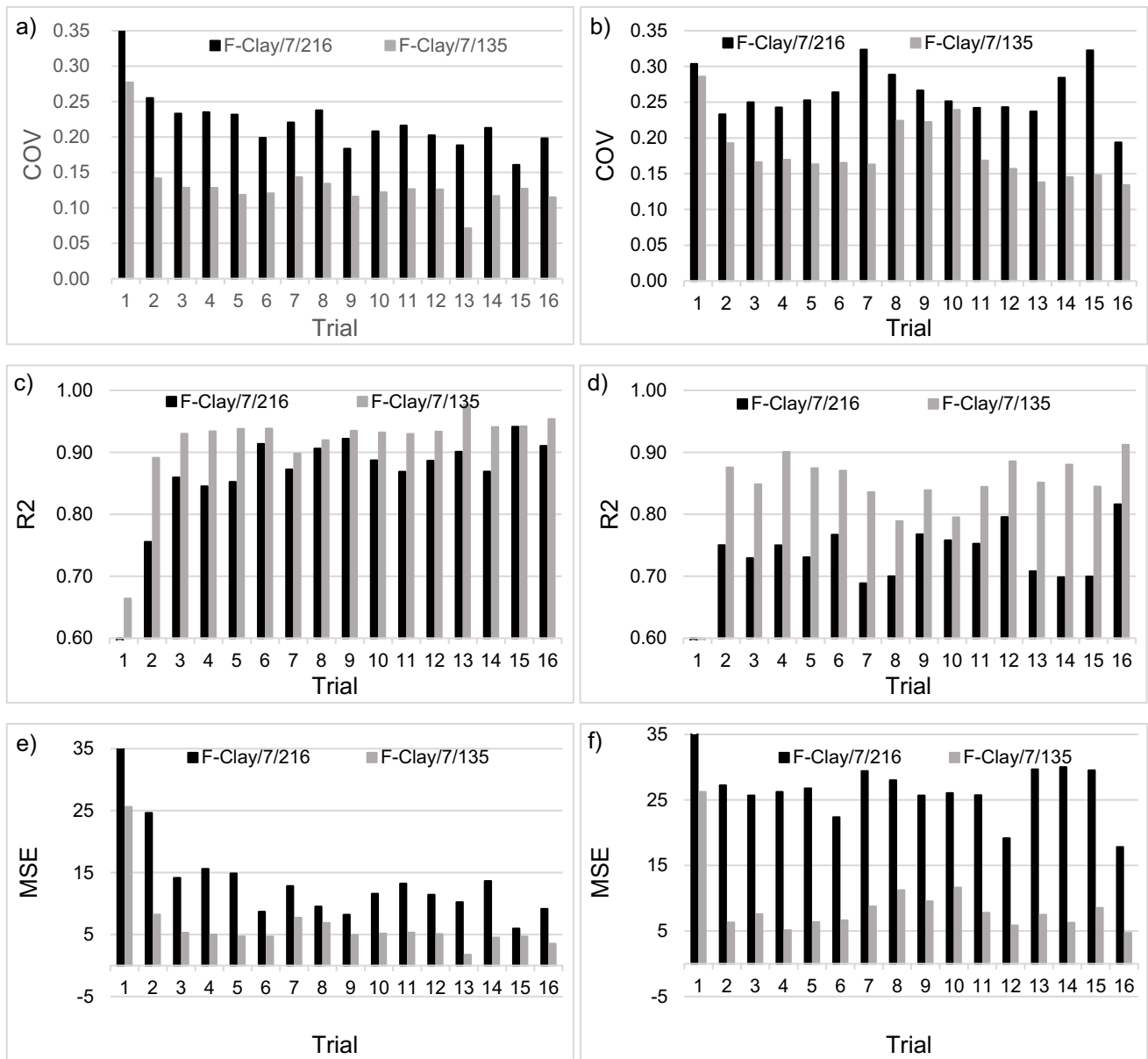


Fig. 6 Performance of ANN trials considering training data (a, c, e) (left) and test data (b, d, f) (right) for databases of F-CLAY/7/216 and F-CLAY/7/135

set for F-CLAY/7/216, while the performance of testing set is not so good.

The performance of models can be visualized by Taylor diagrams (Taylor 2001). They are often used to compare climate models but can well be used to compare how any model represents reality. In the Taylor diagram, three statistical parameters are used and plotted in the same diagram, namely, the Pearson correlation coefficient (R), the centered root-mean-square error (RMSE, i.e., the root of MSE), and the standard deviation (STD). Because of their mathematical relation, these parameters can be plotted together in a quadrant in the following way. The observation, i.e., the

parameter that is modelled, in our case s_{umob} , is plotted on the horizontal axis, at a distance of STD from the origin. The horizontal axis represents $R = 1$, while other R values are related to the angle to the horizontal axis, i.e., the azimuthal angle. The model points are plotted at a distance of their STD from the origin on a line which angle corresponds to the R value. Then, the centered RMSE value will be related to the distance of this point to the observation. As a centered RMSE value is used, i.e., the means are subtracted out before computing their values, the diagram does not provide information about overall biases, but solely characterizes the centered pattern error (Taylor 2001).

As the actual value of STD is not important when comparing the models, a normalized plot is often used. Therein, both RMSE and STD values are normalized with the STD of the observation. Then, the observation point is at unit distance from the origin. Herein, such a normalized plot is used to compare different datasets.

In Fig. 7, the Taylor diagram representation is given for trials 4, 15, and 16 for both datasets F-CLAY/7/216 and F-CLAY/7/135. In addition, trial 15 is given for an ANN training set, giving the best performance for the training data out of the 1000 training sets. In all cases, the performance for training, validation, and testing is given separately, using white fill for training, grey for validation, and black for testing. Trials 4 and 16 give a relatively consistent performance, while the points for training, validation, and testing are relatively closely located, trial 16 showing the best performance. Trial 15 is giving a good performance for the training data, but it declines more for validation and training data, indicating overtraining. The implications of overtraining are clearly shown for the case trial 15 max train, showing the best performance for training data, but worst performance for testing data.

The differences between trials 4 and 16 are further studied by Fig. 8 comparing the predicted data to the observed. Again, white fill is used for training, grey for validation, and black for testing. The dotted lines in both trials 4 and 16 represent $\pm 2^{\text{COVAIL_trial 4}}$.

As can be seen also from this presentation, the difference between these two trials is minimal. However, it can though

be seen that including data on water content and depth slightly improves the performance especially for smaller values. Generally, the value of depth is not related to the strength, as we may have highly overconsolidated clays at shallow depths with high undrained shear strength and soft normally consolidated clays at deeper depths. However, as the dataset consisted of clays with generally low OCR, it is logical that the depth correlates with the strength. As information of water content and depth is available if oedometer data is used, adding them will not require any extra effort. It would though not be advisable to use such model for highly overconsolidated clays, predicting, e.g., the strength of dry crust. For practical engineering, the difference between trials 4 and 16 is insignificant.

Performance of GP models

Various combinations of input parameters were also tried for the GP models, but not as extensively as for the ANN. Interestingly, for F-CLAY/7/216, the best performances were achieved by including data of σ'_v , σ'_p , and OCR, while for F-CLAY/7/135 including LL improved slightly the performance of the testing set, while the performance of training was slightly declined. It is good to note that also for GP the outcome will vary for repetitive trainings and a similar 1000 training trials as for ANN were not tried out. As no clear improvement was found for additional parameters, models including data of σ'_v , σ'_p , and OCR are used for comparisons.

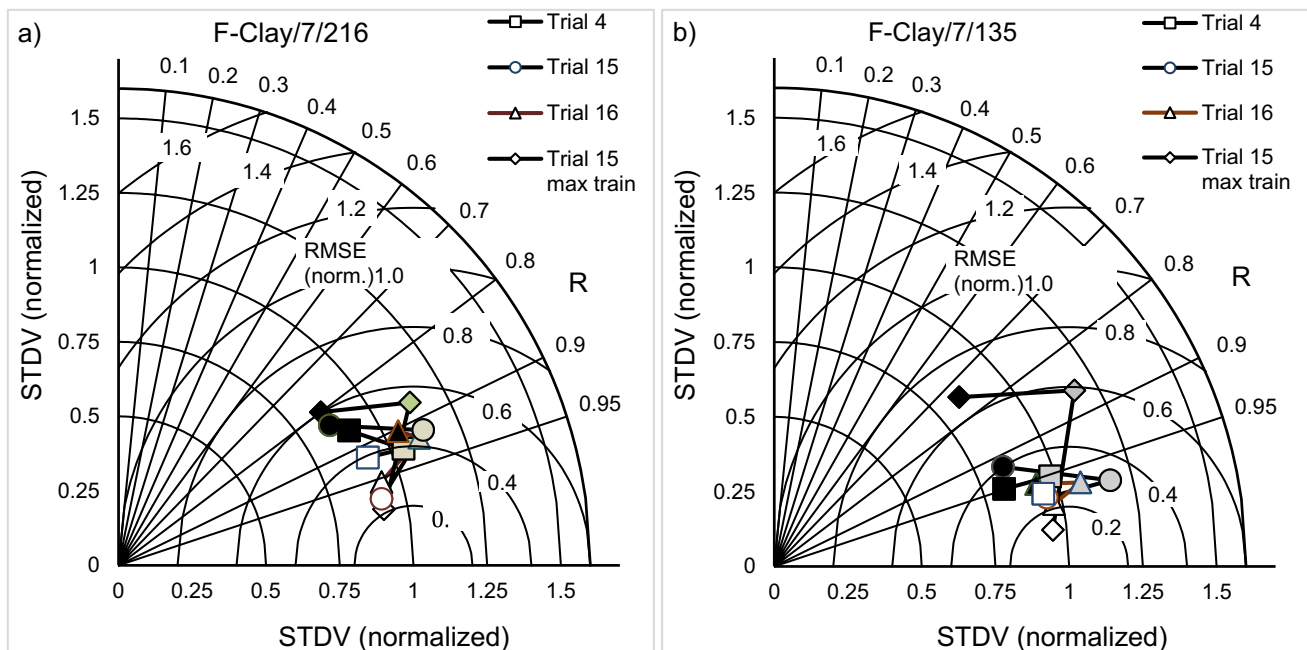
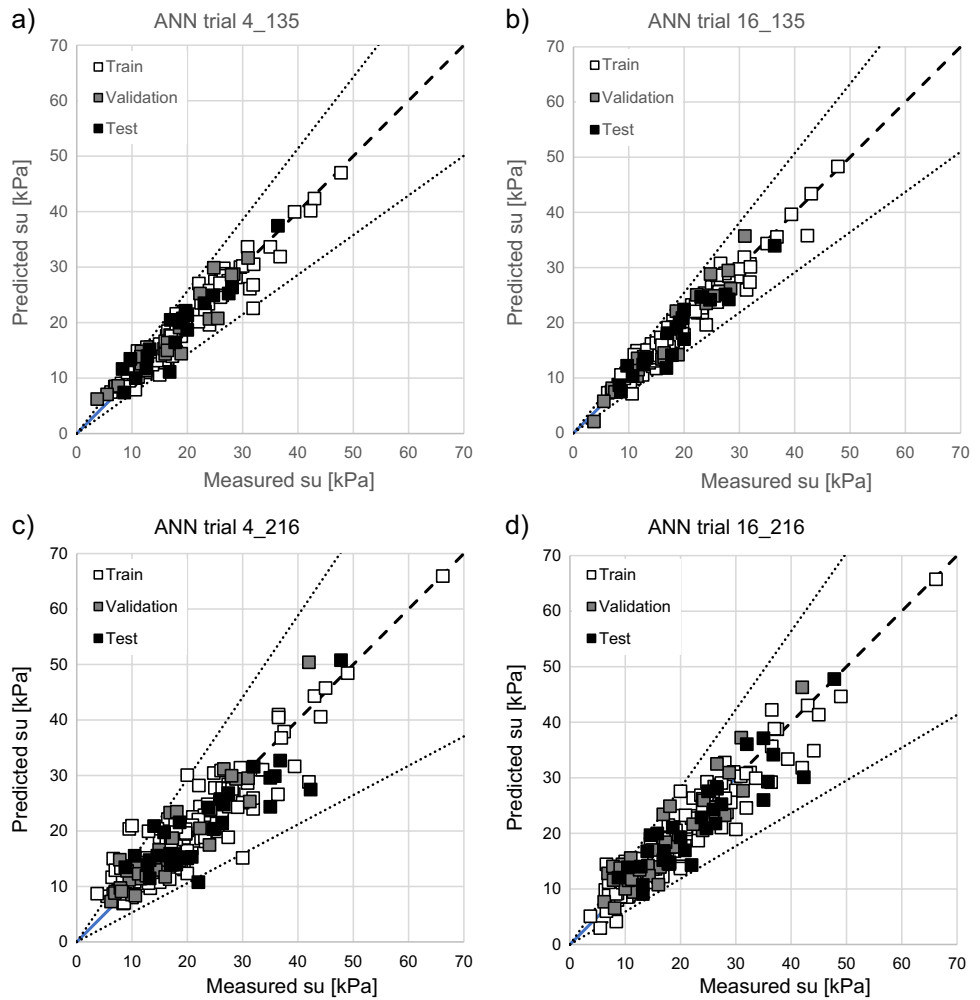


Fig. 7 Taylor diagrams for trials 4, 15, and 16 for databases F-CLAY/7/216 (a) and F-CLAY/7/135 (b). White, grey, and black fills show training, validation, and testing, respectively

Fig. 8 Performance of ANN trials 4 (a) and 16 (b) for F-CLAY/7/135 and trials 4 (c) and 16 (d) for F-CLAY/7/216. In addition to predicted vs. measured data, a 1:1 line and the lines for ± 2 COV are given



The performance of the GP models for F-CLAY/7/216 and F-CLAY/7/135 using σ'_v , σ'_p , and OCR as input is given in Fig. 9, as the ratio of measured vs. predicted undrained shear strength including lines indicating ± 2 COV. The

Fig. 9 Performance of GP model for datasets F-CLAY/7/216 (a) and F-CLAY/7/135 (b). In addition to predicted vs. measured data, a 1:1 line and the lines for ± 2 COV are given

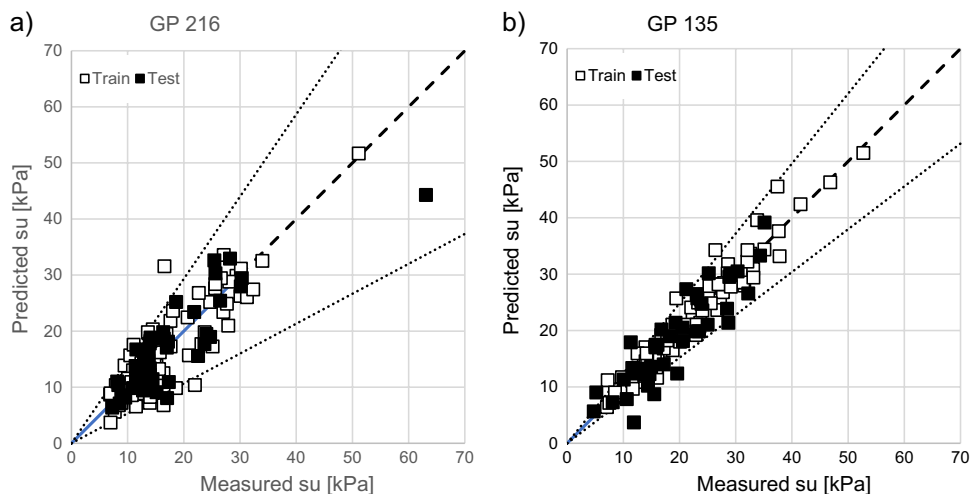


Table 4 Performance of various AI transformation models

Dataset	Method		Bias	COV	R^2	MSE	
F-CLAY/7/216	ANN case 16	Train (70%)	1.009	0.205	0.887	11.105	
		Validation (15%)	1.011	0.198	0.910	9.147	
		Test (15%)	0.941	0.228	0.850	13.554	
		All	1.069	0.194	0.816	17.817	
	GP	Train (70%)	1.002	0.233	0.849	17.082	
		Test (30%)	1.055	0.261	0.701	9.523	
		All	1.018	0.244	0.797	14.807	
	MARS	Train (70%)	1.004	0.245	0.818	20.550	
		Test (30%)	1.031	0.262	0.667	39.583	
		All	0.997	0.241	0.781	23.875	
	F-CLAY/7/135	ANN case 16	Train (70%)	1.019	0.136	0.944	4.023
			Validation (15%)	1.011	0.115	0.954	3.537
Test (15%)			1.040	0.200	0.932	5.534	
All			1.037	0.134	0.912	4.722	
GP		Train (70%)	1.001	0.120	0.941	5.287	
		Test (30%)	0.986	0.248	0.805	6.608	
		All	0.996	0.169	0.908	5.688	
MARS		Train (70%)	1.000	0.138	0.908	8.196	
		Test (30%)	0.977	0.197	0.846	10.304	
		All	0.993	0.158	0.894	8.836	

statistical performance of the models evaluated by their bias, COV, R^2 , and MSE is presented in Table 4.

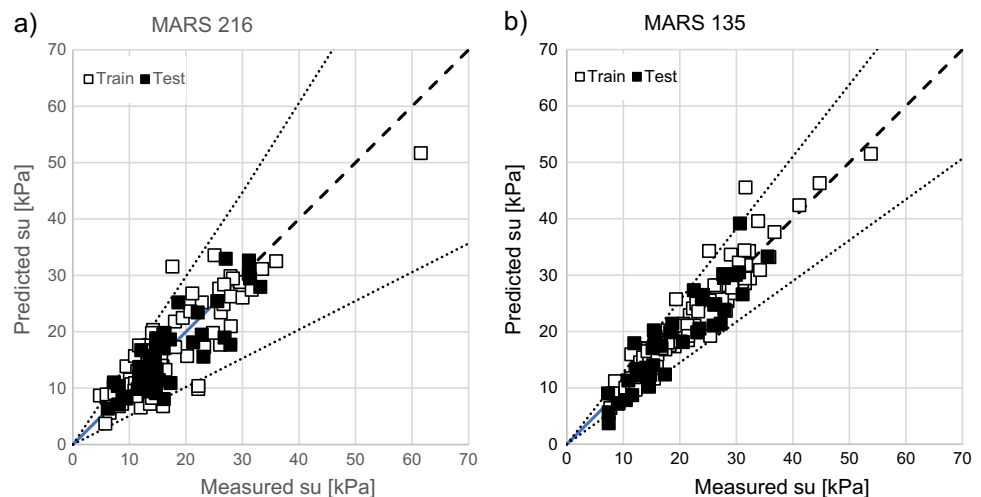
Performance of MARS models

Similarly, as for GP models, different combinations of input parameters were also tried for the MARS models, but not as extensively as for the ANN. Similarly, for MARS models, the best performance was not achieved by including all parameters. In general, it can be reported that the most consistent performance was achieved with a combination including data of σ'_v , σ'_p , and OCR. For F-CLAY/7/216, the

inclusion of LL improved the performance of training set, while for the training set, COV and R values were slightly worse and MSE was slightly better. The differences though were minor. For F-CLAY/7/135, the inclusion of LL and PL data improved the performance of the training set, while the performance of testing set and overall performance dropped. Again, the differences are rather small. Unlike for ANN and GP models, MARS yields the same results on repeated trainings.

The performance of the final MARS models for F-CLAY/7/216 and F-CLAY/7/135 using σ'_v , σ'_p , and OCR as input is given in Fig. 10 as the ratio of measured

Fig. 10 Performance of MARS model for datasets F-CLAY/7/216 (a) and F-CLAY/7/135 (b). In addition to predicted vs. measured data, a 1:1 line and the lines for ± 2 COV for training set are given



to predicted undrained shear strength including lines indicating ± 2 COV. The statistical performance of the models evaluated by their bias, COV, R^2 , and MSE is presented in Table 4.

Discussion

In the development of the ML models, the data was divided into a training (70%) and test (30%) data for MARS and GP methods and to training (70%), validation (15%), and test (15%) data for the ANN. As the division to the different sets was done randomly, it might result in that all relevant data is not covered in the training set. In Table 5, the mean, min, and max values of the training datasets are given with respect to their values for the complete datasets. A min value of 1.24 means thus that for training the smallest value was 1.24 times larger than the minimum value of the whole dataset. As can be seen from Table 5, the training dataset for F-CLAY/7/216 covered almost the entire range of the various parameters with a mean very close to the whole dataset.

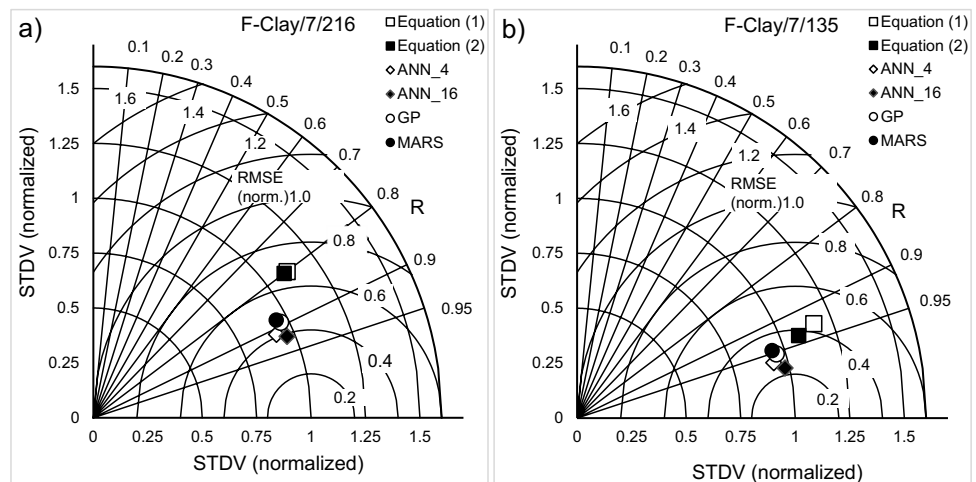
For dataset F-CLAY/7/135, the training data did not cover as well the range of values for all data. Still, the models performed much better than for F-CLAY/7/216 dataset.

As an overview of the performance of different types of models, the Taylor diagram plots for some selected transformation models are presented in Fig. 11 for both datasets F-CLAY/7/216 and F-CLAY/7/135, indicating the overall performance of the models. As can clearly be seen also from this presentation, all models perform better for database F-CLAY/7/135. Based on this, two obvious conclusions can be made. Firstly, the removal of data from F-CLAY/7/216 to obtain F-CLAY/7/135 consisted at least of the major part of true outliers. Secondly, incoherent data deteriorated the performance of all models evaluated. A further conclusion from Fig. 11 is that the ML models perform better than the traditional equations. Generally, the ANN models gave the best performance, although the difference to GP and MARS models was small. It should also be remembered that for the ANN, 1000 training sets were performed while only one was done for the GP model, MARS lacking the influence of different training sets.

Table 5 Range of values used for training of the models given as training dataset value/all data value, where value corresponds to the mean, min, and max values

Variable	F-CLAY/7/216			F-CLAY/7/135		
	Ratio of values between training data and all data					
	Mean	Min	Max	Mean	Min	Max
σ'_v	0.98	1.00	0.83	1.03	1.24	1.00
σ'_p	0.99	1.00	1.00	1.04	1.25	1.00
LL	0.99	1.00	1.00	1.01	1.14	1.00
PL	1.00	1.00	1.00	0.98	1.00	1.00
w	1.00	1.00	1.00	1.00	1.20	0.87
S_t	0.99	1.00	1.00	0.97	2.00	1.00
s_{umob}	0.99	1.00	1.00	1.04	1.68	1.00
OCR	1.01	1.00	1.00	1.02	1.07	1.00
s_{umob}/σ'_p	0.99	1.16	1.00	0.98	0.99	0.93

Fig. 11 Taylor diagrams for modelling of s_{umob} for the various transformation models for datasets F-CLAY/7/216 (a) and F-CLAY/7/135 (b)



By studying the performance of different ANN trials in testing, it can be noted that adding information generally tends to improve the performance in the training dataset. Remembering that due to the initial guess of the optimization model there is some randomness in a single ANN training and that in the chosen training sets the overall performance was optimized, it is understood that the trend could be more apparent. The important conclusion from this is that if one uses all available data to train a model, like often done for traditional transformation models, the predictions may improve by adding variables. However, these may still be false predictors causing overfitting of the data. Then, the model may fail in predicting any additional data, resulting in even dangerous outliers.

It is obvious that σ'_p is by far the most dominant factor in predicting s_u . The inclusion of σ'_v also improves the estimates. Both conclusions are logical with respect to the yielding of clays, i.e., the relationship between the yield surface of clays, in situ stress, and failure. However, it should be noted that for the more scattered dataset F-CLAY/7/216 the inclusion of the effective vertical stress did not improve the performance of the testing data. The situation is more unclear with respect to plasticity properties. While there was clear indication of overtraining when additional parameters were included in the model, it can be argued that including LL could slightly improve the performance of some models. It is good to note that the correction made for s_{uFV} to derive $s_{u(mob)}$ is based on LL. There is no reason to believe that this correction would be perfect, and some influence of LL to $s_{u(mob)}$ is thus likely.

Comparing the results for the two datasets F-CLAY/7/216 and F-CLAY/7/135, it was observed that for none of the methods and different sets of parameters, the dataset F-CLAY/7/216 led to a better transformation model and revealed some connections. A logical conclusion then is that the partially subjective cleaning of data for dataset F-CLAY/7/135 did not remove any potentially important data, but true outliers were removed from dataset F-CLAY/7/216. Acknowledging that the dataset F-CLAY/7/216 is more scattered, it can be further concluded that the risk for overtraining increases for more scattered data. Thus, if the data is scattered, simpler models should be preferred, as also indicated by the relatively good performance of using only σ'_p as input for F-CLAY/7/216 dataset, in comparison to the more extended models. This supports the discussions of results by D'Ignazio et al. (D'Ignazio et al. 2016), Mesri and Wang (Mesri and Wang 2017), and D'Ignazio et al. (D'Ignazio et al. 2017).

The comparison of the performance of traditional and ML-based transformation models is in that sense problematic, that while the whole datasets were used for the traditional models, only 70% was used to train the ML models with additional 15% for the validation of the ANN models.

Still, it is obvious that the ML models give a clearly better performance when comparing the results presented in Tables 2 and 4. For engineering purposes, it is though often more practical to apply traditional transformation models, of which type (2) gives a very fair prediction for soft sensitive clays. For ML models, ANN can be criticized of being a black box while no equation is provided. GP and MARS models provide an equation, but the equations are often rather complex and sensitive, so one needs to be careful in their implementation. At the end, it is much up to the problem in hand if ML models give any additional benefits in predicting s_u . For big database applications, ML offers intriguing possibilities. The authors also strongly believe that ML is a very useful tool to provide new insights over various dependencies and, as in the present study, give explanation to some of the present discrepancies researchers may have on correlations between geotechnical properties.

Conclusions

In this study, transformation models for s_{umob} have been developed based on traditional correlation models as well as ML models. Two different datasets have been used, F-CLAY/7/216 being the larger but more incoherent one while some data has been cleaned from its subdataset F-CLAY/7/135. Based on the study, the following conclusions can be made:

- Based on all models, the cleaning of the F-CLAY/7/216 data into dataset F-CLAY/7/135 removed true outliers.
- As expected, preconsolidation pressure is the most important variable in predicting the mobilized undrained shear strength.
- For both traditional and ML models, the inclusion of effective vertical stress/OCR in addition to the preconsolidation pressure did not improve the prediction for the more scattered dataset F-CLAY/7/216.
- For the more coherent dataset F-CLAY/7/135, the inclusion of effective vertical stress and/or OCR improved the prediction notably for both traditional and ML models.
- The inclusion of depth and w improved slightly the performance of the ANN models, especially for lower s_u values. This can though well be partly due to the used datasets, with a very limited OCR.
- The ANN-based sensitivity study with 1000 different training sets revealed that the outcome of repetitive trainings may vary significantly. The difference was more significant for the more scattered dataset F-CLAY/7/216.
- The ANN-based sensitivity study of different training scenarios revealed that the performance of the training dataset can generally be improved by including more parameters. However, the performance of the testing

data rapidly decreases when false predictors are added. This trend is also more significant for the more scattered dataset F-CLAY/7/216.

- The above indicates that there is a greater risk for over-training for more incoherent data. That is why, it is recommended to use simple models with few input data, when the data is scattered.
- Generally, all ML models performed better than the traditional models with the same input parameters, indicating that the dependency of s_u on these parameters is not as straightforward as in the traditional models.

Data availability

1. The datasets analyzed during the current study are available in the webpage of ISSMGE technical committee 304 at <http://140.112.12.21/issmge/tc304.htm>

2. The datasets generated during the current study are available from the corresponding author on reasonable request.

Acknowledgements The study is made as an independent study. It was initiated by some earlier work by the first author relating to transformation models for undrained shear strength. Initially, the goal was to seek if machine learning could reveal some of the hidden dependencies related to which parameters truly effect on the undrained shear strength (see “Introduction”). As discussed in the paper, the geotechnical community does not agree on which parameters to include in transformation models. In addition, we wanted to find out how much the quality of a database influences on the results and conclusions made thereof by various researchers. As the work progressed, we found that, for example, for ANN, the outcome can vary quite significantly even if exactly the same division of data into training, validation, and testing is used, for subsequent training trials. Based on this, we decided to repeat ANN trainings 1000 times to study its influence. To illustrate the results, as well as the danger of overprediction, we found the Taylor diagram to be very helpful.

Funding Open access funding provided by Tampere University including Tampere University Hospital, Tampere University of Applied Sciences (TUNI).

Declarations

Conflict of interest The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adamowski J, Chan HF (2011) A wavelet neural network conjunction model for groundwater level forecasting. *J Hydrol* 407(1–4):28–40
- Adamowski J, Karapataki C (2010) Comparison of multivariate regression and artificial neural networks for peak urban water-demand forecasting: evaluation of different ANN learning algorithms. *J Hydrol Eng* 15(10):729–743
- Bjerrum L (1972) Embankments on soft ground. Proc., ASCE Specialty Conf. on Performance of Earth and Earth-Support Structures, Purdue University, ASCE, Reston, VA, II, 1–54
- Chandler RJ (1988) The in-situ measurement of the undrained shear strength of clays using the field vane. In Vane shear strength testing in soils: field and laboratory studies. ASTM STP 1014. ASTM, Philadelphia, Pa. pp. 13–44
- Ching J, Phoon KK (2012) Modeling parameters of structured clays as a multivariate normal distribution. *Can Geotech J* 49(5):522–545. <https://doi.org/10.1139/T2012-015>
- Ching J, Phoon KK (2014) Correlations among some clay parameters — the multivariate distribution. *Can Geotech J* 51:686–704. <https://doi.org/10.1139/cgj-2013-0353>
- D’Ignazio M, Phoon KK, Tan SA, Lämsivaara TT (2016) Correlations for undrained shear strength of Finnish soft clays. *Can Geotech J* 53(10):1628–1645
- D’Ignazio M, Phoon KK, Tan SA, Lämsivaara TT, Lacasse S (2017) Reply to the discussion by Mesri and Wang on “Correlations for undrained shear strength of Finnish soft clays.” *Can Geotech J* 54(5):749–753
- Daliakopoulos I, Coulibalya P, Tsani IK (2005) Groundwater level forecasting using artificial neural networks. *J Hydrol* 309(1–4):229–240
- Foroozesh J, Khosravani A, Mohsenzadeh A, Mesbahi AH (2014) Application of artificial intelligence (AI) in kinetic modeling of methane gas hydrate formation. *J Taiwan Inst Chem Eng* 45(5):2258–2264
- Friedman JH (1991) Multivariate adaptive regression splines. *Ann Stat* 19(1):1–67
- Hansbo S (1957) A new approach to the determination of the shear strength of clay by the fall-cone test. In: Proceedings of the royal Swedish geotechnical institute No 14, Royal Swedish Geotechnical Institute, Stockholm, pp 1–47
- Haykin S (2005) Neural networks: a comprehensive foundation, Prentice-Hall Publications
- Helene Lund KV (1977) Methods for reducing undrained shear strength of soft clay. in Proceedings of the Swedish Geotechnical Institute, Report No. 3. Linköping
- Jamiolkowski M, Ladd CC, Germain JT, Lancellotta R (1985) New developments in field and laboratory testing of soils. In: Proceedings of the 11th International conference on soil mechanics and foundation engineering. August Aimé Balkema, San Francisco, CA, pp 57–153
- Karaagac B, Inal M, Deciz V (2009) Artificial neural network approach for predicting optimum cure time of rubber compounds. *J Mater Des* 30(5):1685–1690
- Karlsrud K, Hernandez-Martinez FG (2013) Strength and deformation properties of Norwegian clays from laboratory tests on high-quality block samples. *Can Geotech J* 50(12):1273–1293. <https://doi.org/10.1139/cgj-2013-0298>
- Khosrojerdi S, Vakili M, Yahyaei M, Kalhor K (2016) Thermal conductivity modeling of graphene nanoplatelets/deionized water nanofluid by MLP neural network and theoretical modeling using experimental results. *J Int Commun Heat Mass Transf* 74:11–17
- Koza JR (1994) Genetic programming as a means for programming computers by natural selection. *Stat Comput* 4, 87–112 (1994). <https://doi.org/10.1007/BF00175355>

- Ladd CC, Foot R (1974) New design procedure for stability of soft clays. *J Geotech Eng Div* 100:763–786. [https://doi.org/10.1016/0148-9062\(74\)90494-X](https://doi.org/10.1016/0148-9062(74)90494-X)
- Larsson R (1980) Undrained shear strength in stability calculation of embankments and foundations on soft clays. *Can Geotech J* 17:591–602. <https://doi.org/10.1139/t80-066>
- Larsson R, Sällfors G, Bengtsson P-E, Alén C, Bergdahl U, Eriksson L (2007) Utvärdering av skjuvhållfasthet i kohesionsjord. Evaluation of shear strength in cohesive soil (In Swedish). Swedish Geotechnical Institute. Revision of SGI Information No 3, Linköping
- Mesri G (1975) Discussion on “new design procedure for stability of soft clays.” *J Geotech Eng Div* 101(4):409–412
- Mesri G, Wang C (2017) Discussion of “Correlations for undrained shear strength of Finnish soft clays.” *Can Geotech J* 54(5):745–748
- Ratahallintokeskus (2005) (RHK julkaisu), B15, Ratahallintokeskus, Helsinki. (In Finnish. In English: Guidelines for embankments stability calculation by the Finnish Transport Agency, publication B15)
- Sekulic S, Kowalski BR (1992) MARS: a tutorial. *J Chemom* 6(4):199–216
- Shahin MA, Jaksa MB, Maier HR (2001) Artificial neural network applications in geotechnical engineering. *Aust Geomech J* 36(1):49–62
- Skempton AW (1954) Discussion of the structure of inorganic soil. *J Am Soc Civil Eng* 80(478):19–22
- Swedish Geotechnical Institute. (1969) Reduction in shear strength with reference to liquid limit and sulphide content. In: Summary from technical meeting 11th December 1969. Stockholm
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res* 106(D7):7183–7192
- Zhang W, Phoon K-K (2022) Editorial for Advances and applications of deep learning and soft computing in geotechnical underground engineering. *J Rock Mech Geotech Eng* 14(3):671–673. <https://doi.org/10.1016/j.jrmge.2022.01.001>
- Zhang W, Wu C, Zhong H, Li Y, Wang L (2020) Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci Front*. <https://doi.org/10.1016/j.gsf.2020.03.007>
- Zhang W, Xin Gu, Tang L, Yin Y, Liu D, Zhang Y (2022) Application of machine learning, deep learning and optimization algorithms in geoenvironment and geoscience: comprehensive review and future challenge. *Gondwana Res* 109:1–17