Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore. ©2025 The Author(s), CC-BY 4.0 licence.
doi: 10.3850/GRF-25280825_isgsr-156-P210-cd

A DATA-DRIVEN APPROACH FOR SOIL PARAMETER DETERMINATION USING SUPERVISED MACHINE LEARNING

Haris Feli•

Institute of Soil Mechanics, Foundation Engineering and Computational Geotechnics, Graz University of Technology, Austria. E-mail: h.felic@tugraz.at

IslamMarzouk

Institute of Soil Mechanics, Foundation Engineering and Computational Geotechnics, Graz University of Technology, Austria. E-mail: islam.marzouk@tugraz.at

Franz Tschuchnigg

Institute of Soil Mechanics, Foundation Engineering and Computational Geotechnics, Graz University of Technology, Austria. E-mail: franz.tschuchnigg@tugraz.at

Soil constitutive models have significantly advanced over the years, often with an increase in parameters. Accurate determination of these parameters is critical, as inaccuracies can lead to unreliable numerical simulations. Conventional calibration practicetypically relies on laboratory testing, which is often impractical for applications, particularly at early project stages. An ongoing research projectofthe Computational Geotechnics Group at Graz University of Technology focuses on developing an *Automated Parameter Determination* frameworkthat integrates a graph-based approach to derive constitutive model parameters from in-situ tests. This framework uses established correlations to identify parameters based on in-situ tests. However, the multiplicity of correlations for a given parameter introduces inherent challenge in selecting the recommended value. An alternative methodinvolves using advanced regression algorithms to enhance the robustness of parameter determination through data-driven techniques.This approach can improve the quality of numerical simulations and minimizes the uncertainty in the parameter calibration process. In this paper, supervised machine learning regression models are employed to predict soil parameters, including saturated unit weight, undrained shear strength, and small-strain shear modulus (via shear wave velocity) using cone penetration test data as input. The performance of these models is benchmarked using data from two Norwegian GeoTest Site locations and are validated based on laboratory-derivedvalues, in-situ measurements, and traditional correlation-based methods. The findings demonstrate the potential of advanced regressionmodels to generalize soil parameter predictions across test sites, significantly improving reliabilitywhile reducing site-specific biases.

Keywords: Constitutive soil parameter, machine learning, data-driven,site characterization, Cone penetration test

1. Introduction

Constitutive soil models approximate the non-linear behaviour of soils, which makes model selection and parameter determination challenging. Generally, increasing model complexity raises the number of parameters. One such parameter is the stress-dependent small-strain shear modulus G_0, which depends on soil density Áand shear wave velocity v_s.The latter can be obtained through in-situ (e.g. seismic cone penetration test (CPT))or lab tests (e.g. bender elements). Other parameters, such as saturated unit weight $^3{}_{sat}$ and undrained shear strength s_u, are derived from lab tests (e.g. index tests for ^{3}, triaxial tests for s_u) or in-situ tests (e.g. vane shear tests for s_u).Ground conditions and basic soil properties in practical engineering are often initially assessed using local geological knowledge and experience from similar sites. However, site investigations provide more accurate, location-specific quantitative data.In geotechnical practice, in-situ data are often used in correlations to estimate further soil parameters, requiring statistical analysis and engineering judgement (Phoon et al. 2022). An ongoing research project focuses on developing an *Automated Parameter Determination* (APD) framework using a graph-based approach, using established correlations to identify constitutive model parameters from in-situ tests (Marzouk et al. 2024).Despite its potential, the multiplicity of correlations for a given parameter makes it difficult to determine a single, recommended value. An alternative involves using advanced machine learning (ML) regression algorithms to enhance the robustness of parameter determination through data-driven approaches.This paper explores the application of ML models to improve the accuracy of predicting soil parameters. The proposed framework uses CPT data to predict lab parameters such as $^3{}_{sat}$ and s_u, as well as in-situ measurements ofv_s. The study details the database, ML training process, ML model performance,and validation using data from two Norwegian GeoTest Sites (NGTS)(L'Heureux and Lunne 2020)**.**

2. Databases for Machine Learning

2.1. *Databases for laboratory parameters*

The database for laboratory parameters comprises pairs of CPT measurements and corresponding soil parameters. These soil parameters are determined through lab tests conducted on soil samples extracted from boreholes located near the CPT tests. The database includes 384 pairs for ^{3} and 84 pairs for s_u. Data sources include publicly available data from Ballina data, NGTS (both accessible at geocalcs.com/datamap), and from the Netherlands(see Lengkeek and Breedeveld 2022), and private data from Norway and Austria.

2.2. *Databases for in-situ parameters*

The in-situ database comprises 46 seismic CPTs and 254 seismic CPTu tests, resulting in 21,760 data points. Data sources for this database include the New Zealand Geotechnical Database (Scott et al. 2015), PremstallerGeotechnik(Oberhollenzer et al. 2021), Dutch site data (github.com/snakesonabrain/ isc7_datasets), and Taiwan site data (https://data.mendeley.com/ datasets/v7frv3k2d3/1)

3. Machine Learning Models

The created ML models are discussed in this section, covering feature selection, the training workflow, and model evaluation. This paper focuses exclusively on the use of one decision tree algorithm,namely XGBoost (eXtreme Gradient Boosting Decision Tree), which is an enhanced version of the Gradient Boosting Decision Tree (GBDT) algorithm. This algorithm focuses on improving computational speed and model accuracy by introducing more regularization terms and parallelized tree building.

3.1. *Feature selection*

In ML, a modelis a function that produces an output based on a given input (Deisenroth et al. 2021).Here, the input matrixincludes variables of depth, tip resistance q_c, sleeve friction f_s, friction ratio R_f, as also utilized in Feli• et al.(2024). Preliminary studies demonstratedthat usingdepth, q_c, f_s, and $R_f$leads to goodmodel performance.In this framework, separate models are developed to predict each soil parameter:^{3} $_{sat}$, s_u, and$v_{s.}$

3.2. *Machine learningworkflow training*

The training workflow, following a process described byDeisenroth et al.(2021), begins with splitting the dataset into an 80% training set and a 20% test set. The training set is exclusively used for model training, while the test set is reserved for testing the model after model training.The initial training step starts using an initial set of hyperparameters. After training, the model predictions are compared to the validation set (a subset of the training data), and its performance (validation loss) is computed with an objective function. To minimize overfitting, cross-validation is applied, where the data is split into K subsets, with K-1 used for training and one for validation. The model's performance is averaged over these K iterations; in this study, ten-fold cross-validation is employed. Additionally, early stopping is used to halt training if the validation error does not improve for ten consecutive iterations. Hyperparameter optimization is performed using the Differential Evolution algorithm from pymoo, with a population size of 25. This algorithm identifies the optimal hyperparameters by maximizing the coefficient of determinationR^2(see section 3.3) as the objective function. The optimization process runs for up to 1000 iterations or until a maximum relative error of $10^{-4}$is achieved. Once the hyperparameters are optimized, the model is retrained using the combined training and validation data. The model performance is then evaluated on the 20 % test set. If the test loss is satisfactory, a final training phase is conducted with the entire dataset using the best hyperparameters obtained from the optimization.

3.3. *Evaluation of model performance*

For evaluation and optimization, the coefficient of determination R^2 serves as the objective function. R^2 is expressed as:

$$R^2 = 1 - \sum(y_i - f_i)^2 / \sum(y_i - \bar{y})^2 \qquad (1)$$

Here, y_i denotes the true values, f_i represents the predicted values, and $\bar{y}$ is the mean of the true values. R^2measures the predictive accuracy of a model, typically ranging from 0 to 1. A higher R^2value indicates betteralignment between the model's predictions and the observed data, while a lower R^2 suggests poorer alignment.

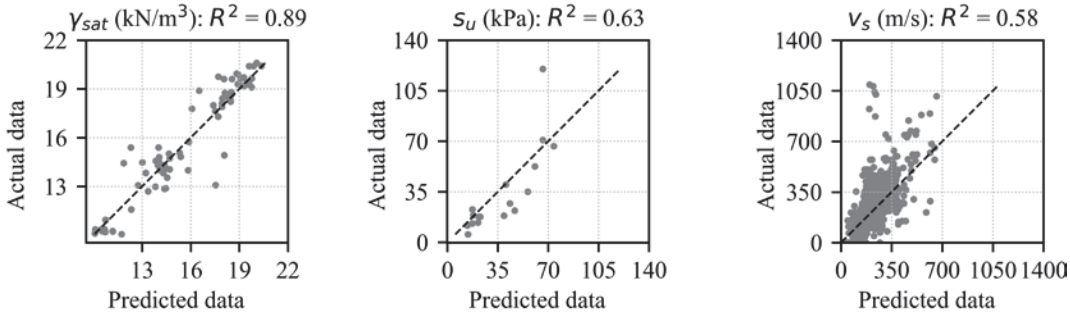

Fig. 1. ML model performance after hyperparameter optimization with ten-fold cross validation and early stopping.

Fig. 1 illustrates the performance of the ML model after hyperparameter optimization, focusing on the test loss evaluation. Each soil parameter ($^3_{sat}$, s_u, v_s) in Fig. 1 shows the predicted data plotted against the actual data (as grey scatter points). The dashed 45° line represents perfect alignment between the actual and predicted values. The scatter distributions indicate strong overall performance for most predictions, with some instances of overestimation and underestimation. For example, some $v_s$predictions between 0-350 m/s are either under- or overestimated compared to the actual data. The R^2 values for the three soil parameters range from 0.58 to 0.89.

4. Application of Machine LearningModels

This section presents the performance of the trained ML model using real CPT data from NGTS (L'Heureux and Lunne 2020). These sites encompass various soil types, including clay, silt, quick clay, sand, and permafrost. For this study, only the silt and clay sites are considered. The ML model's performance is evaluated by comparing its predictions with laboratory data (index, triaxial, and direct shear tests), in-situ tests (seismic dilatometer tests, SCPTu, CPTu) from the selected test sites, and the following established correlations:

C1	Robertson and Cabal (2010)	$\gamma_{sat} = \gamma_w(0.27(\log R_f) + 0.36(\log q_t/p_a) + 1.236)$	(2)
C2	Lengkeek and Brinkgreve(2022)	$\gamma_{sat} = 19.5 - 2.87 * (\log(9000/q_t)/\log(20/R_f))$	(3)
C3	Mayne(2016)	$s_u = (q_t - u_2)/N_{ke}; N_{ke} = 8$	(4)
C4	Mayne(2016)	$s_u = (q_t - \sigma_v)/N_{kt}; N_{kt} = 12$	(5)
C5	Mayne(2006)	$v_s = 51.6 * \ln(f_s) + 18.50$	(6)
C6	Hegazy and Mayne(1995)	$v_s = 12.02 * q_c^{0.319} * f_s^{-0.0466}$	(7)

4.1 Norwegian GeoTest Sites – Silt site

Fig. 2 presents the CPTu data from the silt site in Halden, alongside ML predictions, lab and in-situ sitedata, and correlations for each soil parameter:$^3_{sat}$, s_u, v_s, and G_0. G_0 was calculated as $v_s^2 \cdot ^3_{sat}/9.81$ for both correlations and ML predictions. The grey hatched regions in each figure indicate the upper and lower bounds of the selected correlations. For$^3_{sat}$, v_s,and G_0,these regions reveal scatter,and a significant offset compared to the lab and in-situ data. $s_u$also exhibits a larger rangeand offset relative to the lab data. For s_u, the total stress $\sigma_v$using Eq. (5) was calculated based on the average $^3_{sat}$from Eq. (2) and (3). The ML predictions for $^3_{sat}$, v_s, and G_0 show a very good agreement with the lab and in-situ data across the entire depth profile. However, predictionsfor $s_u$tend to overestimate the lab values, although the increasing trend of s_u with depth is well captured compared to correlations and lab data. The reason for this overestimation is part of ongoing research.Comparing both methods (correlations, and ML predictions), ML models tend to be less site-specific as the training database contains no data from NGTS, whereas correlations inherently incorporate site-specific characteristics, making it often difficult to determine which correlation is most suitable for a given site.

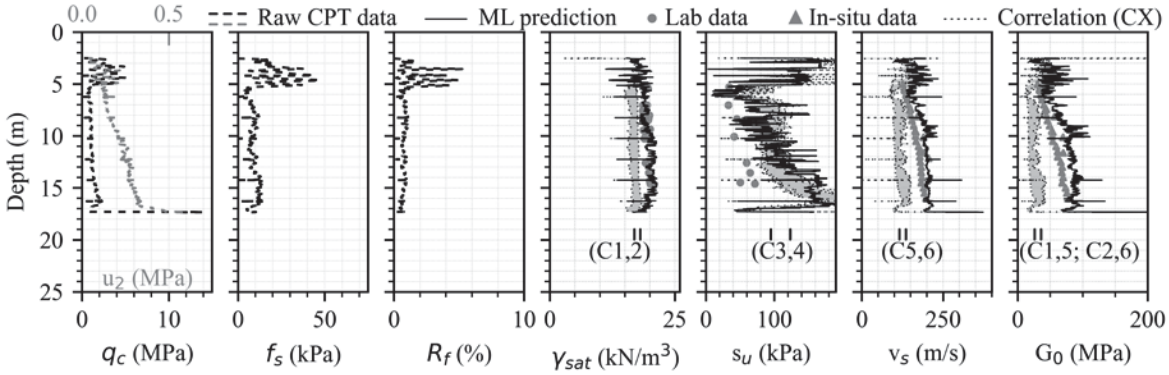

Fig. 2. Raw CPTu site data for HALC11 (dashed black line), along with the ML prediction (solid black line), lab and in-situ data (represented by circular and triangular markers, respectively), and correlations (dotted black line).

4.2 Norwegian GeoTest Sites – Clay site

Fig. 3 illustrates SCPTu data from the clay site in Onsøy, ML predictions, lab and in-situ site data, and correlations for each soil parameter: $^{3}_{sat}$, s_u, v_s, and G_0. G_0 was calculated similar to the silt site. The grey hatched regions represent the bounds of the selected correlations. $^{3}_{sat}$ and $s_u$correlations align well with the lab data, showing minimal scatter. Forv_s,and G_0, minor scatter and slight offsets are observed compared to the in-situ data. ML predictions for $^{3}_{sat}$closely match lab data along depth. Predictions for $v_s$and G_0 align well with the in-situ data between 1–7 m, although anoverestimated trend occurs below this depth.The overestimation is most likely due to sparse datapoints (CPT data)within the database for this material. $s_u$predictions also alignwell with the lab data, but exhibit a sensitive, zig-zag pattern with increasing depth.Overall, the ML models appear to be less sensitive to site-specific conditions, comparable to the findings at the silt site,whereas correlations may be influenced by site-specific characteristics.

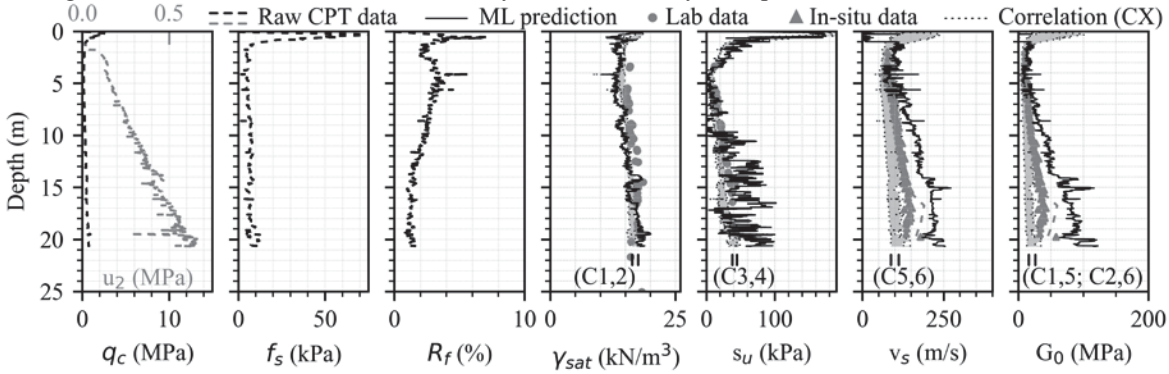

Fig. 3. Raw SCPTu site data for ONSC08 (dashed black line), along with the ML prediction (solid black line), lab and in-situ data (represented by circular and triangular markers, respectively), and correlations (dotted black line).

5. Conclusion

This study demonstrates the potential of supervised machine learning (ML) models to enhance soil parameter determination from cone penetration test data, addressing the limitations in traditional empirical correlations. By employing advanced regression algorithms, such as XGBoost, data-driven approaches significantly improveprediction accuracy for soil parameters, including saturated unit weight, undrained shear strength, and small-strain shear modulus (derived from shear wave velocity). Validation using data from the Norwegian GeoTest Sites highlights the robustness of these ML models in capturing soil parameters across different locations. While slight overestimations occur at certain depths, the models provide generalizable predictions and can effectively reduce uncertainty in parameter calibration. These advancements are particularly valuable in scenarios requiring robust and reliablesoil characterization, especially where limited data is available. The proposed framework will be incorporated into the *Automated Parameter Determination*research project as an additional information source to enhance parameter accuracy. Future work will focus on database extensions, managing outliers, and investigating potential biases that may affect the ML performance. The open-source availability of both datasets and algorithms on GitHub ensures transparency and fosters further collaborations.

Acknowledgement

The authors express their gratitude to Simon Oberhollenzer for providing a subset of the laboratory database.

Data repository

The GitHub repository is available here: https://github.com/harifel/ISGSR25_DataDrivenSiteCharacterization/

References

Deisenroth, M. P.,Ong, C. S.,and Faisal, A. A. (2021). *Mathematics for machine learning*. Cambridge University Press.

Feli•, H.,Marzouk, I.,Peterstorfer, T.,Tschuchnigg, F. (2024). Data-driven site characterization - Focus on small-strain stiffness. In M. Arroyo, A. Gens*, 7th ISC – Ground models, from big data to engineering judement*.

Hegazy, Y. A. and Mayne, P. W. (1995). Statistical correlations between Vs and cone penetration data for different soil types. In Swedisch Geotechnical Society (Eds.), *International Symposium on Cone Penetration Testing, CPT' 95*.

L'Heureux, J.-S.,Lunne, T. (2020). Characterization and Engineering properties of Natural Soils used for Geotesting, *AIMS GeosciencesVolume (6)*.https://doi.org/10.3934/geosci.2020004.

Lengkeek, H. J.,Breedeveld, J. (2022). Eemdijk full-scale test on dike reinforced by sheet pile. Datasethttps://doi.org/10.4121/19213890.v1.

Lengkeek, H. J. and Brinkgreve, R. (2022). CPT-based unit weight estimation extended to soft organic clays and peat: An update. In L. Tonni, G. Gottardi (Eds.), *Cone Penetration Testing 2022*.https://doi.org/10.1201/9781003308829-71.

Marzouk, I.,Brinkgreve, R.,Lengkeek, A.,and Tschuchnigg, F. (2024).APD: An automated parameter determination system based on in-situ tests.*Computers and Geotechnics176*.https://doi.org/10.1016/j.compgeo.2024.106799.

Mayne, P. (2014). Interpretation of geotechnical parameters from seismic piezocone tests. In Robertson, P. K., Cabal, K. L. (Eds.), *Proceedings from the 3rd International Symposium on Cone Penetration Testing. CPT' 14*.

Mayne, P. (2016). Evaluating effective stress parameters and undrained shear strengths of soft-firm clays from CPTu and DMT. In Lehane B. M., Acosta-Martinez H. E., Kelly R. (Eds.), *Geotechnical and Geophysical Site Characterization 5*.

Oberhollenzer, S.,Premstaller, M.,Marte, R.,Tschuchnigg, F.,Erharter, G. H.,and Marcher, T. (2021). Cone penetration test dataset Premstaller Geotechnik, *Data in Brief 34*. https://doi.org/10.1016/j.dib.2020.106618.

Phoon, K.-K., Ching, J., and Shuku, T. (2022). Challenges in data-driven site characterization. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16*.https://doi.org/10.1080/17499518.2021.1896005.

Robertson, P. K., & Cabal, K. L. (2010). Estimating soil unit weight from CPT 2010. InRobertson, P. K., P.W. Mayne (Eds.), *Proceedings of the 2ndInternationalSymposium on Cone Penetration Testing*.

Scott, J. W.,van Ballegooy, S.,Stannard, M.,Lacrosse, V.,et al.(2015). The Benefits and Opportunities of a Shared Geotechnical Database. In Cubrinovski (Ed.), *6th International Conference on Earthquake Geotechnical Engineering*.