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-166-P377-cd

Machine learning predictions on an extensive geotechnical dataset of laboratory tests in Austria

Enrico Soranzo[1]

[1] *Department of Civil Engineering and Natural Hazards, University of Natural Resources and Life Sciences, Vienna, Austria. E-mail: enrico.soranzo@boku.ac.at*

This study presents and analyzes an extensive geotechnical dataset spanning over 20 years, including particle size distributions, Atterberg limits, Proctor tests, permeability tests and direct shear tests. The dataset, covering 24 features, also includes some locations and allows correlation analysis and preliminary predictive modeling. Previous research has explored links between properties like liquidity index and undrained shear strength, particle size distribution and friction angle and liquid limit and plasticity index with residual friction angle. Python-based exploratory analysis and machine learning revealed cross-correlations among variables and predictive models for cohesion and friction angle achieved moderate accuracy. The dataset's breadth and repeated testing enable the development of predictive models with varying reliability, highlighting its potential as a resource for geotechnical studies.

Keywords: Atterberg limits; Direct shear tests; Exploratory data analysis; Friction angle; Geotechnical dataset; Machine learning models; Particle size distribution; Permeability tests; Predictive models; Proctor tests; Soil parameters.

1. Introduction

Correlation equations among soil parameters have significant practical utility, as they can provide estimates that would otherwise necessitate additional testing, thereby leading to cost savings. Previous studies have attempted to establish relationships between various soil properties, such as linking the liquidity index to undrained shear strength (Ching and Phoon, 2012), particle size distribution to friction angle (Kaydelen et al., 2009), liquid limit and plasticity index to residual friction angle (Das et al., 2011). The standards primarily correlate soil properties based on field tests (European Committee for Standardization, 2010).

In this study, we present an extensive geotechnical dataset derived from laboratory tests conducted over a period of more than 20 years. The dataset includes particle size distributions, Atterberg limits, Proctor tests, permeability tests and direct shear tests. Where available, the location of the soil samples has been documented. In total, 24 variables have been recorded.

For some soil samples, multiple tests have been performed (Table 1), allowing for the assessment of correlations between variables and the training of preliminary machine learning models.

2. Methods

Soil samples were predominantly collected from the Eastern federal states of Austria, including Vienna, Lower Austria and Burgenland (Fig. 1).

Initially conducted under Austrian standards and later international standards, laboratory methods included mechanical sieving and hydrometer tests for particle size, shear box tests using 10 cm cylindrical samples and cylindrical permeameters for permeability testing (Fig. 2). As an application of the dataset, Artificial Neural Networks (ANNs) were developed to predict friction angle (ϕ) and cohesion (c). The input data consisted in: the percentage content of Co, Gr, Sa, Si and Cl; d_{max}, C_U, C_C, d_{10}, d_{50}; ρ_s, ρ_{Pr}, w_{opt}; w_L, w_P, w_S, I_P, w, I_C, I_A; $\log(k_{10})$; latitude and longitude of the sampled soils. For φ, the highest positive correlations were with d_{max} (0.59), Gr (0.58) and d_{10} (0.53). Negative correlations with I_P (-0.66), Cl (-0.62) and Si (-0.53) were observed. c correlated positively with ρ_{Pr} (0.12) and I_C (0.11) and negatively with $\log(k_{10})$

Table 1. Pairwise numerosity of test results.

	Location	PSD	PROC	ATT	PERM	DST
Location	646	618	395	60	111	74
PSD	618	719	438	82	120	97
PROC	395	438	471	40	98	79
ATT	60	82	40	89	20	39
PERM	111	120	98	20	124	21
DST	74	97	79	39	21	137

Abbreviations: PSD: Particle Soil Distributions; PROC: Proctor tests; ATT: Atterberg's limits; PERM: Permeability test; DST: Direct Shear Test.

(-0.051) and w_L (-0.053). Data preprocessing involved K-Nearest Neighbors (KNN) imputation for missing values (Troyanskaya et al., 2001), Local Outlier Factor (LOF) methods for outlier removal (Breunig et al., 2000) and Z-score normalization. The ANNs, trained via bootstrap resampling, incorporated regularization (L_2 and L_1-L_2), LeakyReLU activations and dropout layers to prevent overfitting, optimized with RMSprop. The ANN had 3 hidden dense layers each with 32 neurons. Hyperparameters were tuned manually by comparing the test results of 61 different combinations of value. Overfitting was also controlled with an early stopping callback. The ensemble averaged predictions from 16 base ANN models.

3. Results

The performance of the model was evaluated using the R^2 and MAE performance metrics, as shown in Fig. 3. For the friction angle prediction, the training R^2 value of 0.837 indicates a strong fit between the observed and predicted values during training, while the test R^2 of 0.700 shows moderate generalization to the test data. The MAE values are also indicative of reasonable accuracy.

For the cohesion prediction, the model performed even better, with a high training R^2 value of 0.932 and a corresponding MAE of 2.603 kPa. The test performance, with an R^2 of 0.837 and

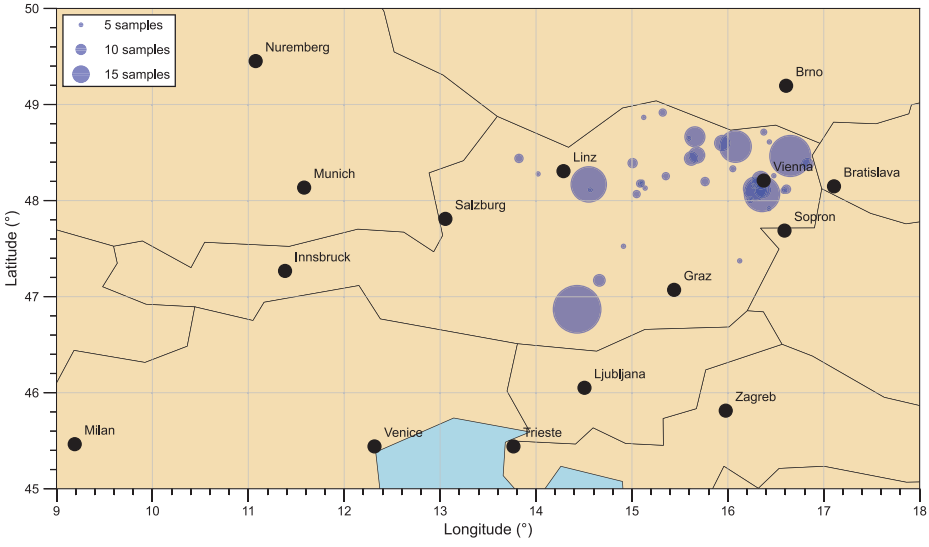

Fig. 1. Location of the samples tested in the laboratory.

MAE of 3.890 kPa, also demonstrates good generalization.

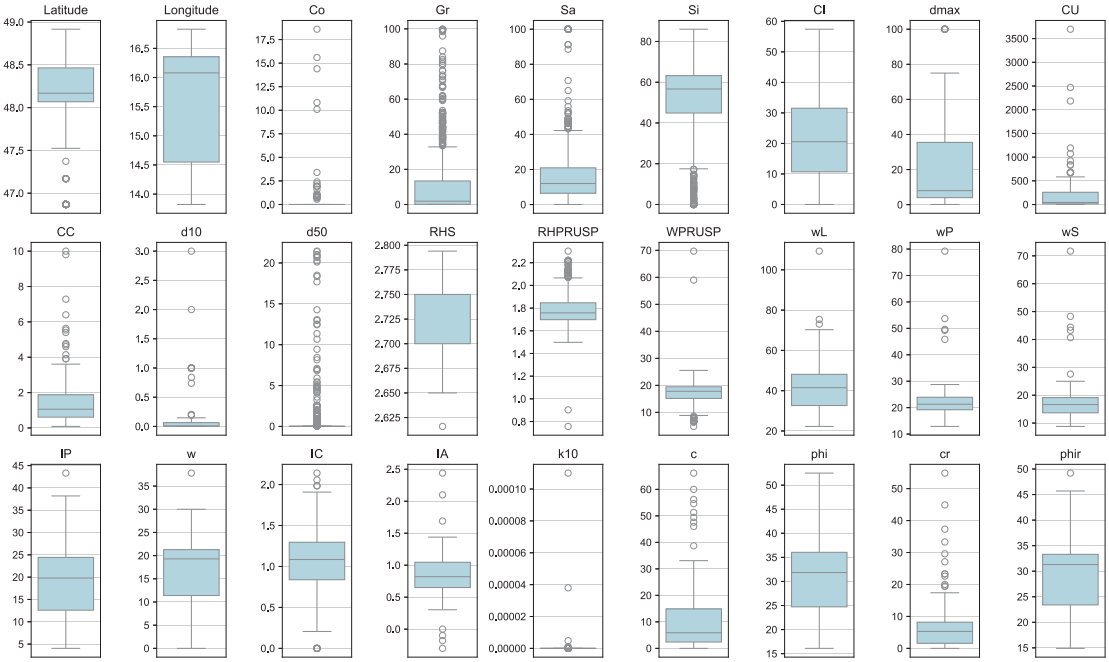

Fig. 2. Descriptive statistics of the dataset by means of boxplots.

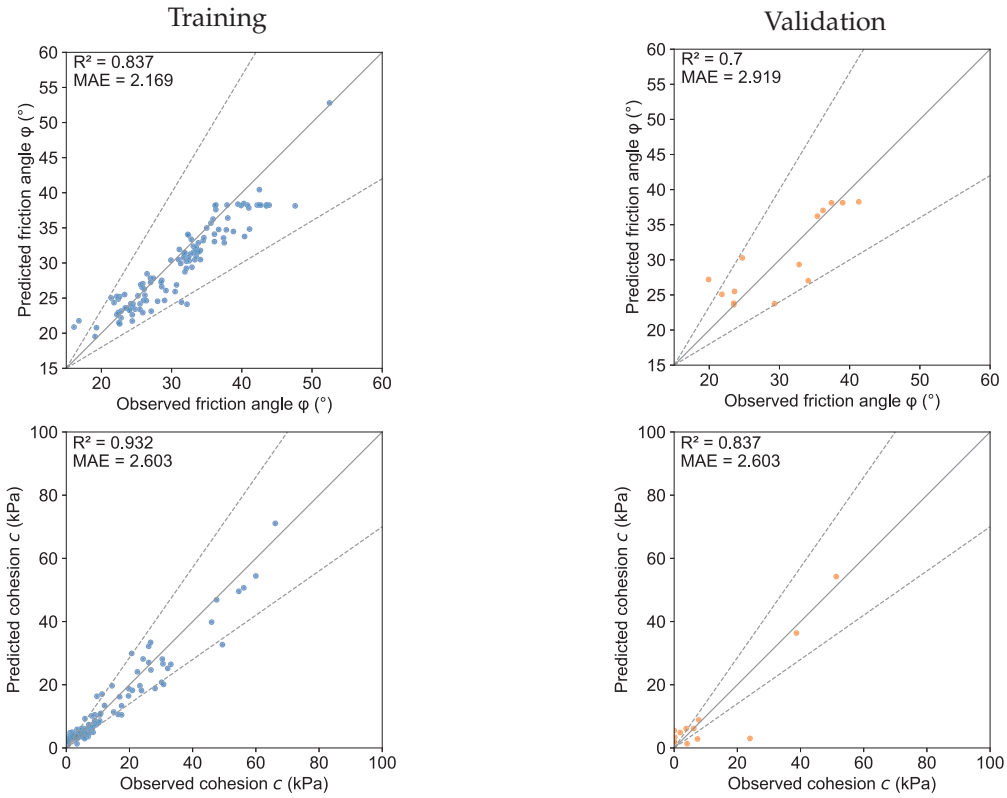

Fig. 3. Results of the model training and validation.

4. Discussion and Conclusions

This study introduces an extensive geotechnical dataset compiled from laboratory testing in Austria. The dataset's significance lies in its ability to enable cross-correlation between geotechnical parameters, such as linking particle size distribution to friction angle or Atterberg limits to cohesion. These correlations are crucial for geotechnical design, allowing practitioners to estimate parameters without the need for extensive and costly laboratory tests. However, challenges such as missing data and variability in testing standards had to be addressed using KNN imputation LOF methods for detecting anomalies.

To exemplify the dataset's utility, a machine learning model was trained to predict friction angle and cohesion. While the primary focus of this study is the dataset itself, the model's success underscores its potential as a benchmark for validating computational methods. The results demonstrated that moderately complex algorithms can yield meaningful predictions using the dataset.

As a publicly available resource, this dataset holds promise for the geotechnical engineering community, serving as a foundation for refining empirical correlations, developing predictive models and advancing research in soil mechanics and geotechnical design.

Acknowledgement

This study was funded by the European Union under the MSCA Staff Exchanges project 101182689 Geotechnical Resilience through Intelligent Design (GRID). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The author would like to thank Anand D'Souza for his help in preparing the data.

Supplementary material

The dataset can be accessed via the following link: https://doi.org/10.5281/zenodo.14251191.

References

Breunig, M.M., Kriegel, H.-P., Ng, R.T. and Sander, J. (2000) "LOF: identifying density-based local outliers". Proceedings of the ACM SIGMOD International Conference on Management of Data, 93–104.

Ching, J. and Phoon, K.-K. (2012) "Modeling parameters of structured clays as a multivariate normal distribution". Canadian Geotechnical Journal, 49(5), 522–545.

Das, S.K., Samui, P., Khan, S.Z., and Sivakugan, N. (2011) "Machine learning techniques applied to prediction of residual strength of clay". Central European Journal of Geosciences, 3, 449–461.

European Committee for Standardization (2010). "Ground investigation and testing: EN 1997-2".

Kayadelen, C., Günaydın, O., Fener, M., Demir, A., and Özvan, A. (2009) "Modeling of the angle of shearing resistance of soils using soft computing systems". Expert Systems with Applications, 36(9), 11814–11826.

Soranzo, E. (2024) "Dataset for "Machine learning predictions on an extensive geotechnical dataset of laboratory tests in Austria"". 9th International Symposium for Geotechnical Safety and Risk (ISGSR2025), Zenodo. doi: 10.5281/zenodo.14251191.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B. (2001) "Missing value estimation methods for DNA microarrays". Bioinformatics, 17(6), 520–525.