Third Workshop on the Future of
Machine Learning in Geotechnics
Florence, Italy
October 15-17 2025

LIMITATIONS TO GEOTECHNICAL DATASETS FOR MACHINE LEARNING: THE CASE OF P-S LOGGING

Bruno Stuyts
Department of Civil Engineering, Ghent University, Belgium. E-mail: bruno.stuyts@ugent.be
Esben Dalgaard
SolidGround ApS, Denmark. E-mail: ed@solidground.xyz

Keywords: Elastic properties, CPT, correlations, compression wave, shear wave.

1 Introduction

Establishing correlations between geotechnical parameters has been a focus topic since the start of the geotechnical engineering profession (Terzaghi et al., 1996). Even before the advent of machine learning (ML), geotechnical parameter correlations were established with a specific focus on cone penetration testing (CPT) (e.g. Robertson, 2009). The success of establishing reliable CPT-based correlations depends greatly on the quality of the underlying dataset. Effects of drilling disturbance, data interpretation and geotechnical variability all have a potential to obscure the underlying soil mechanical relations. The uncertainty on the data can also negatively affect the potential for success of an ML effort. Without proper guidance from underlying physical knowledge, training models on noisy data may return nonsensical results.

In this contribution, the effects of working with a low-quality dataset are illustrated. The importance of properly accounting for the uncertainties on the data is highlighted. The dataset of choice consists of measurements of compression wave velocity V_p and shear wave velocity V_s on sediments from the Southern North Sea with P-S logging. CPT-based correlations for V_p are not widely available but have applications in seismic inversion (Dalgaard et al., 2024). A simplified analytic model is compared to a purely data-driven method to highlight the differences between both approaches.

2 P-S logging measurements

P-S logging is a measurement technique for determining the elastic properties of the soil surrounding an uncased geotechnical borehole. Although determining the shear wave velocity with the seismic CPT is more frequently employed (Stuyts et al., 2024), P-S logging is a technique that allows the measurement of elastic properties in layers where the CPT refuses (e.g. very dense sands or weak rock). The measurement consists of lowering a long probe into an uncased borehole with a supporting drilling fluid and triggering a seismic source that is acoustically separated from two geophone sets positioned on the probe at a given offset (Figure 1). When the source is activated, the travel time of P- and S-waves through the soil can be inferred from the measurements.
Since the quality of the measurement depends on the borehole deformation during the test, caliper and natural gamma ray measurements are often also performed in the uncased borehole.

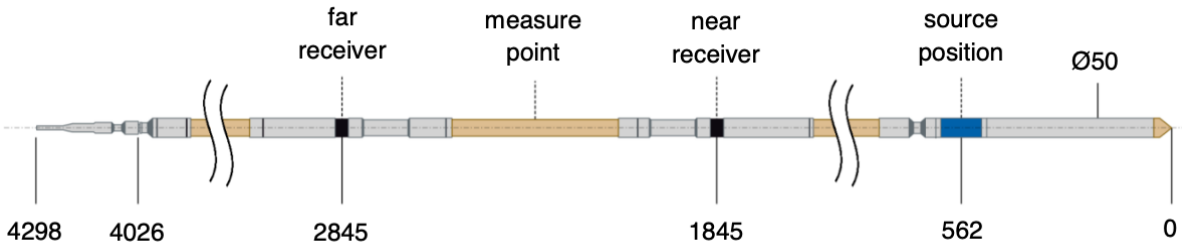

Figure 1. P-S logging probe with dimensions from the bottom of the probe in mm (Source:pinta.bsh.de)

3 Dataset

3.1 Overview

The P-S logging dataset consists of measurements at 8 offshore wind farm project sites with open geotechnical data from Germany, the Netherlands and Belgium (Bundesamt für Seeschifffart und Hydrographie, 2025; Netherlands Enterprise Agency, 2025; FOD Economie, 2023). Only P-S logging boreholes with CPT data available within 100m are selected for combining CPT and P-S logging data. The CPT and P-S logging measurements are not performed at exactly the same location and therefore there is uncertainty on the similarity of soil conditions between the P-S logging and CPT location. While no sudden geological variations such as faults were noticed at the project sites, lateral variability may still play a role.

The combination of P-S logging and CPT data results in a total of 1899 measurements in 65 boreholes. The soils are predominantly cohesionless with varying degrees of overconsolidation. The only exception is the Belgian Princess Elisabeth Zone where overconsolidated Tertiary clays dominate. A soil behavior type index I_c according to Robertson (2009) was assigned to each data point based on the collocated CPT.

Figure 2 shows an overview of the dataset. The scatter in the data is immediately obvious. Nevertheless, cohesive soils ($I_c > 2.7$) result in V_p values close to 1500m/s near the surface and show a slight increase with depth. For cohesionless soils ($I_c < 2.3$) the V_p values are higher. They increase with increasing I_c and also show an increase with vertical effective stress. In any case, the V_p for water (1450m/s) forms a lower bound.

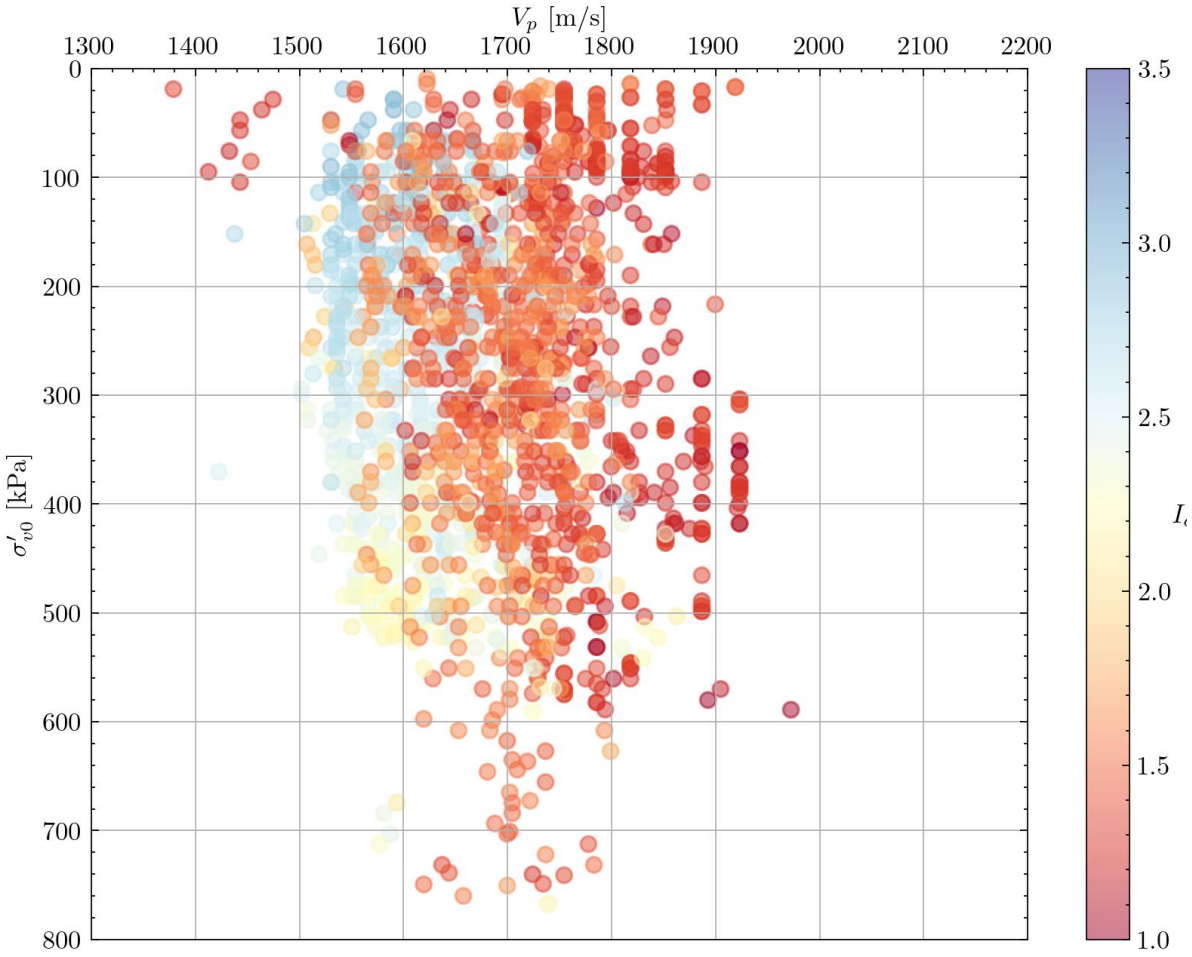

Figure 2. Compression wave velocity as function of vertical effective stress and soil behavior type index

3.2 Interpretation uncertainty

While Figure 2 presents V_p values, it should be noted that these data are interpreted. A human agent picks the arrival time of the compression wave from the recorded waveform. The data quality has a major impact on the certainty of the picked arrivals. As the P-S logging has two receivers, the signals of the near and far receiver should both show a clear first arrival. When multiple recordings are performed at a given depth, the arrival times

should be consistent. Usually, the human agent can judge the quality of the first arrival picking in a qualitative manner. This information is not always recorded, making it impossible to distinguish between low- and high-quality data. Moreover, quality criteria for P-S logging processing are not uniform across the different contractors. A universally adopted quality scoring metric would add value to a ML workflow as high-quality data could be given higher weights in any subsequent assessment.

3.3 Test method uncertainty

Even though the seismic CPT cannot measure V_p, it can be instructive to compare V_s measurements from P-S logging to those obtained with the seismic CPT. For the IJmuiden Ver and Nederwiek wind farm zones in the Netherlands, there are a number of P-S logging boreholes with collocated seismic CPTs. Figure 3 shows that V_s from seismic CPT tends to be higher than that from P-S logging at the same depth. A possible reason for this is that insertion of the CPT rods leads to a stress increase in the soil, with associated increases of the stress-dependent stiffness. Borehole drilling, on the other hand, will relax the stresses in the soil, leading to stiffness reduction.

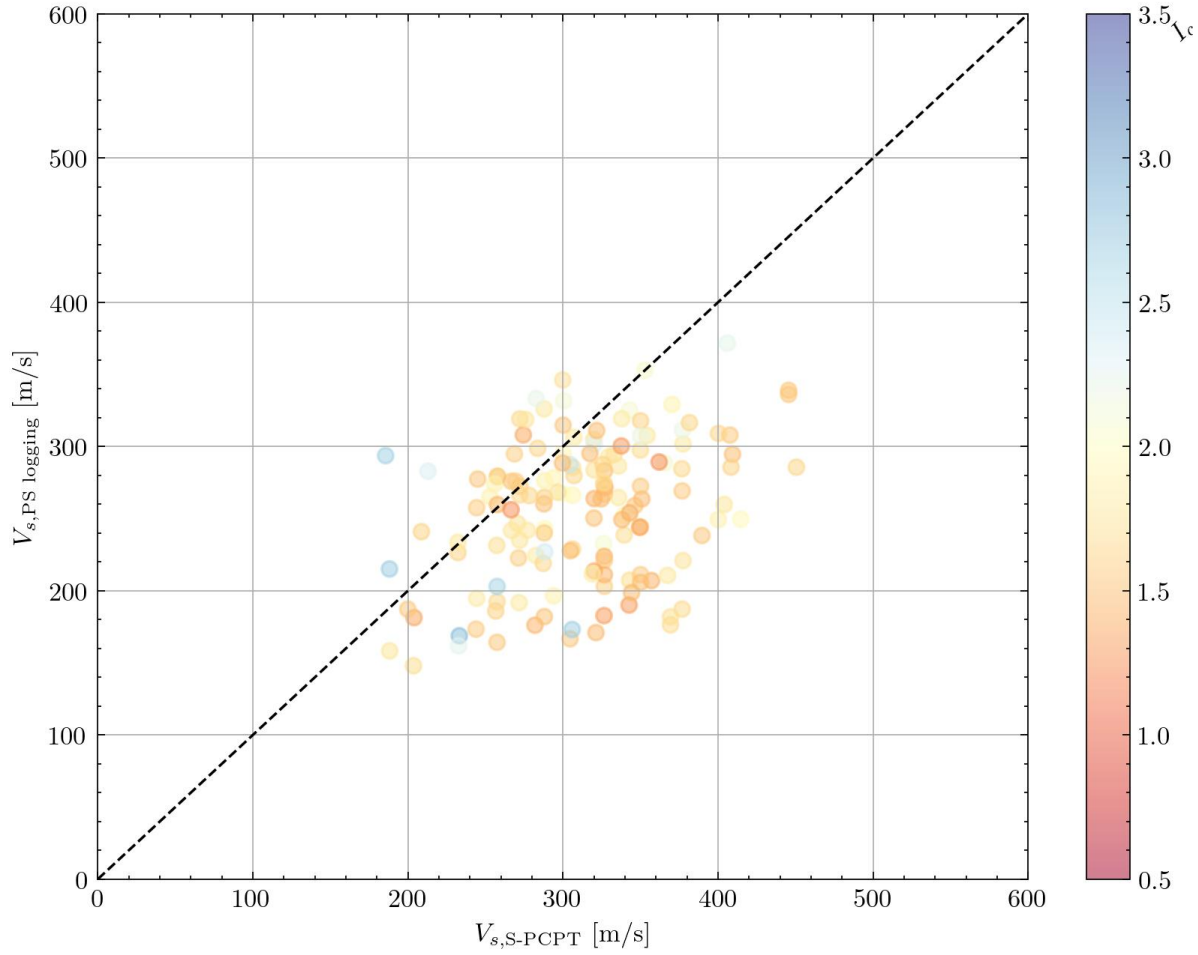

Figure 3. Comparison of V_s measurements with the seismic CPT and P-S logging

The measurements of borehole deformation with the caliper log can be examined to check the impact of borehole deformation on the measured elastic properties. Figure 4 shows that both expansion (caliper measurement > 228mm) and contraction (caliper measurement < 228mm) of the borehole were recorded. Although a clear trend of the ratio $V_{s,S-PCPT}/V_{s,P-S\ logging}$ with borehole deformation could not be established, the ratio of V_p/V_s (both from P-S logging) does show some coherence. For increasing borehole contraction, the range of observed V_p/V_s ratios shrinks. For an expanded borehole, a wider range of V_p/V_s is noticed.

Based on these observations, it is clear that there is significant epistemic uncertainty on the relation between borehole deformation and the value of V_p and V_s. The true underlying trends are poorly understood and are not observable from the data. Any ML model that will aim to capture the effect of borehole deformation will not be able to identify clear trends from the data.

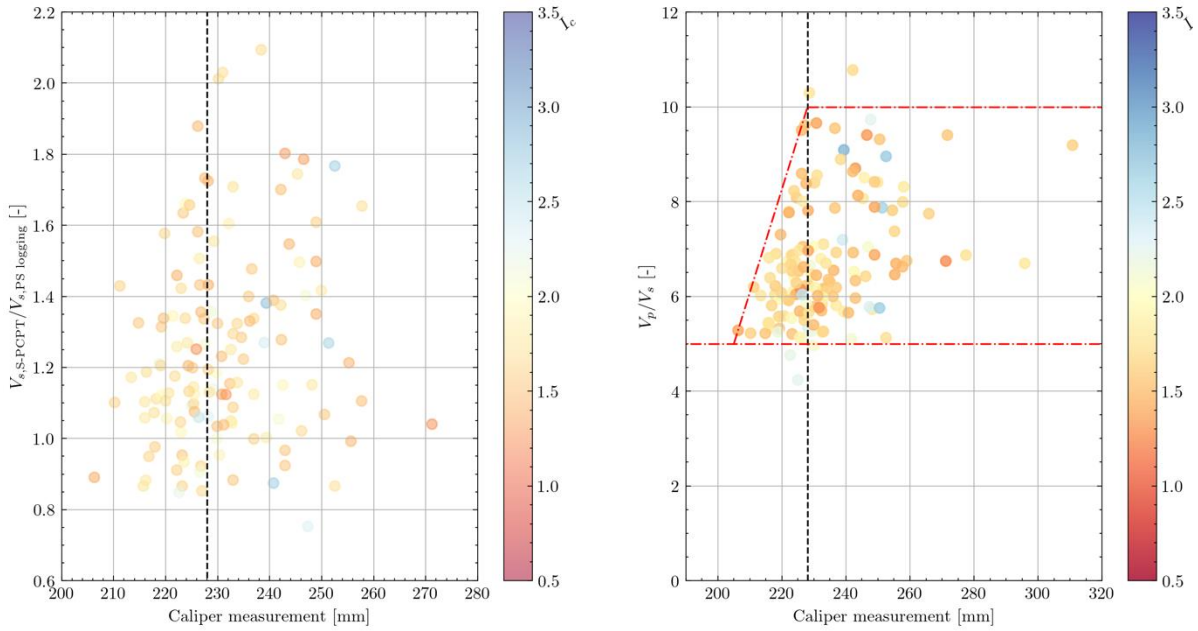

Figure 4. Variation of the ratio $V_{s,S-PCPT}/V_{s,P-S\ logging}$ and V_p/V_s with borehole deformation (diameter of drillbit is 228mm).

4 Model building

4.1 Simplified analytical model

Based on the observations from Figure 2, a simplified model was built that interpolates between two linear trends. Figure 5 shows this model where the black dashed line shows the average trend for cohesive soil ($I_c = 3$) and the purple dash-dotted line shows the average trend for sandy gravel ($I_c = 1$). Interpolation between the two lines is performed based on the value of the soil behavior type index. The model is greatly simplified but it does capture the observed general trends.

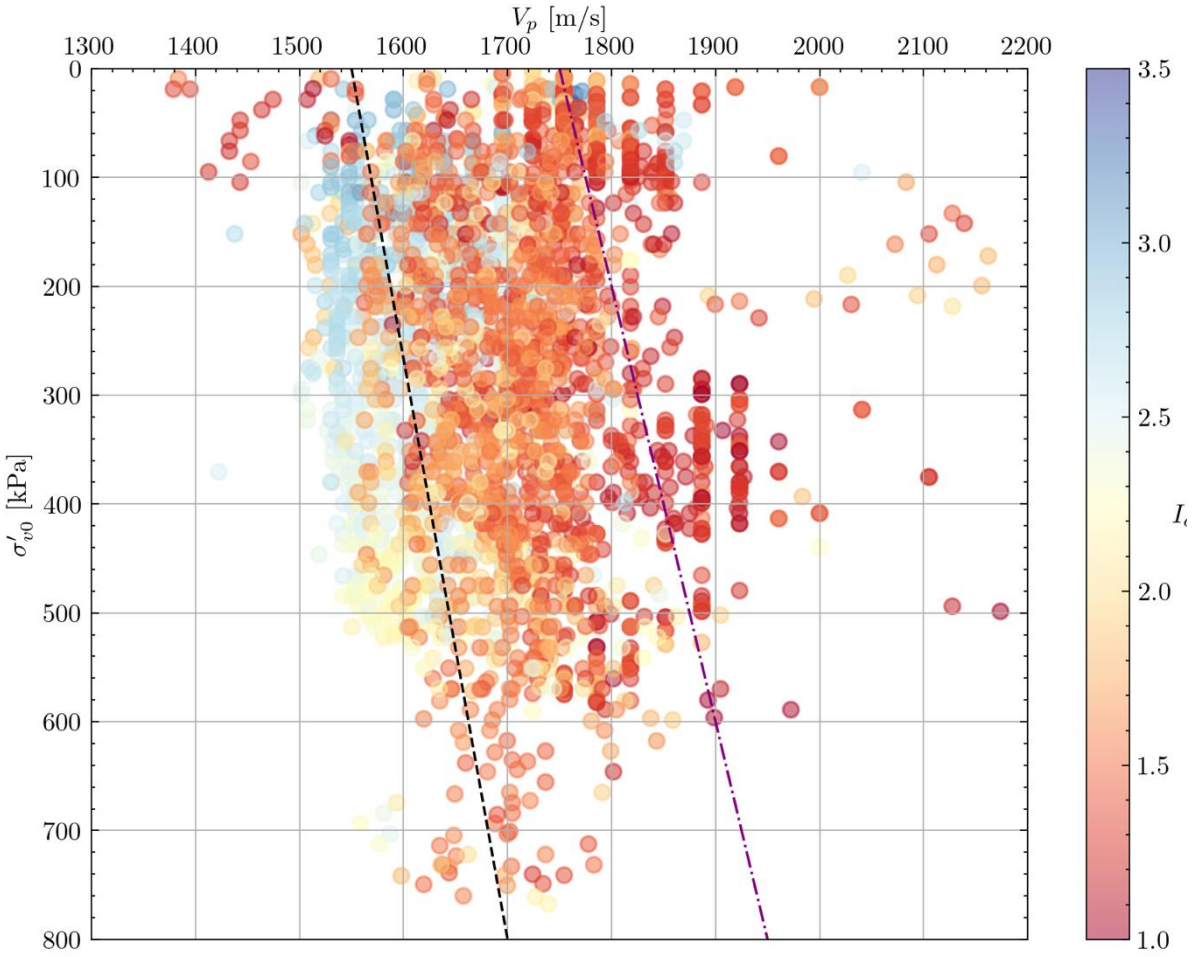

Figure 5. Simplified correlation capturing the dependence of V_p on vertical effective stress and I_c.

The performance of the correlation was checked against the available data. Figure 6 shows a histogram of the ratio of calculated to measured V_p together with a scatterplot of calculated and measured V_p (the black line represents a perfect prediction). The correlation shows nearly neutral bias and a coefficient of variation of 10%. The R^2-value was very low (0.03) showing that the model does not capture the underlying variability. As a lot of the variability in the dataset is the consequence of the uncertainty on the data, high R^2 scores would be hard to achieve anyway.

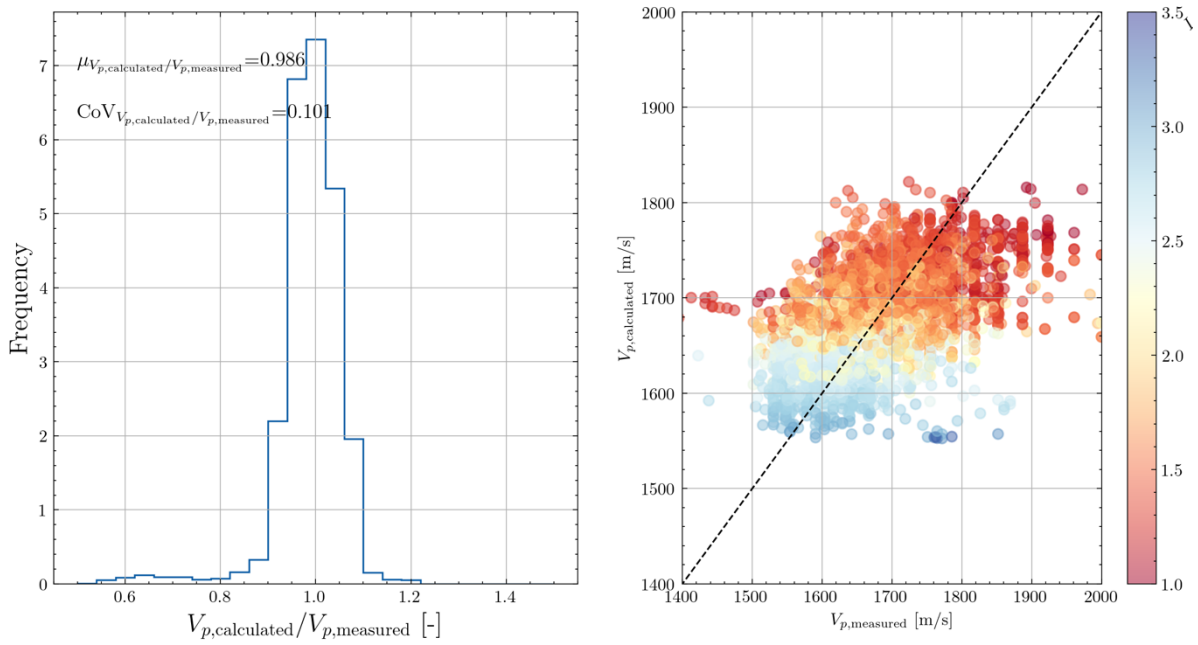

Figure 6. Performance of the simplified correlation for V_p.

It is also interesting to check the dependence of the ratio of calculated to measured V_p on the input feature values. Ideally, the mean should be stationary, and the variance should not change with the feature values. Figure 7 shows that this is broadly the case for the proposed correlation.

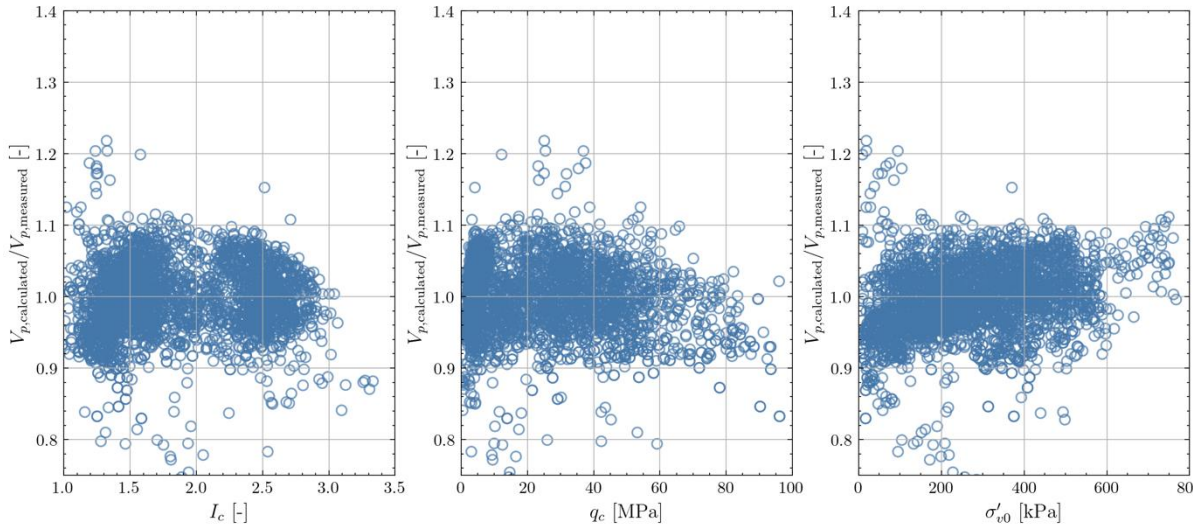

Figure 7. Bias and scatter of the simplified correlation for V_p.

4.2 Machine learning model

Machine learning can also be used to train models based on the available CPT and V_p data. The cone tip resistance q_c, normalized cone tip resistance q_{c1N}, soil behavior type index I_c and vertical effective stress σ'_{v0} are used as features. The compression wave velocity V_p is the target. The data were split into a training and test set using a randomized 80%/20% split.

An XGBoost regression model was trained on the data with tree stubs (maximum tree depth of 2). The number of estimators was set at 200 and a learning rate of 0.1 was selected.

The performance of the ML model is shown in Figure 8, with accuracy metrics summarized in Table 1. The histogram reveals a mean and coefficient of variation comparable to those of the analytical model (Figure 6). However, the scatterplot of predicted versus measured V_p indicates slightly greater dispersion for the ML model.

The R^2 score is higher on the training set than on the test set, indicating there is a degree of overfitting. It is also debatable whether the variance in the dataset can actually be meaningfully represented by a model.

Table 1. XGBoost model performance.

Data	$\mu_{V_{p,calculated}/V_{p,measured}}$ (-)	$CoV_{V_{p,calculated}/V_{p,measured}}$ (-)	R^2 (-)
Train	1.004	0.07	0.50
Test	1.002	0.12	0.40
All data	1.004	0.09	0.47

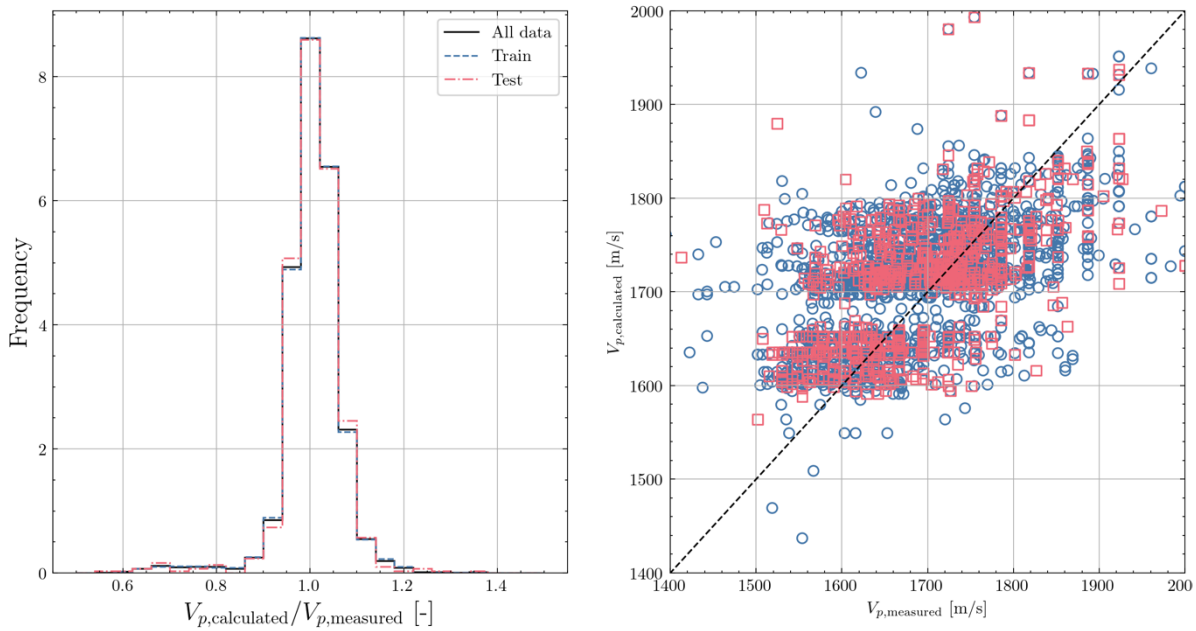

Figure 8. XGBoost model for V_p performance of the train and test sets.

Based on the results shown in Figure 6 and Figure 8, both the simplified analytical model and the ML model seem to have comparable predictive power. To further verify its performance, the XGBoost model was applied to two example soil profiles:

- A uniform clay profile with $I_c = 3$ and $q_{c1N} = 20$. The remaining feature q_c can be derived when the vertical effective stress and normalized cone resistance are known
- A uniform dense sand profile with $I_c = 1.3$ and $q_{c1N} = 200$

The V_p predictions as a function of vertical effective stress are checked. The remaining feature q_c can be derived when the vertical effective stress and normalized cone resistance are known.
The predicted lines (Figure 9) show that reasonable predictions are made between 120kPa and 560kPa. For lower and higher stress levels, the predictions become unreliable, and the model cannot be used there.
This highlights the need for thorough model evaluation, including checks on the performance for generic feature combinations. Without such checks, the model cannot be used with confidence.

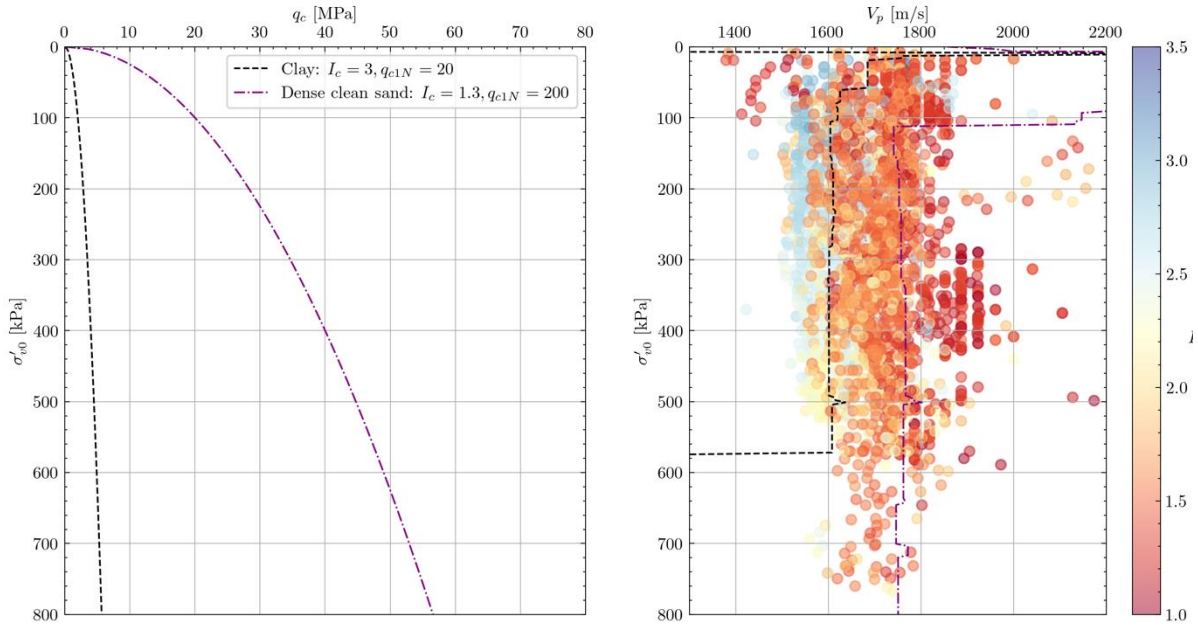

Figure 9. XGBoost model predictions for two example soil profiles.

5 References

Bundesamt für Seeschifffart und Hydrographie, 2025. Data Hub Preliminary Investigation -of Sites [WWW Document]. URL https://pinta.bsh.de/

Dalgaard, E., Hansen, H., Horn, F., Jakobsen, A., Kuppens, S., Stuyts, B., 2024. Energy Island-CPT Interpretation and Soil Classification from UHRS Pre-Stack Data, in: Fifth EAGE Global Energy Transition Conference & Exhibition (GET 2024). European Association of Geoscientists & Engineers, pp. 1–4.

FOD Economie, 2023. Digital Database Princess Elizabeth Zone.

Netherlands Enterprise Agency, 2025. Wind Farm Zones [WWW Document]. URL https://offshorewind.rvo.nl/

Robertson, P.K., 2009. Interpretation of cone penetration tests-a unified approach. Canadian Geotechnical Journal 46, 1337–1355.

Stuyts, B., Weijtjens, W., Jurado, C.S., Devriendt, C., Kheffache, A., 2024. A Critical Review of Cone Penetration Test-Based Correlations for Estimating Small-Strain Shear Modulus in North Sea Soils. Geotechnics 4, 604–635. https://doi.org/10.3390/geotechnics4020033

Terzaghi, K., Peck, R.B., Mesri, G., 1996. Soil mechanics in engineering practice - Third Edition. John Wiley & Sons.