A Data-Driven Approach to Predict Shear Wave Velocity from CPTu Measurements: An Update

Iman Entezari[1#], *James* Sharp[1], and *Paul W.* Mayne[2]

[1]*ConeTec Group, Burnaby, Canada*
[2]*Georgia Institute of Technology, Atlanta, Georgia USA*
[#]*Corresponding author: ientezari@conetec.com*

ABSTRACT

This paper is an extension of our previous research investigating the potential of machine learning models to estimate shear wave velocity (V_s) from piezocone penetration test (CPTu) measurements. The aim of this update is to examine the effect of incorporating geographical information, namely latitude and longitude, as input parameters to the machine learning models. New models are developed by incorporating both CPTu parameters and spatial coordinates as input features and are compared to models developed with only CPTu parameters. Furthermore, SHAP (SHapley Additive exPlanations) analysis is employed to assess the importance of different features and variables in the developed machine learning models. The results show improvement in prediction performance when adding geographical data, indicating the influence of geological variations on V_s. The paper shows the potential of using geospatial information to improve the data-driven approach for estimating soil properties from CPTu tests when large worldwide datasets are available.

Keywords: Cone Penetration Test; CPTu; SCPTu; Shear Wave Velocity; Machine Learning.

1. Introduction

The assessment of shear wave velocity (V_s) in soil deposits is a crucial element in the domain of geotechnical earthquake engineering. Field-based seismic geophysical tests or empirical correlations may be used to measure or estimate V_s. Empirical correlations offer a practical means to estimate V_s from in-situ tests such as the piezocone penetration test (CPTu). Studies such as Hegazy and Mayne (1995), Mayne (2006), Robertson (2009), Andrus et al. (2007), and McGann et al. (2015) have proposed and examined various empirical correlations to predict V_s based on these in-situ tests.

Recently, machine learning (ML) techniques have emerged as a powerful tool for data analysis and prediction in various fields, including geotechnical engineering. ML methods can learn complex and nonlinear patterns from large datasets without relying on predefined assumptions or equations. Several researchers have applied ML methods to estimate various soil properties from CPTu data (e.g. Wang et al. 2019, Xiao et al. 2021, Entezari et al. 2021 & 2022, Assaf et al. 2023). These studies have shown that ML methods can outperform traditional empirical correlations.

Geographical information, such as latitude and longitude, can capture the spatial variations and trends of soil properties across different regions and locations. This information can be useful for enhancing the generalization and robustness of the ML models, especially when dealing with large and diverse datasets. Furthermore, geographical information can provide insights into the underlying factors and mechanisms that influence the relationship between CPTu and V_s, such as soil type, depositional environment, age, cementation, and weathering. Therefore, it is important to investigate the potential of geographical information in the ML models for estimating V_s from CPTu data.

In our prior research (Entezari et al., 2022), a substantial dataset comprising more than 100,000 paired V_s-CPTu observations derived from seismic piezocone (SCPTu) soundings conducted worldwide was utilized to develop machine learning models for the prediction of V_s based on CPTu data. The machine learning models were exclusively trained using fundamental CPTu parameters, including corrected tip resistance (q_t), sleeve friction (f_s), dynamic porewater pressure (u_2), and depth (z). The performance of the models was assessed with a specific focus on the influence of soil microstructure. In this study, the impact of adding latitude and longitude as additional features to the models is assessed. Furthermore, SHAP (SHapley Additive exPlanations) analysis is used to investigate the importance of each feature on the predicted results.

2. Dataset description

The dataset used in this study is the same dataset used by Entezari et al. (2022) updated with new data as well as latitude and longitude of the test results compiled using the ConeTec's geospatial database. The dataset includes more than 180,000 paired V_s with q_t, f_s, u_2, z, latitude, and longitude datapoints collected from 2017 to early 2023 compiled from 10,386 individual SCPTu soundings across diverse soil types, stress histories, and geological environments worldwide. The geographical distribution of the dataset is shown in Fig. 1, where the density of the data is represented by the heatmap.

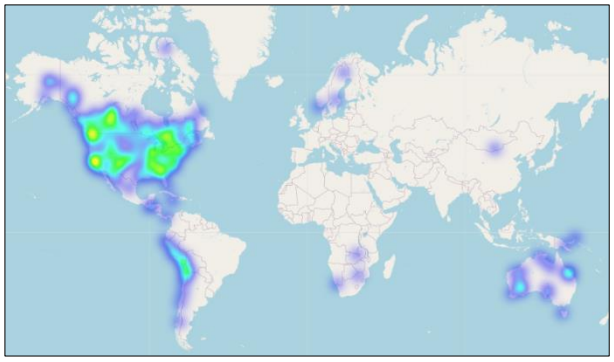


Figure 1. A heatmap representation of the dataset showing the relative distribution and density of the data pairs.

The dataset is split into training and test sets. The training set is used to calibrate the model whereas the test set is used to evaluate the model performance. The data collected in 2017- 2021 timeframe is used as the training set and data collected from 2022 to early 2023 provides the test set. This allows for an unbiased performance assessment and avoids leakage of information which is specifically important to assess the performance of the models with latitude and longitude as input features. The number of paired V_s-CPTu data points for the training and test sets are listed in Table 1.

Fig. 2 depicts the test set on the normalized tip resistance (Q_{tn}) versus small-strain rigidity index (I_G) plot, showing the soils classified as cemented and uncemented based on the empirical K_G^* threshold of 330 (Robertson 2016). In this study, the impact of soil microstructure on the ML models is investigated through developing models for three categories: all-soils, uncemented, and cemented soils.

Table 1. Number of data pairs in the training and test sets.

	All	Uncemented ($K_G^* \leq 330$)	Cemented ($K_G^* > 330$)
Training set	142,114	85,285	56,829
Test set	39,028	24,012	15,016

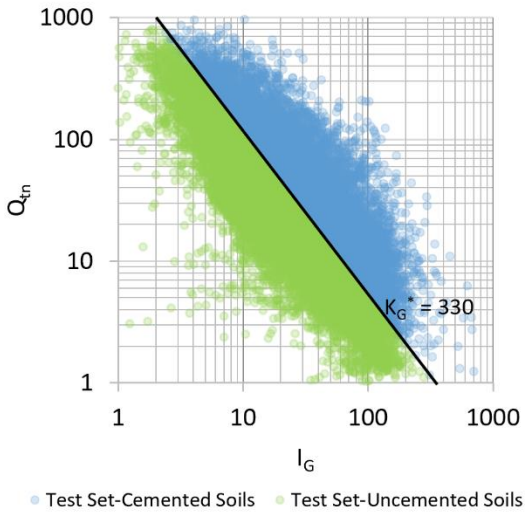


Figure 2. The test set plotted in the Q_{tn}-I_G chart.

3. ML modelling

3.1. XGBoost

The ML algorithm used in this study to develop models for predicting V_s from CPTu data is XGBoost (Chen and Guestrin 2016). XGBoost is a scalable and efficient implementation of gradient boosting trees that has gained popularity and success in various data science applications. XGBoost can handle both regression and classification tasks and offers several advantages such as regularization, parallelization, distributed computing, and missing value handling.

In this study, XGBoost is used to develop both sets of models, the ones with only basic CPTu parameters (q_t, f_s, u_2, and z) and the ones with adding latitude and longitude as additional features. Microsoft Azure AutoML (2023) is employed to optimize hyperparameters of the XGBoost models such as the learning rate, number of trees, tree depth, and the minimum loss reduction required to make a further partition. The optimization process is conducted through the widely used k-fold (k = 5) cross-validation technique (Kohavi 1995) using the training set.

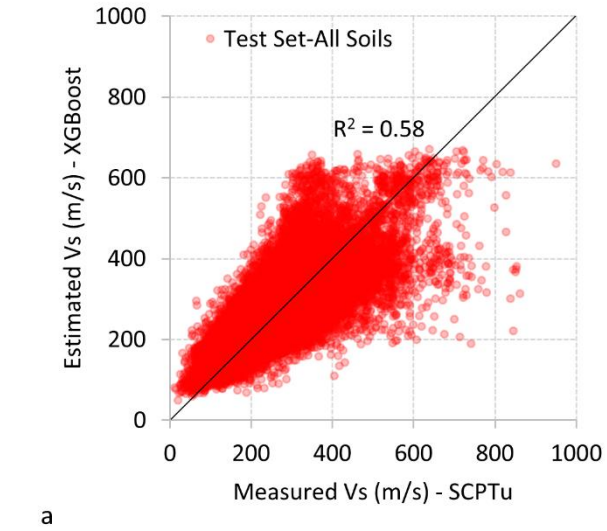
3.2. Performance assessment

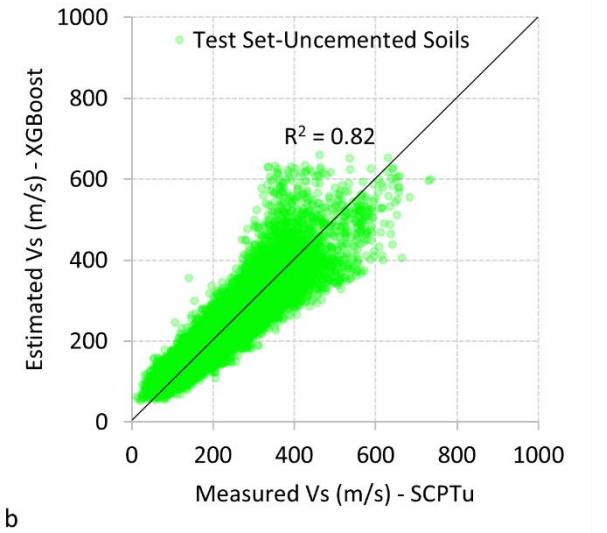
The evaluation of the performance of the developed models involves the examination of cumulative distribution function (CDF) of errors on the test set. Error calculation is based on the disparity between the measured V_s obtained from the SCPTu and the V_s predicted by the XGBoost models. The bias of the prediction is determined at the 50th percentile in the CDF. Assuming normal distribution of errors, the CDF values at 15.9% and 84.1% correspond to ±1 standard deviation. The average of these CDF values at 15.9% and 84.1% is adopted as the comprehensive error measure for the model.

4. Results

4.1. Models with basic CPTu

The test set results of the XGBoost models that use basic CPTu parameters to predict V_s from SCPTu are shown in Figs. 3a to 3c for all-soils, uncemented soils, and cemented soils, respectively. The all-soil model uses all the data pairs in the training set, while the models specific to uncemented and cemented soils are developed using their respective fractions within the training set. The error analysis using CDF of errors on the test set shows that the errors of the estimated results are ±55.53, ±32.31, and ±51.22 (m/s) for all-soils, uncemented, and cemented soils, respectively (Table 2). Evidently, the uncemented soil model outperforms both the all-soils and cemented soils models.





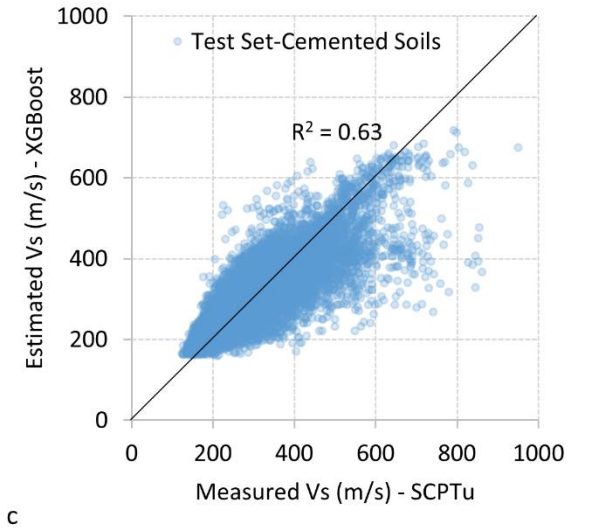
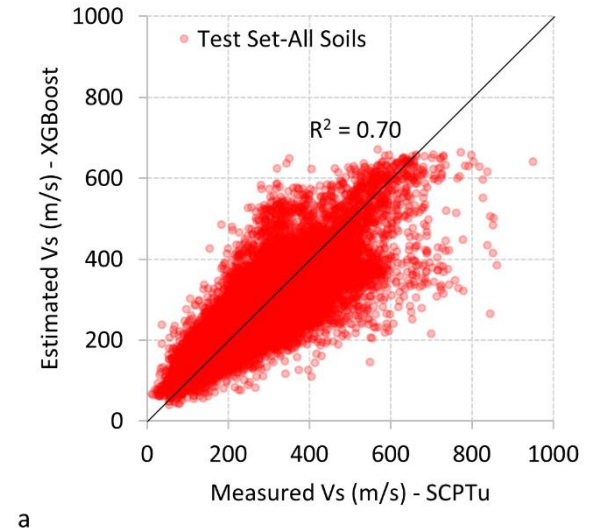
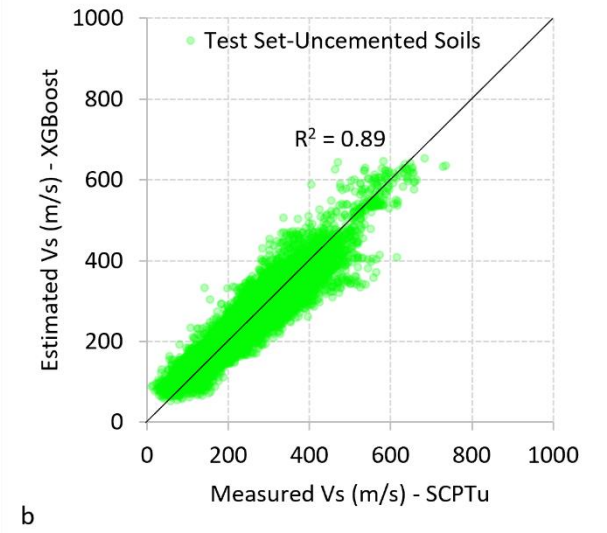


Figure 3. Relationship between measured and estimated V_s on the test set using (a) all-soils, (b) uncemented soils, and (c) cemented soils models developed with basic CPTu parameters.





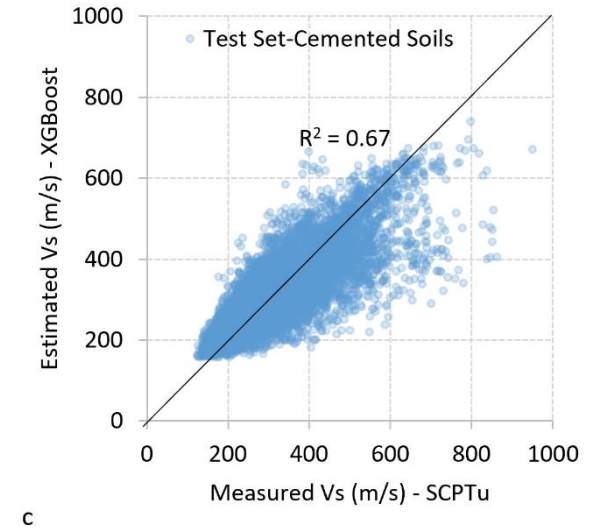


Figure 4. Relationship between measured and estimated V_s on the test set using (a) all-soils, (b) uncemented soils, and (c) cemented soils models developed with basic CPTu parameters along with latitude and longitude.

Table 2. Performance of different models.

	Bias±Error (m/s)	R^2
Models with q_t, f_s, u_2, z		
All Soils	-15.39±55.53	0.58
Uncemented	2.17±32.31	0.82
Cemented	-14.8±51.22	0.63
Models with q_t, f_s, u_2, z, lat, lon		
All Soils	-4.73±41.9	0.70
Uncemented	2.66±26.48	0.89
Cemented	-6.44±46.32	0.67

4.2. Models with the inclusion of latitude and longitude

Figs. 4a to 4c illustrate the performance of models developed with both basic CPTu data and geographical coordinates (latitude and longitude) on the test set. By analyzing the CDF of errors on the test set, the overall errors are found to be ±41.9, ±26.48, and ±46.32 for all-soils, uncemented, and cemented soils, respectively. Similar to the models developed solely with basic CPTu parameters, the uncemented soil model exhibits superior performance compared to the all-soils and cemented soils models.

An enhancement in performance across all three models is observed through a comparative analysis between models incorporating geographical coordinates and those relying exclusively on the basic CPTu parameters. Although the overall improvement may not be statistically significant, the integration of geographical coordinates appears to contribute to a nuanced refinement of the estimated results. Analyzing the outcomes presented in Figs. 3 and 4, a notable reduction in the dispersion of data points is observed, particularly within the high shear wave velocity region (V_s > 350 m/s), when comparing models developed with the inclusion of latitude and longitude to those developed with only basic CPTu parameters. This effect is more pronounced in the uncemented soil model. This could imply that the inclusion of geographical locations may enhance the estimated results in relation to specific geographic locations.

4.2.1. Applicability of the models with geographical coordinates

While the integration of geographical coordinates has demonstrated enhanced model performance, it is important to acknowledge the inherent sensitivity of ML models to the similarity between their training and test datasets. Generalizations drawn from the improved performance of latitude and longitude-inclusive models are dependent on the availability of comprehensive training data from corresponding geographic regions. Models incorporating geographic coordinates effectively capture regional variations, making them particularly reliable in areas with robust data coverage during the training phase. Conversely, in regions with limited training data, the effectiveness of these models may be compromised. Therefore, the applicability of such models is most reliable within the geographic scope well-represented in the training set.

In the context of this study, it is essential to conduct an analysis to evaluate the proximity of test set data points to those in the training set to identify whether errors and outliers in the predicted results may arise due to the absence of sufficient training data in corresponding geographic areas. This remains to be investigated in future work.

5. SHAP analysis

SHAP analysis is a method to explain the predictions of machine learning models (Lundberg and Lee 2017). SHAP values are based on the concept of cooperative game theory that allocates a contribution to each player in a coalition game. In the context of machine learning, features are considered as players and the prediction is the payoff of the game. In this paper, SHAP analysis is employed to compare the performance and interpretability of the uncemented soils models developed with basic CPTu variables, and with basic CPTu and latitude/longitude variables. SHAP analysis is applied on the test set to examine how the features contribute to the model predictions.

5.1. Feature importance

The SHAP summary plots (Lundberg and Lee 2017, Lundberg et al. 2020) depicting the uncemented soils models are showcased in Fig. 5. In the summary plots, every point represents a SHAP value (impact on model output) corresponding to a feature and a specific instance within the test set. The color gradient from blue to red indicates the feature value, ranging from low to high, respectively. To provide a clear distribution of SHAP values per feature, overlapping points are jittered along the y-axis. The arrangement of features follows their importance, where features with larger absolute SHAP values implies a more pronounced influence on predictions.

As depicted in Fig. 5a, the uncemented soil model developed utilizing basic CPTu data highlights depth (z) as the primary influential feature, succeeded by q_t, f_s, and u_2 in descending order of significance. The observed trends indicate that elevated values of z, q_t, and f_s are generally associated with an increase in the predicted outcomes. The summary plot also shows that u_2 exhibits the least importance among the features, and there is an absence of an apparent pattern regarding its impact on the estimated results.

In the uncemented soil model incorporating latitude and longitude as input features, the significance of features is as follows: z, q_t, latitude, f_s, u_2, and longitude as shown in Fig 5b. Higher values of z, q_t, and f_s correspond to an increase in the predicted results, while elevated latitude values are generally associated with a reduction in the predicted outcomes. Also, u_2 and longitude exhibit minimal impact on the model outcome, with no discernible relationship observed between these features and the model results.

The significance attributed to latitude as an influential feature in contrast to the minimal impact of longitude, could be due to specific characteristics or patterns in the

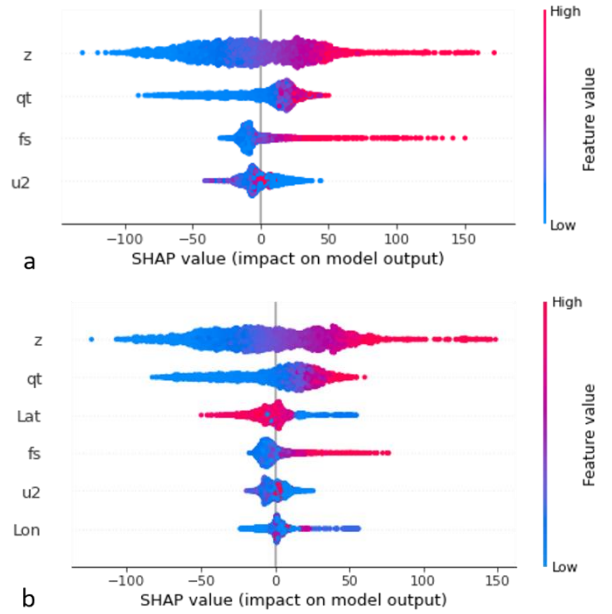


Figure 5. SHAP summary plots for the uncemented soil model developed using (a) basic CPTu parameters, and (b) basic CPTu parameters along with latitude and longitude.

dataset where changes in latitude correspond to significant variations in the predicted results. SHAP partial dependence plots (PDP) presented in next section can explain the details of the relationship between latitude and longitude and the predicted results.

5.2. Partial dependence plots (PDP)

SHAP PDP visually depicts the isolated impact of a single variable on model predictions, holding other variables constant. These plots offer valuable insights into the effects of individual features, aiding in the interpretation of complex ML models.

In Fig. 6, the PDP for z, latitude, and longitude for the uncemented soil model developed with basic CPTu along with latitude and longitude are presented. These plots effectively illustrate how alterations in a specific feature would influence the model predictions. Sample observations are depicted through thin gray lines, showcasing the variability in predictions associated with changes in the feature. The average impact is represented by the thick gray line. The effect of changing the feature for a singular instance is delineated in blue. The vertical dashed blue line denotes the value of the feature for the specified instance, while the horizontal dashed blue line shows the corresponding predicted result for that singular instance. As shown in Fig 6a, there exists a positive correlation between z and the predicted results, signifying that, on average, an increase in depth, while maintaining other variables constant, results in an elevated predicted V_s. Subsequently, Fig. 6b shows that the impact of latitude on predicted results is complex and region-specific. Notably, latitude values proximate to -16.5 degrees exhibit the highest predicted V_s outcomes. Further examination of the dataset identifies this region is characterized by deep water tables. Deeper water tables are associated with an increase in effective stress, resulting in elevated V_s values. In essence, soils in this specific region, with equivalent depths to those in another area, experience higher effective stress, leading to a subsequent rise in predicted V_s values. Therefore, it is anticipated that, for this particular region, depth has a more pronounced impact on the estimated results compared to soils located at other latitudes. The PDP for longitude, as depicted in Fig. 6c, shows a generally negligible impact on the predicted results. The near-flat average impact curve underscores the insignificant influence of longitude on the predicted outcomes.

To investigate the interaction between depth and latitude, the SHAP dependence plot for depth is presented in Fig. 7. Notably, data points located near latitude -16.5 are distinguished by orange color. These specific points exhibit a distinct behavior, indicating an increased influence of depth on the estimated results when compared to the remainder of the dataset. This distinctive pattern may be attributed to deeper water tables and thus higher effective stresses in the corresponding region, as previously discussed.

It should be noted that relationships observed associated with the impact of latitude and longitude on the predicted results are specific to the dataset employed in this research. While the dataset encompasses global information, the limited data availability from certain regions restricts the generalizability of the conclusions drawn in relation to the impact of geographical location on the V_s estimation – instead it is a function of where data are available in the dataset.

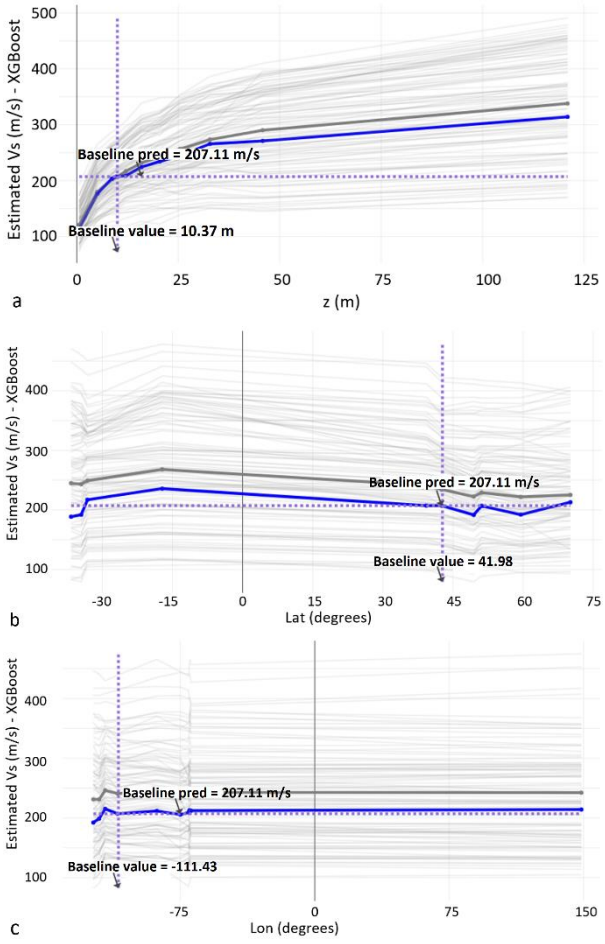


Figure 6. SHAP partial dependence plots to show the impact of (a) depth, (b) latitude, and (c) longitude on the estimated V_s results for the uncemented soil model developed using basic CPTu parameters along with latitude and longitude.

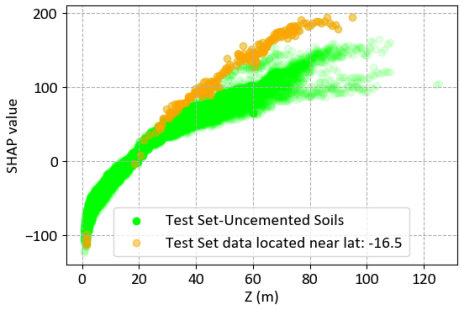


Figure 7. SHAP dependence plot to exhibit the impact of depth on the model output.

6. Conclusions

In this extended study, we built upon our previous research exploring the potential of ML models to estimate V_s from CPTu measurements. The focus of this update was to investigate the impact of incorporating geographical information, specifically latitude and longitude, as additional input parameters to the ML models. The newly developed models, featuring both CPTu parameters and spatial coordinates, were compared with models based solely on CPTu parameters.

Performance assessment on an extensive test set of nearly 40,000 data pairs revealed that models with inclusion of latitude and longitude enhance the predictions results. While the overall enhancement was observed to be marginal, it was more pronounced on the prediction results of specific regions. Furthermore, SHAP analysis was employed to study the importance of each input parameter on the model output. The SHAP analysis offered valuable insights into the contribution of each feature to the model predictions. Notably, for uncemented soils model, depth emerged as a primary influential parameter, because in this study depth is used as a proxy for in-situ effective stress. Moreover, the influence of latitude on predicted outcomes exhibited region-specific patterns, whereas longitude demonstrated minimal impact, ranking as the least consequential feature in the model.

While the integration of geographical coordinates demonstrated improved model performance, the reliability of models depends on the availability of comprehensive training data from corresponding geographic regions. Applicability is most reliable within well-represented geographic scopes, necessitating a proximity analysis to assess data representation and potential errors in less-covered areas.

This approach holds promise for geotechnical applications, underscoring the relevance of spatial context in soil property assessments. As this research progresses, integrating geospatial information into ML models emerges as a compelling avenue for advancing geotechnical predictions and furthering the comprehension of soil behavior, particularly when leveraging large and diverse datasets on a global scale. The ML models can be exported to standalone applications or integrated into commercial software packages for utilization within the geotechnical community.

References

Assaf, J., Molnar, S., El Naggar, M.H. 2023. "CPT-V_s correlations for post-glacial sediments in Metropolitan Vancouver," Soil Dynamics and Earthquake Engineering, Vol. 165.

Andrus, R.D., Mohanan, N.P., Piratheepan, P., Ellis, B.S., and Holzer, T.L. 2007. "Predicting shear wave velocity from cone penetration resistance," Proceedings of the 4th International Conference on Earthquake Geotechnical Engineering, Thessaloniki, Greece.

Chen, T. and Guestrin, C. 2016. "Xgboost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August: 785–794.

Entezari, I., Sharp, J., and Mayne, P.W. 2021. "Soil unit weight estimation using the cone penetration test and machine learning," Proceedings of GeoNiagara 2021, Niagara Falls, Canada.

Entezari, I., Sharp, J., & Mayne, P.W. 2022. "A data-driven approach to predict shear wave velocity from CPTu measurements," Cone Penetration Testing 2022 (Proc. CPT'22, Bologna), CRC Press, Leiden: 374–380.

Hegazy, Y.A. & Mayne, P.W. 1995. "Statistical correlations between V_s and cone penetration data for different soil types," Proceedings of CPT '95, Linkoping, Sweden, Vol. 2: 173–178.

Kohavi, R.1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection," International Joint Conference on Artificial Intelligence (IJCAI), Vol. 14. Montreal, Canada, 1137–1145.

Lundberg, S.M., Lee, S.I. 2017. "A unified approach to interpreting model predictions," Proceedings of Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December: 4765–4774.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. 2020. "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, 2: 56-67.

Mayne, P.W. 2006. "In-situ test calibrations for evaluating soil parameters," Characterization and Engineering Properties of Natural Soils II, Vol. 3, Taylor & Francis, London: 1601-1652.

McGann, C.R., Bradley, B.A., Taylor, M.L., et al. 2015. "Development of an empirical correlation for predicting shear wave velocity of Christchurch soils from cone penetration test data," Soil Dynamics and Earthquake Engineering, 75: 66–75.

Microsoft. 2023. "What is automated machine learning (AutoML)?" https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml?view=azureml-api-2.

Robertson, P.K. 2009. "Interpretation of cone penetration tests – a unified approach," Canadian Geotechnical Journal, 46(11): 1337–1355.

Robertson, P.K. 2016. "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system-an update," Canadian Geotechnical Journal, 53: 1910-1927.

Wang, H., Wang, X., Wellmann, J.F., and Liang, R.Y. 2019. "A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data," Canadian Geotechnical Journal, 56: 1184–1205.

Xiao, T., Zou, H.F., Yin, K.S., et al. 2021. "Machine learning-enhanced soil classification by integrating borehole and CPTU data with noise filtering," Bulletin of Engineering Geology and the Environment, 80: 9157–9171.