Data-Driven, Reliable Translation of Shear-Wave Velocity to CPT Cone-Tip Resistance Using Machine Learning

Revelo Obando, E.; Ghose, R.; Hicks, M.

DOI
10.3997/2214-4609.202420112

Publication date
2024

Document Version
Final published version

Citation (APA)
Revelo Obando, E., Ghose, R., & Hicks, M. (2024). *Data-Driven, Reliable Translation of Shear-Wave Velocity to CPT Cone-Tip Resistance Using Machine Learning*. Paper presented at Near Surface Geoscience 2024, Helsinki, Finland. https://doi.org/10.3997/2214-4609.202420112

Data-driven, reliable translation of shear-wave velocity to CPT cone-tip resistance using machine learning

E. Revelo Obando[1], R. Ghose[1], M. Hicks[1]

[1] Dept. of Geoscience and Engineering, Delft University of Technology

Summary

The absence of information on lateral variability in the soil is detrimental to estimating accurately the local site response in the event of an earthquake. To address this problem, the use of densely sampled seismic data together with sparsely distributed but detailed vertical soil profiles obtained from cone penetration tests (CPTs) is advantageous. This study explores the adaptation of suitable machine learning (ML) approaches to derive reliable, site- and depth-specific correlations between seismic shear-wave velocity (Vs) and cone-tip resistance (qc). Such correlation could be successfully established by combining information from seismic CPT surveys with available borehole information for the Groningen region in the Netherlands. It is found that, even over substantial distances, ML-based techniques offer site- and depth-specific correlations between Vs and qc.

Data-driven, reliable translation of shear-wave velocity to CPT cone-tip resistance using machine learning

Introduction

Cone penetration testing (CPT) is one of the most widely used methods to obtain the geotechnical engineering properties of soils and delineating the soil stratigraphy. Cone-tip resistance (q_c) is used to obtain the undrained shear strength of saturated cohesive soils and the friction angle of sands. CPTs provide very detailed soil variability information in the vertical direction. CPT information is crucial in designing foundations, assessing the risk of soil liquefaction, and understanding the bearing capacity of the soil, among others. However, CPTs are often sparsely located in the lateral direction, and the interpolation between two locations is generally inaccurate due to the lack of information.

Numerous past studies have shown that seismic shear-wave velocity (V_s) and q_c exhibit correlation with each other in the near-surface soils. This correlation greately improves when it is made in a layer-specific manner. In other words, if the values of V_s and q_c of different layers are used together, the observed correlation is much inferior. Using the available SCPT database for the Groningen region of the Netherlands, we also found a good depth-specific correlation between V_s and q_c in many locations, as illustrated in Fig. 1. High q_c generally corresponds to high V_s values. Because high-resolution seismic methods, like those involving full-waveform inversion, can potentially provide detailed and reliable distribution of V_s in the near-surface soil layers, finding a data-driven approach to correlate V_s to q_c will allow interpolation of the q_c value in between two CPT locations, and finally provide a reliable 2D and 3D spatially continuous field of CPT q_c. This promises a potential breakthrough by decreasing uncertainty in geotechnical design and safety assessments, when this uncertainty is caused by the lack of knowledge due to incomplete site-investigation data. Also, insufficient knowledge regarding spatial variability of q_c inevitably leads to increased conservatism in design and mitigation, and thereby to increased costs.

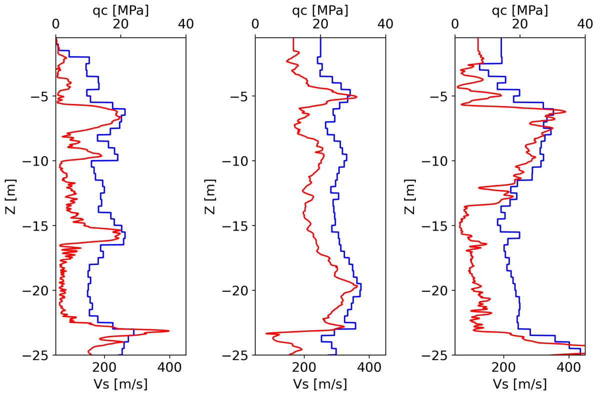

Figure 1 *SCPTs from Groningen, the Netherlands. The blue lines denote V_s values, and the red lines are the q_c values at the same location.*

The relationship between V_s and q_c is nonlinear, and it is governed by a complex combination of the effects of multiple, local (layer/depth-specific) soil properties. This hinders the possibility of a physics-driven translation of V_s to q_c. In the present research, we attempt to derive site- and depth-specific correlations between V_s and q_c through developing novel approaches utilizing machine learning (ML). Some early attempts to correlate V_s and q_c using ML was made in the context of marine geotechnical site investigation in combination with other CPT parameters like sleeve friction and friction ratio. It requires, however, much greater efforts to make such correlation successful between V_s to q_c for site investigations in the land environment, where the spatial heterogeneity in the soil is more conspicuous.

Method: deriving q_c from V_s using ML

In the past, empirical correlations between V_s and q_c were found in numerous field studies. Such correlations also considered the effect of factors such as soil density, soil type, overburden pressure, and effective stress. These earlier correlations were mostly derived through regression analysis. More recently, pattern recognition and ML techniques have been used in determining such correlation.

We make use of the SCPT database from Groningen, in order to investigate the feasibility of predicting q_c from V_s through special adaptation of the ML approaches. The utilized database consists of 45 irregularly distributed SCPTs in Groningen. Using this information, we create a new database containing high-quality V_s profiles, spatial coordinates of the SCPTs, and relevant geological information. Geological information is obtained from TNO's 3-D GeoTOP model for the Dutch subsurface. The geology is expressed as a Boolean vector to indicate at each depth level which lithology is more likely e.g., clay, sandy clay, gravel, etc. For each depth level, we define the feature vector $\mathbf{v}$ as:

$$\mathbf{v} = [V_s, X, Y, h, L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8]^T, \tag{1}$$

where V_s is the shear-wave velocity at a specific depth level, X and Y are the spatial coordinates of the SCPT, h is the actual depth level, and the L corresponds to geological features, which are expressed as one-hot vectors as follows:

$$L_1 = [1,0,0,0,0,0,0,0]^T, \quad L_2 = [0,1,0,0,0,0,0,0]^T, \ldots, \quad L_8 = [0,0,0,0,0,0,0,1]^T. \tag{2}$$

This means that from the 8 possible values of L, at each depth level, one element takes on the value of 1, and the others are set to 0. This applies for all the SCPTs at every depth level defined in the data. The new dataset serves as the training dataset for ML. The target of the prediction is q_c. A common preprocessing step is applied before splitting the data for training, validation, and the ML tests to normalize the values - in order to provide the input vectors in a standard scale and to facilitate the training of a chosen ML algorithm:

$$v_{i,new} = \frac{v_i - v_{i,min}}{v_{i,max} - v_{i,min}}, \tag{3}$$

where v_i represents the scaled ith element of the vector $\mathbf{v}$, $v_{i,min}$ is the minimum value of the selected feature, $v_{i,max}$ is the maximum value, and $v_{i,new}$ is the new scaled value. We exclude some SCPTs that are judged as outliers. Fig. 2 illustrates the locations of the SCPTs in the province of Groningen. The average distance between the SCPTs is more than 200 m.

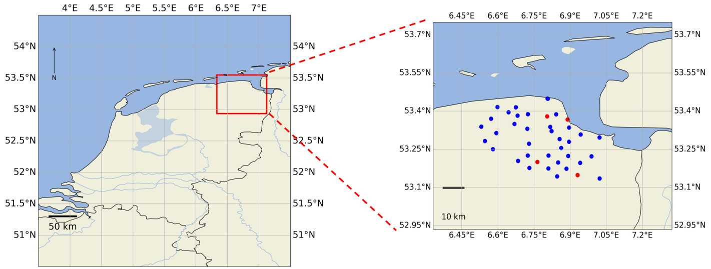

Figure 2 *SCPT database from Groningen. Left: locations of SCPTs in the northern tip of the Netherlands. Right: zoomed-in area in Groningen. Blue dots - SCPTs used for training. Red dots - SCPTs used for prediction.*

We adapt three different ML techniques: Support Vector Regressor or SVR (Boser et al., 1992), Random Forest Regressor or RFR (Breiman, 2001), Extreme Gradient Boosting or XG Boost (Chen and Guestrin, 2016). We chose these algorithms over deep learning algorithms (e.g., convolutional or recurrent networks) because these algorithms generally perform better in case of tabular datasets of relatively small sizes. When training tabular data using deep learning algorithms, as the data are propagated deeper into the layers of the network, it becomes diffucult to assign a physical interpretation to the learning process. Furthermore, the tabular data contain vectors for which many values are set to zero to describe the different lithologies, which also affects the training process. Ensemble tree-based algorithms such as XG Boost and Random Forest are, therefore, more suited for applications requiring training tabular data for regression tasks (Shwartz-Ziv and Armon, 2022).

The selected ML algorithms are tested using the Python toolbox Scikit Learn. GroupKfold cross-validation was used to avoid overfitting in the training process. A total of 5 Kfolds were used during the validation stage, and 3 SCPTs were set aside for testing. We conduct a random search to determine the best hyperparameters leading to each technique's highest prediction accuracy. For this, we select a sufficiently large range of possible hyperparameters for each technique. Subsequently, we train the data and select the hyperparameters that yield the best possible predictions.

Results

Figs. 3, 4, and 5 show the results of the q_c predictions at locations that are not used in the training. The label on top of each figure denotes the SCPT used for the prediction. The locations of the SCPTs used for training and prediction are shown in Fig. 2. At the locations (red dots) shown in Fig. 2, we have successfully predicted q_c with reasonable accuracy. The coefficient R2 is used as a measure of accuracy. We find that XG Boost offers the best results among the three adapted ML methods. XG Boost requires less computation time than SVR. It can also handle sparse data better than RFR due to the way how the algorithm builds the next trees based on the values of the previous ones. Further, we include regularization during the training of the XG boost, which also aids in obtaining a higher prediction accuracy. It is important to note that, in spite of the considerable distance between the SCPTs and the very large size of the region in which the SCPTs are located and the ML algorithms are trained, our adaptation of the ML techniques offers a good prediction of q_c. If such a prediction is performed in a site-specific manner, we anticipate the accuracy to be even higher.

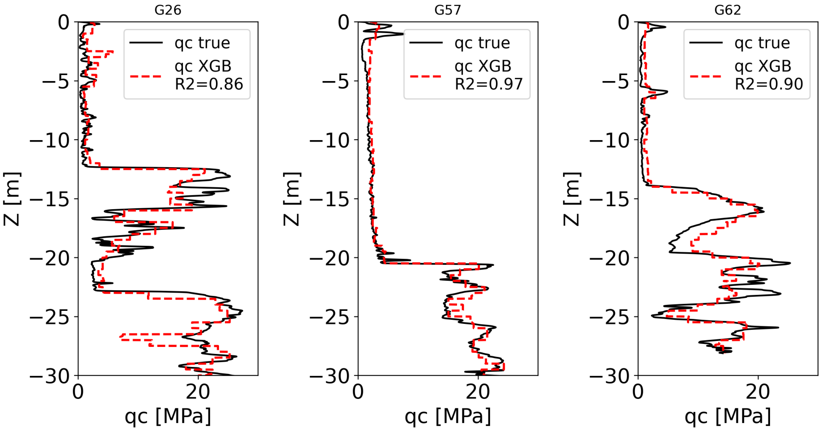

Figure 3 *True and predicted q_c using XG Boost algorithm.*

Conclusions

In this research, we attempt to derive site- and depth-specific correlations between V_s and q_c through developing novel approaches utilizing ML. For this purpose, we have specifically adpated three different ML techniques: Support Vector Regressor, Random Forest Regressor, Extreme Gradient Boosting. We have utilized the high-quality SCPT database available in Groningen in order to investigate the feasibility of predicting q_c from V_s. Organization of the SCPT data-coordinates and layer-specific geology in a

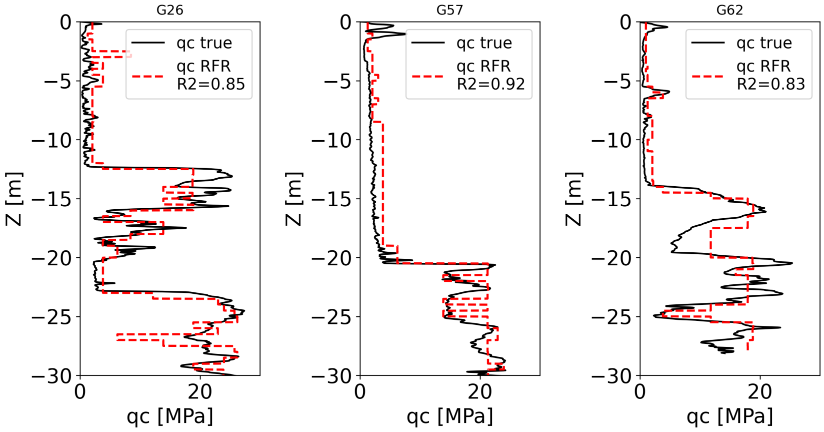

Figure 4 *True and predicted q_c using RFR algorithm.*

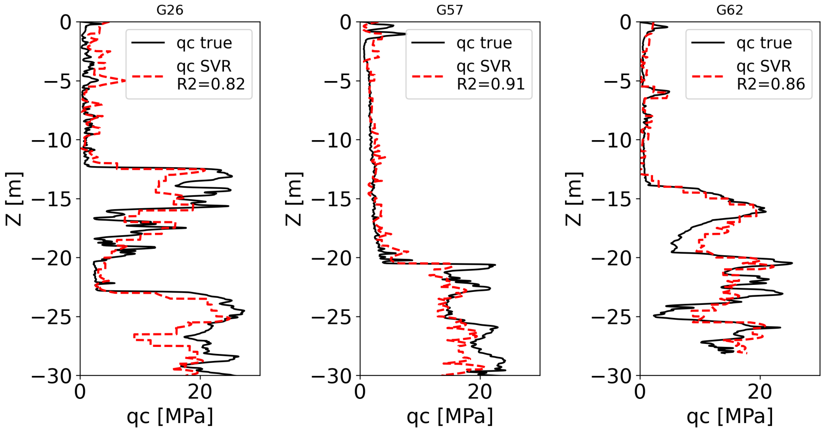

Figure 5 *True and predicted q_c using SVR algorithm.*

specific vector format were beneficial before data normalization. Among the adapted techniques, the Extreme Gradient Boosting approach offers the best prediction. Considering the large size of the region used in data training, the results obtained so far appear promising. In the next step we will test this approach in a more local scale, and will utilize V_s estimated from both SCPT and high-resolution surface seismic surveys.

Acknowledgements

The authors are grateful for the financial support provided by the Dutch Research Council (NWO) under the DeepNL programme (project 3DSOIL).

References

Boser, B.E., Guyon, I.M. and Vapnik, V.N. [1992] A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92. Association for Computing Machinery, New York, NY, USA, 144–152.

Breiman, L. [2001] Random forests. *Machine learning*, **45**, 5–32.

Chen, T. and Guestrin, C. [2016] XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 785–794.

Shwartz-Ziv, R. and Armon, A. [2022] Tabular data: Deep learning is not all you need. *Information Fusion*, **81**, 84–90.