MACHINE LEARNING-BASED PEDO TRANSFER FUNCTION FOR ESTIMATING THE SOIL COMPRESSION INDEX

Pham Nguyen Linh Khanh[a,*], Nguyen Huu Bao Ngan[a]

[a]*School of Civil Engineering and Management, International University, Vietnam National University HCM City, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam*

Article history:
Received 17/11/2022, Revised 12/02/2023, Accepted 21/02/2023

Abstract

Soil compression index ($\mathrm{C_c}$) plays a vital role in describing the settlement behaviors of geotechnical infrastructures. The conventional Oedometer test broadly used to determine $\mathrm{C_c}$ is time-consuming and expensive, which challenges incorporating the high spatial variability of $\mathrm{C_c}$. Alternatively, this study utilized the pedo transfer function (PTF) concept to develop a predictive model on the extreme gradient boosting (XGB) framework for estimating $\mathrm{C_c}$ with high accuracy and low effort. The presented XGB-PTF implemented on the database is acquired from 40 boreholes in Ho Chi Minh city and its vicinity to learn and recognize the correlation patterns of $\mathrm{C_c}$ and the easily-obtainable soil parameters (i.e., grain size distribution, unit density, moisture content, Atterberg limits). Rigorous evaluation with standard regression metrics demonstrated the efficiency and excellent performance of the XGB-PTF (e.g., low root-mean-squared error of 0.089 and a high coefficient of determination of 0.903). Furthermore, the presented framework showed its superiority over the current empirical equations in estimating $\mathrm{C_c}$ by higher prediction accuracy and applicability to the broader range of soil types. Given efficiency, flexibility, and dynamics, the presented model is expected to be a versatile approach to quantizing and advancing the knowledge of soil characteristics over a regional area.

Keywords: compression index; pedo transfer function; extreme gradient boosting; soil mechanics.

https://doi.org/10.31814/stce.nuce2023-17(1)-06

1. Introduction

Accuracy estimate of ground settlement is a typical geotechnical problem that has drawn massive attention during the history of soil mechanics. In practice, the state-of-the-art Terzaghi approach linearizes the correlation between void ratio (e) and effective stress to approximately describe the primary soil settlement behaviors with a certain degree of success. Given the increased effective stress, the expected settlement can be sufficiently estimated utilizing the compression index ($\mathrm{C_c}$) with acceptable accuracy for engineering applications. However, the Oedometer experiment for determining $\mathrm{C_c}$ is time-consuming and labor-intensive, which challenges incorporating the high spatial variability of $\mathrm{C_c}$.

Several experimental studies have been carried out to explore the variation patterns of $\mathrm{C_c}$ concerning different soil types. Closed-form equations have been proposed to estimate $\mathrm{C_c}$ based on fitting the obtained experimental data [1, 2]. Those published equations generally relate $\mathrm{C_c}$ in terms

[*] Corresponding author. *E-mail address:* pnlkhanh@hcmiu.edu.vn (Khanh, P. N. L.)

of liquid limit (LL), water content (w), and void ratio (e_0) (e.g., $C_c = 0.009(LL - 10)$ for clays [3]; $C_c = 0.011wC_c = 0.54\mathrm{e} - 0.19$ for peat [4]; $C_c = 0.54\mathrm{e} - 0.19$ for clays [5]). However, despite the practicality, merely being suitable for some specific soil texture restrains the generalized capacity of those empirical models.

In this regard, machine learning (ML) based pedo transfer function (PTF) is a promising alternative to the generic PTF. For details, PTF contributes as mathematical links between the easily-obtainable parameters (i.e., basic soil properties [6]) and the parameter of interest (e.g., C_c) that later allows exploiting the ML advantages in data mining to increase the model performance. Furthermore, the potential of ML-PTF has been accredited in describing various geotechnical applications ([7–10]). Recently, Zhang [11] developed the Bayesian Neural network-based model to forecast soil compressibility and undrained shear strength of clayey. Similarly, Scott Kirts et al. [12] utilized the support vector machine (SVM) model to predict the soil compressibility for coarse-grained, fine-grained, and organic peat. The obtained results once demonstrated the great potential of this approach in geotechnical design. Yet proper academic attention is required to enhance the practical applications of this approach further, as well as advance the knowledge of the correlation between given attributes and soil compressibility behaviors.

The primary objective of this research is to develop an ML-PTF on the extreme gradient boosting (XGB) framework [13] capable of estimating soil compression index with high precision and low effort. Furthermore, advancing the quantitative knowledge of which soil structural indicators determine soil compressibility using correlation analysis. The XGB-PTF was implemented on the Ho Chi Minh (HCM) soil database based on 40 boreholes system collected from different projects in HCMC and its vicinity.

2. Database and Correlation Analysis

2.1. Compression Index Database

This study developed the ML-PTF model for forecasting the C_c value base on the database of boreholes investigated in Ho Chi Minh city and its vicinity. Fig. 1 presents the approximate position of the boreholes. The soil specimen was collected at every 2 m to 3 m of the borehole, whose depth ranges from 30 m to 50 m, for determining the fundamental soil parameters. Consequently, the database of interest contains 600 data points with 13 attributes.

Data validity is utilized to eliminate errors that are likely to occur during measuring or documenting. The data points are expected to satisfy: (i) the particle size criterion and (ii) the physical relationship between soil parameters. The validity process resulted in a clean dataset containing 433 data points.

Moreover, the database includes features directly extracted from grain size distribution (fine gravel (FG), medium gravel (MG), coarse gravel (CG), very fine sand (VFS), fine sand (FS), medium sand (MS), coarse sand (CS), very coarse sand (VCS), fine silt (FS) and coarse silt (CS), fraction of clay (Fclay), silt (Fsilt) and sand (Fsand), natural moisture content (w), natural gravity of soil (γ), dry unit weight (γ_d), degree of saturation (S_r), the specific gravity of soil (G_s), void ratio (e_0), liquid limit (LL), plastic limit (PL), compression index (C_c). The secondary features were estimated based on interpreting the particle size distribution (i.e., D_{10} and D_{60} are grain sizes of 60% and 10% passing the sieved soil, respectively, C_u is a measure of the uniformity of grain size in the soil).

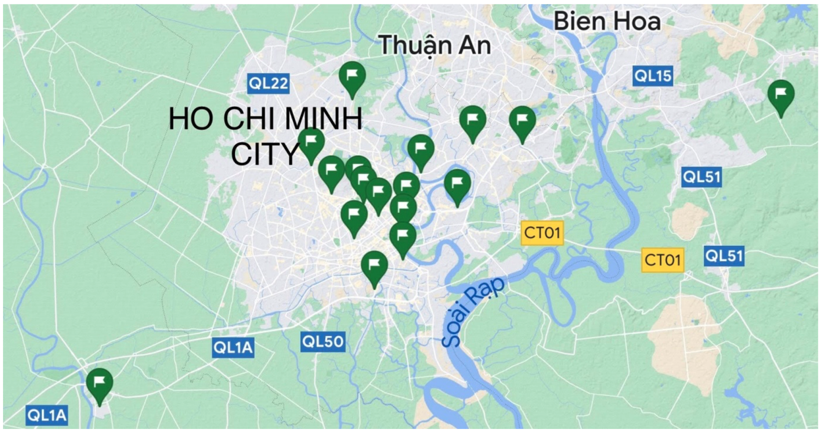


Figure 1. Approximate locations of 40 boreholes

2.2. *Data Properties*

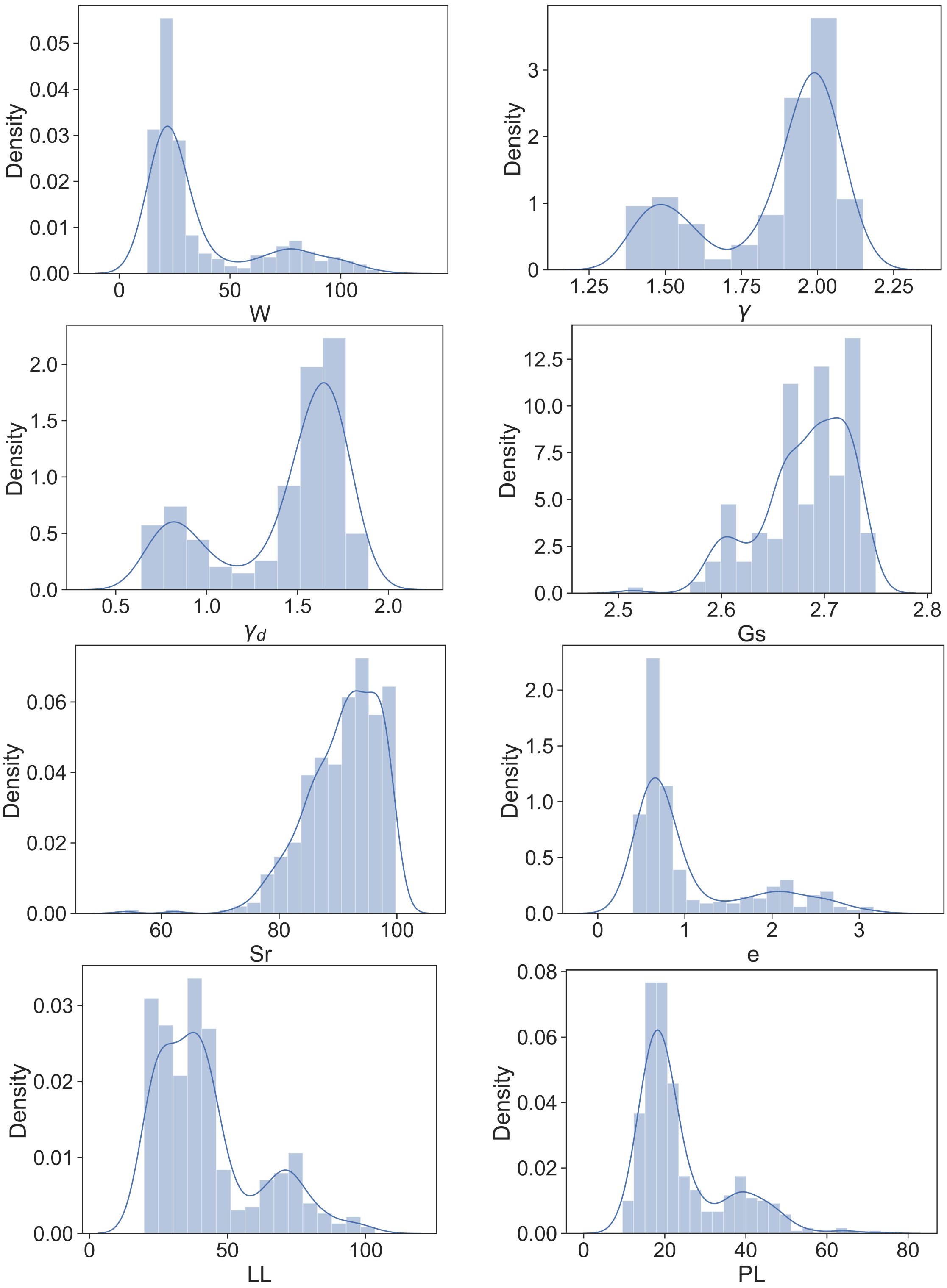
Table 1 summarizes the statistical descriptions of primary indicators in the database.

Table 1. Statistical descriptions of data attributes

Statistical description	w (%)	γ (g/cm^3)	γ_d (g/cm^3)	G_s (g/cm^3)	S_r (%)	e	LL (%)	PL (%)	C_c (cm^2/kG)
Number	433	433	433	433	433	433	433	433	433
Mean	36.7	1.86	1.42	2.68	90.9	1.05	42.4s_r	24.0	0.27
Std.	26.2	0.22	0.35	0.04	6.25	0.68	18.5	10.9	0.40
Min	12.5	1.37	0.64	2.51	53.9	0.40	19.8	9.57	0
Q_1 (25%)	19.3	1.66	1.2	2.66	87.0	0.60	27.2	16.9	0.06
Median	24.1	1.95	1.57	2.69	91.8	0.72	38.5	20.1	0.10
Q_3 (75%)	42.5	2.01	1.69	2.72	96	1.19	48.1	27.4	0.20
Max	117	2.15	1.89	2.75	99.8	3.17	103	73.3	2.41

(Q_1 and Q_3 are first and third quartiles)

As shown in Fig. 2, the water content had the broadest range among the other factors, which slanted toward the smaller value in the 19.3 - 42.53% range. Moreover, the degree of saturation had an extent from 53.9 to 99.8%, which slanted toward the larger value varying between 87 and 96%. In particular, 75% of the data had a degree of saturation higher than 87%. The natural gravity of soil showed a possible bimodal distribution, varying between 1.37 and 2.15 g/cm^3, which was primarily centered in the range of 1.66 - 2.01 g/cm^3. Similarly, the dry unit weight of soil also performed a possible bimodal distribution, varying between 0.64 and 1.89 g/cm^3, which was primarily centered in the range of 1.2 - 1.69 g/cm^3. The specific gravity indicates the lowest standard deviation, only 0.04, with a narrow range from 2.51 to 2.75 g/cm^3. The soil's void ratio ranged from 0.4 to 3.17, which slanted toward the smaller value varying between 0.6 and 1.19. Only 25% of the data had a void ratio higher than 1.19. The liquid limit had a high standard deviation value (18.54%) compared to other factors and has a broad range from 19.8 to 103.4%, which is slanted toward the smaller value in the range of 27.21 - 48.12%. Only 25% of the data had a liquid limit higher than 48.12%. The plastic limit also had a broad vary from 9.57 to 73.3%, which slanted toward the smaller value varying from 16.9 - 27.4%. The compression index ranged from 0 to 2.41 cm^2/kG, which slanted toward the smaller value varying from 0.06 to 0.2 cm^2/kG.



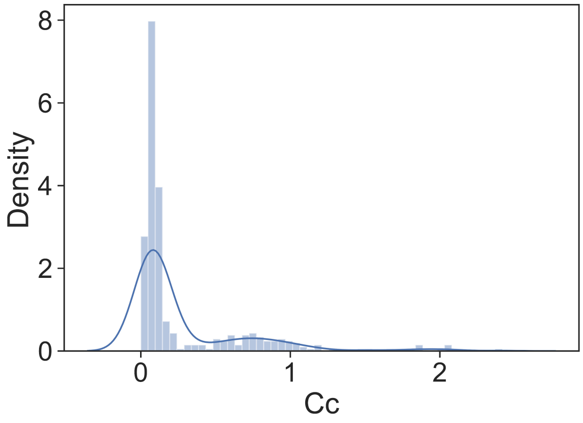


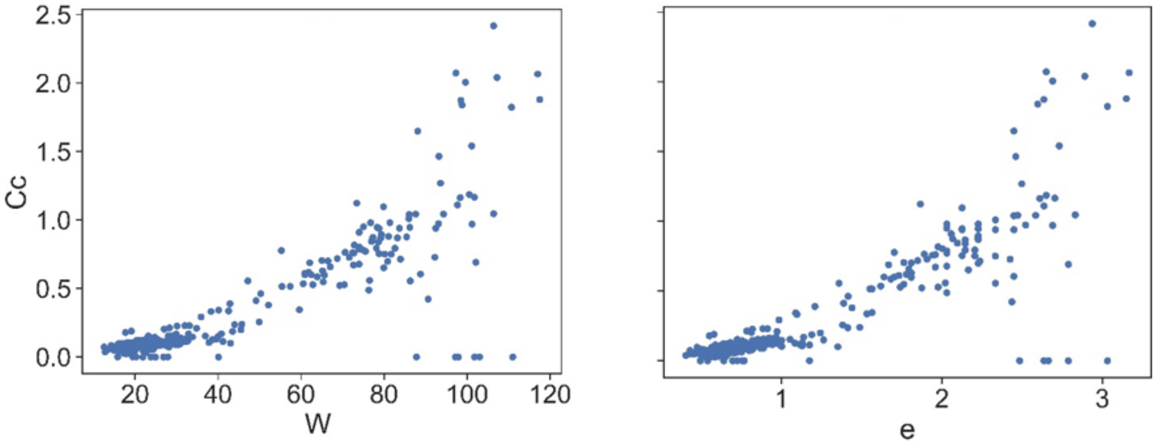
Figure 2. Distribution of factors on their ranges

2.3. Data Analysis

The Pearson correlation coefficient (R) was concisely utilized to estimate the correlation between each couple of input factors:

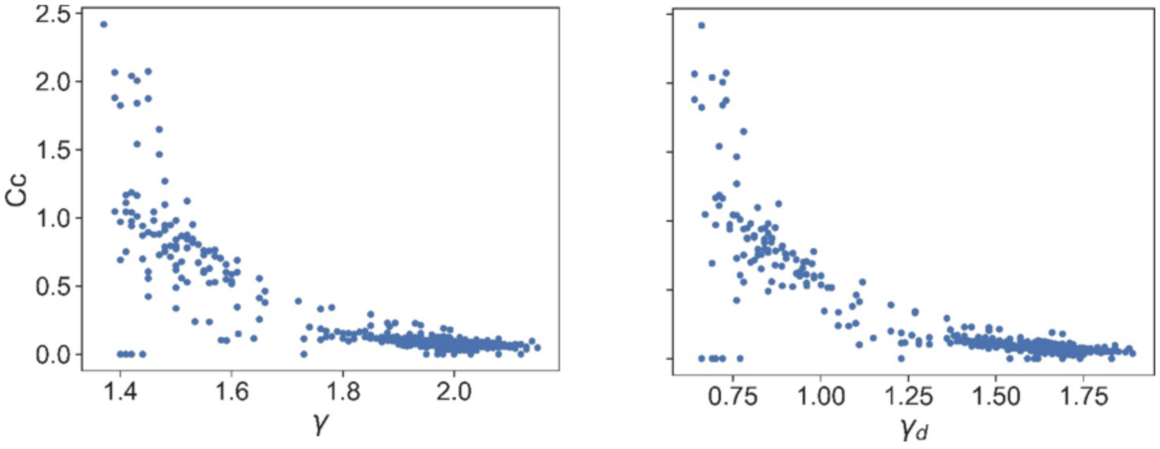
$$R = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2}} \tag{1}$$

The perfectly positive linear correlation gets the absolute R-value of 1; meanwhile, for two factors with no linear relationship, R equals 0. Fig. 3 shows the scatter plot for the correlation of C_c with the most relevant features (i.e., W, e, $\gamma_{d,}$ and γ). A relatively strong correlation was observed in the pair of C_c with W (R = 0.861) and e (R = 0.859). These observations are consistent with the empirical equations proposed in the literature. In contrast, in the case of $\gamma_{d,}$ and γ the negative R values were obtained, indicating the negative correlation with C_c of these parameters, expressed in the downward trend of the scatterplot.



(a)

(b)



(c)

(d)

Figure 3. Geographical scatterplots in pairs: (a) C_c and W (P = 0.86); (b) C_c and e (P = 0.86); (c) C_c and γ_d (P = -0.82); (d) C_c and γ (P = -0.80)

3. Methodology

3.1. Boosting learning

Boosting is a well-known branch of ensemble learning used to improve model performance. A chain of "weak" learners is sequentially added to the ensemble in a stepwise fashion to yield a potentially better one (as presented in Fig. 4).

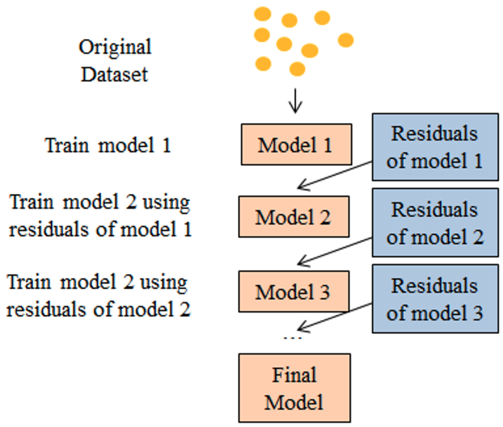


Figure 4. Boosting learning concepts

3.2. Extreme Gradient Boosting (XGB)

Extreme gradient boosting is engineered from the well-known boosting algorithm, especially emphasized in [13] for decreasing the great amount of risk of overfitting problems and enhancing model efficiency. Overfitting is a state-of-the-art issue to all the machine learning models, in which the model performs excellently on the training data but poorly on the previously unseen data (i.e., test set).

Owing to the simplicity, the Decision tree (DT) is chosen as the weak learner that is sequentially added to the ensemble stepwise fashion to enhance the overall performance. Given an instance x_i, the prediction $\hat{y}_i$ of the target y_i is obtained by utilizing predetermined K DTs as below:

$$\hat{y}_i = \sum_{k=1}^{K} \hat{f}_k(x_i) = \sum_{k=1}^{K} \hat{f}_k(x_i) + \rho_K \hat{f}_K(x_i) \tag{2}$$

where $F = \left\{ \mathrm{f(x)} = \mathrm{w}_{q(x)} | q : R^m \rightarrow T; \mathrm{w} \in R^T \right\}$ is the family of DT f(x), q is the structure of each tree that represents an instance to the corresponding leaf index. T represents the number of leaves in the tree; each leaf contains a continuous score w; ρ represents the learning rate.

Chen & Guestrin resettled the high risk of overfitting issue inherent in boosting models by regularizing the original objective functions (Eq. (3)) to favor the less complexity model (e.g., simple DT structure with low w on each leaf)

$$L(\hat{y}) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\hat{f}_k)$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \mathrm{w}_j^2 \tag{3}$$

Here l is a differentiable convex loss function that computes the difference between and y_i; n is the number of data; γ and λ are regularized hyperparameters; $\Omega(f)$regularized functions.

During the training process, the new DTs are added to the ensemble in the direction determined by the gradient descent concept to minimize the objective function (Eq. (3)). For details, the $k^{th} \hat{f}_k$ is trained with the pseudo database $\left\{ x_i, r_{ik} = \dfrac{\partial L^k}{\partial \hat{f}(x)} \middle| \hat{f}(x) = \hat{f}_k(x) \right\}$ to focus on specific rows. The objective function becomes as follows:

$$L^k = \sum_{i=1}^{n} l(y_i, \hat{y}^{(k-1)} + \rho_k \hat{f}_k(x_i)) + \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T^k} (\mathrm{w}_j^k)^2 \tag{4}$$

3.3. Hyperparameter tuning

The hyperparameters are used to manage the distance between the testing and training errors and enhance model performance. Considering the computational expense, the Bayesian optimization [14] coupled with the K-fold cross-validation was adopted to tune the hyperparameters of the XGB. Table 2 summarizes the XGB hyperparameters.

3.4. Train-Test set

Machine learning models conduct specific tasks in accordance with the patterns extracted from the databases. The training or learning process is the procedure of identifying the regularity and pattern of the database. Once the learning phase is finished, the trained model can appropriately execute a given task on the formerly unseen inputs, and this capability is experienced as generalization. This study utilized 80% of the database (i.e., training data) for training the model, and the remaining 20% (i.e., test set) was to evaluate the model's generalized capacity.

4. Results and Discussion

4.1. Evaluation metrics

Standard evaluation metrics for regression models, including root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and coefficient of determination (R),

Table 2. A summary of the hyperparameters of XGB

Hyperparameters	Descriptions	Search space/ Distribution	XGB
Number of estimators (K)	Determine the number of DTs	[10, 500]/ Uniform	162
Regularized parameters (λ, γ)	Penalty the model complexity	[1×10^{-5}, 5]/ Log-Uniform	0.427 0.168
Learning rate (ρ)	Adjust the generalization capacity	[0.01, 1]/ Log-Uniform	0.3
Maximum depth	Control the maximum depth of DT	[1, 20]/ Uniform	4
Subsample ratio	Subsample ratio of training instances	[0, 1]/ Uniform	0.912
Column subsample by tree	Subsample ratio of columns when constructing each DT	[0, 1]/ Uniform	0.683

were utilized to evaluate predictive performance. The high R-values, accompanied by low RMSE and MAE values, prove the outstanding presentation of the developed XGB.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{obs.} - y_i^{pred.})^2}$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{pred.} - y_i^{obs.}\right)^2}{\sum_{i=1}^{n}\left(y_i^{obs.} - \bar{y}^{obs.}\right)^2}; MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{obs.} - y_i^{pred.}\right| \quad (5)$$

where n is data samples; $y_i^{pred.}$ and $y_i^{obs.}$ are observed and predicted C_c.

In addition, five empirical equations for predicting C_c in terms of LL, w, or in-situ void ratio (e_0) [15] were utilized to validate the XGB model performance.

4.2. Results

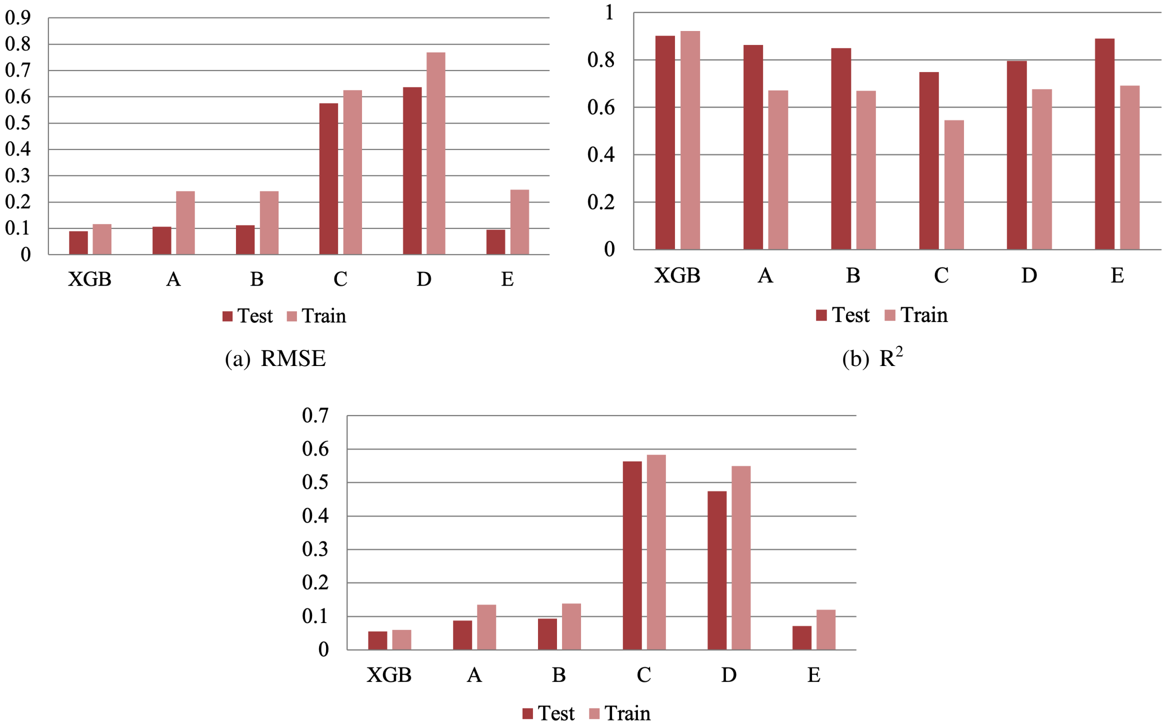
Table 3 and Fig. 5 summarize the evaluation results for the performance of the presented XGB-PTF, along with the broadly used empirical equations. The RMSE and MAE values of XGB were the lowest in comparison with all empirical models on the same dataset, whereas the R^2 value of XGB was higher than all empirical models that demonstrate the superior performance of the developed XGB.

Further elaboration on the model performance was carried on by analyzing the residual errors $y^{pred.} - y^{obs.}$ of the presented models. In Table 4, the mean value of residual error in the case of XGB is -0.006, and the standard deviation is 0.09, which is relatively low compared to the remaining. This means the predicted value of XGB does not fluctuate greatly and is adaptable to many types of soils. Furthermore, the interquartile range ($Q_3 - Q_1$) of the residual error distribution of XGB was close to the zeros line, ranging from -0.045 to 0, as shown in Fig. 6. The amplitude obtained in the case of XGB is smaller than the rest, especially box D, which ranges from -0.197 to -0.55. Box D has a large

Table 3. RMSE, R^2, MAE comparison between XGB and five empirical equations

Regression Metrics	Criteria	XGB	A	B	C	D	E
RMSE	RMSE → 0	0.089	0.106	0.111	0.575	0.636	0.095
R^2	R > 0.8	0.903	0.864	0.849	0.749	0.796	0.891
MAE	MAE → 0	0.055	0.087	0.092	0.563	0.474	0.072

Note: A: $C_c = 0.37(e_0 + 0.003LL - 0.34)$ [16]; B: $C_c = 0.009w + 0.002LL - 0.10$ [16]; C: $C_c = 0.141G_s(\frac{\gamma_w}{\gamma_d})^{\frac{12}{5}}$; D: $C_c = 1.15(e_0 - 0.35)$ [5]; E: $C_c = -0.156 + 0.411e_0 + 0.00058LL$ [17].



(a) RMSE

(b) R^2

(c) MAE

Figure 5. Comparison between XGB and empirical equations

amplitude because it only considers one factor, e_0, which will lead to low performance compared to other models. Consequently, these results again show the predictive ability of the XGB model.

In Table 5, the XGB model, which applies to various types of soil, performs better values of RMSE and R^2 than SVM, which present inconsistent results. In particular, SVM models have a grateful prediction only for coarse-grained, whereas the prediction for fine-grained and organic peat is not good at all. It is acceptable that XGB is a preeminent model for prediction C_c.

Fig. 7 indicates the impact on the expected accuracy of given features on the model. Among them, the moisture content showed the most substantial influence, expressed by the highest value of feature importance. This is supported by the experimental results reported in the literature, which were later developed into the empirical equation describing the linear correlation between C_c and W.

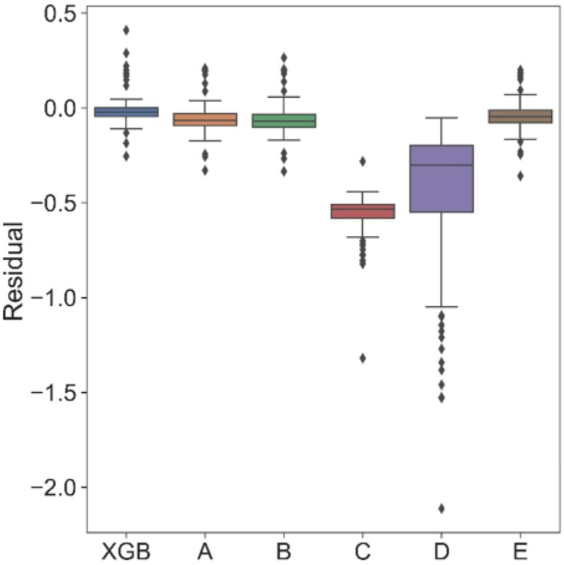


Figure 6. Residuals comparison between XGB and empirical equations

4.3. Discussion

The evaluation results demonstrated that the presented XGB-PTF outperformed the empirical formulas broadly used for predicting C_c. Note that the XGB-PTF consistently performed excellently on various soil types. The flexibility in considering multiple input features for predicting C_c of the XGB-PTF, instead of one or two independent variables like those empirical equations, responds to the improvement in the prediction accuracy. Also, the remarkable capacity to learn complex data patterns allows the XGB-PTF to be workable on a wide range of soil types.

The feature importance indicates the critical role of W in predicting C_c. This obtained result is supported by the theoretical interpretation from soil mechanics and experimental results reported in the literature. Nevertheless, limited in the relatively small database may mislead the understanding of the impacts of the remaining features (such as Atterberg limits and attributes describing grain size distributions). Therefore, further studies should be carried on to explore the potential influence of those relevant factors.

Table 4. Residuals statistical summarization on the test set

	XGB	A	B	C	D	E
Mean	-0.006	-0.053	-0.056	-0.563	-0.474	-0.041
Std	0.090	0.092	0.097	0.117	0.427	0.086
Min	-0.255	-0.330	-0.335	-1.320	-2.113	-0.359
Q_1	-0.045	-0.093	-0.102	-0.581	-0.550	-0.078
Median	-0.023	-0.067	-0.071	-0.534	-0.301	-0.047
Q_3	0.000	-0.031	-0.034	-0.508	-0.197	-0.012
Max	0.410	0.209	0.264	-0.282	-0.052	0.200

Table 5. RMSE and R^2 comparison between XGB and SVM model used in predicting compression index C_c (Scott Kirts et al., 2017)

	XGB		SVM		
	Train	Test	Coarse grained	Fine grained	Organic peat
RMSE	0.117	0.089	0.111	0.391	1.090
R^2	0.923	0.903	0.910	0.650	0.770

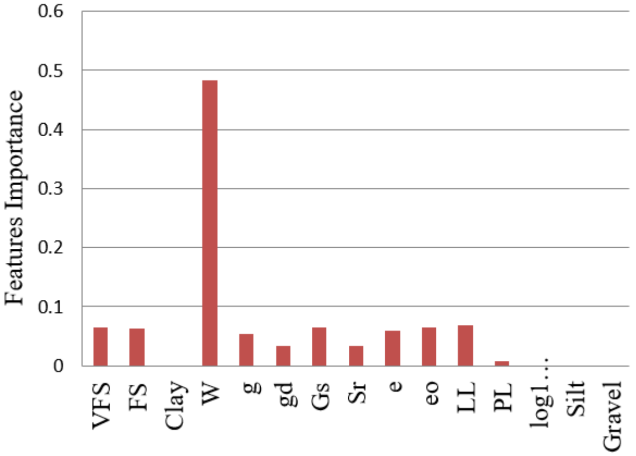


Figure 7. Features Importance of model

5. Conclusions

This study developed the XGB-PTF for predicting compression index. The XGB-PTF was implemented on the geological data from a system of boreholes around Ho Chi Minh city and its vicinity. Also, the hyperparameters of XGB were tunned with the aid of Bayesian optimization coupled with the K-fold cross-validation. Standard evaluation metrics (RMSE, MAE, and R^2) were utilized to evaluate the performance of the developed model. For comparison, five empirical formulas were utilized to evaluate the performance of XGB. The statistical metrics demonstrated excellent performance of the model over the empirical formulas in predicting the compression index of soil (e.g., $RMSE = 0.089$ and $R^2 = 0.903$). Consequently, using machine learning models, especially XGB-PTF, is highly suggested to develop reliable models for identifying the compression index C_c and for advanced application in geotechnical infrastructures.

Acknowledgment

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2022-28-01.

References

[1] Jefferson, I., Smalley, I. (1997). Soil mechanics in engineering practice. *Engineering Geology*, 48(1-2): 149–150.

[2] Einav, I. (2007). Soil mechanics: breaking ground. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1861):2985–3002.

[3] Terzaghi, K., Peck, R. B., Mesri, G. (1996). *Soil mechanics in engineering practice*. John Wiley & Sons.

[4] Cook, P. M. (1956). Consolidation characteristics of organic soils. In *Proc. of 9th Canadian Soil Mechanics Conf*, volume 41, 82–87.

[5] Nishida, Y. (1956). A Brief Note on Compression Index of Soil. *Journal of the Soil Mechanics and Foundations Division*, 82(3).

[6] Looy, K. V., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., Vereecken, H. (2017). Pedotransfer Functions in Earth System Science: Challenges and Perspectives. *Reviews of Geophysics*, 55(4):1199–1256.

[7] Pham, K., Kim, D., Yoon, Y., Choi, H. (2019). Analysis of neural network based pedotransfer function for predicting soil water characteristic curve. *Geoderma*, 351:92–102.

[8] Pham, K., Won, J. (2022). Enhancing the tree-boosting-based pedotransfer function for saturated hydraulic conductivity using data preprocessing and predictor importance using game theory. *Geoderma*, 420:115864.

[9] Hung, D. V., Thang, N. T. (2022). Predicting dynamic responses of frame structures subjected to stochastic wind loads using temporal surrogate model. *Journal of Science and Technology in Civil Engineering (STCE) - HUCE*, 16(2):106–116.

[10] Doan, Q. H., Thai, D.-K., Tran, N. L. (2020). A hybrid model for predicting missile impact damages based on k-nearest neighbors and Bayesian optimization. *Journal of Science and Technology in Civil Engineering (STCE) - HUCE*, 14(3):1–14.

[11] Zhang, P., Yin, Z.-Y., Jin, Y.-F. (2022). Bayesian neural network-based uncertainty modelling: application to soil compressibility and undrained shear strength prediction. *Canadian Geotechnical Journal*, 59(4): 546–557.

[12] Kirts, S., Panagopoulos, O. P., Xanthopoulos, P., Nam, B. H. (2018). Soil-Compressibility Prediction Models Using Machine Learning. *Journal of Computing in Civil Engineering*, 32(1).

[13] Chen, T., Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

[14] Snoek, J., Larochelle, H., Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

[15] Al-Khafaji, A. W. N., Andersland, O. B. (1992). Equations for Compression Index Approximation. *Journal of Geotechnical Engineering*, 118(1):148–153.

[16] Azzouz, A. S., Krizek, R. J., Corotis, R. B. (1976). Regression Analysis of Soil Compressibility. *Soils and Foundations*, 16(2):19–29.

[17] Rendon-Herrero, O. (1980). Universal Compression Index Equation. *Journal of the Geotechnical Engineering Division*, 106(11):1179–1200.