The 2025 World Congress on
Advances in Structural Engineering and Mechanics (ASEM25)
BEXCO, Busan, Korea, August 11-14, 2025

Development of machine learning-based three-dimensional spatial interpolation for underground space

*Khiem Nguyen Tran Gia[1)] and Jongmuk Won[2)]

[1), 2)] *Department of Civil, Urban, Earth, and Environmental Engineering, UNIST, UNIST-gil 50, Ulsan 44919, Korea*
[2)] *jwon@unist.ac.kr*

ABSTRACT

Obtaining soil property data in underground spaces is challenging due to limited access and the inherent spatial variability of soils. This study proposes a machine learning-based three-dimensional Geotechnical Distance Field (3D-GDF-ML) framework for reliable and computationally efficient geotechnical interpolations. A synthetic Gaussian random field of cone tip pressure was generated to evaluate the performance and uncertainty of the developed 3D-GDF-ML. Several ML algorithms were selected to evaluate the best-performing ML model for the proposed framework. It was found that random forest was the best-performing ML algorithm for 3D-GDF-ML-based geospatial interpolation. In addition, optimal interval of site investigation for reliable interpolation using 3D-GDF-ML model indicates the applicability of developed framework for sparse site investigation dataset.

1. INTRODUCTION

Quantifying the spatial variability of soil properties is fundamental to assessing the safety and serviceability of geoinfrastructures in underground space. However, sparse and highly variable site data make direct characterization of these properties challenging (Louppe et al., 2013). Consequently, accurate geospatial interpolation of soil properties is crucial for design and risk assessments of underground structures. Classical geostatistical techniques such as Kriging (Isaaks & Srivastava, 1989), are the most popular, as they provide the best linear unbiased estimates and clear covariance framework. However, the accuracy of these methods diminishes when critical parameters (e.g., scale of fluctuation) must be inferred from limited horizontal data, and their reliance on linear assumptions can mask complex local trends. More recently, machine-learning (ML)-based algorithms such as random forests (RF), extreme gradient boosting (XGB), extra trees (ET), gradient boosting (GB), and K-nearest neighbors (KNN), especially tree-based ensemble models, have gained popularity in geospatial

[1)] Graduate Student
[2)] Professor

interpolation due to their capabilities of delivering smooth interpolation from highly nonlinear profiles. Nevertheless, when trained on raw Cartesian coordinates, these ML algorithms may struggle to capture intrinsic spatial correlations and coordinate-to-distance transformations, which can cause decrease in predictability in data-sparse settings. To overcome this limitation, a geotechnical distance field (GDF) model (Xie et al., 2022) was introduced to explicitly quantify spatial relationships for efficient and accurate geotechnical interpolation. However, the proposed GDF model was only for two-dimensional data. Therefore, this study proposes a three-dimensional geotechnical distance field (3D-GDF) framework to interpolate soil properties in three-dimensional domain. The uncertainty quantification of the developed framework and the optimal interval between site investigation for reliable interpolation using 3D-GDF are also discussed.

2. METHODOLOGY AND RESULTS

2.1 Background and proposed three-dimensional spatial interpolation framework

The proposed 3D-GDFs framework follows the concept introduced in Xie et al., (2022) for two-dimensional GDF (2D-GDF) function. The concept consists of three main components but is extended to account for a 3D domain: 1) 3D surface distance field, 2) 3D testing distance field, and 3) 3D corner distance field. The surface distance field represents the distance from the ground surface in the depth (Z) direction, which enables capturing the characteristics of depth-dependent soil properties. The testing distance fields represent the horizontal distance (X or Y) from each testing line along depth, which facilitates quantification of horizontal distance from the measured soil properties. Corner distance fields store the shortest distance to each domain corner, thereby providing smooth and continuous interpolation throughout the 3D domain. A visualization with nine measured dataset locations (e.g., CPTs locations) in a 5 × 5 × 4 grid in Figure 1 yields eighteen distance fields whose Euclidean values populate the training matrix. For example, the vector for sample A lists the GDF values at a location with known soil properties [0; 0; 2; 4; …; 5.66; 19; 0; 4; 19.42; 19.42; 4; 5.66; 19.82]. Once the ML-based model is trained, unknown properties (e.g., sample B in Figure 1) can be predicted from the GDF values at corresponding location of unknown soil properties [0; 1; 1; 3; ...; 5; 19.03; 1; 3; 19.24; 19.44; 4.12; 5; 19.65]. The proposed 3D-GDF offers three key advantages: (i) enhanced ability to capture complex subsurface geometry and depth-dependent variability, (ii) direct estimation and visualization of fully three-dimensional property fields crucial for geostructural and underground-space design, and (iii) adaptability to heterogeneous geological settings, which broadens the applicability across diverse geotechnical problems while ensuring a consistent, data-efficient platform for uncertainty quantification and the spatial optimization of site investigation.

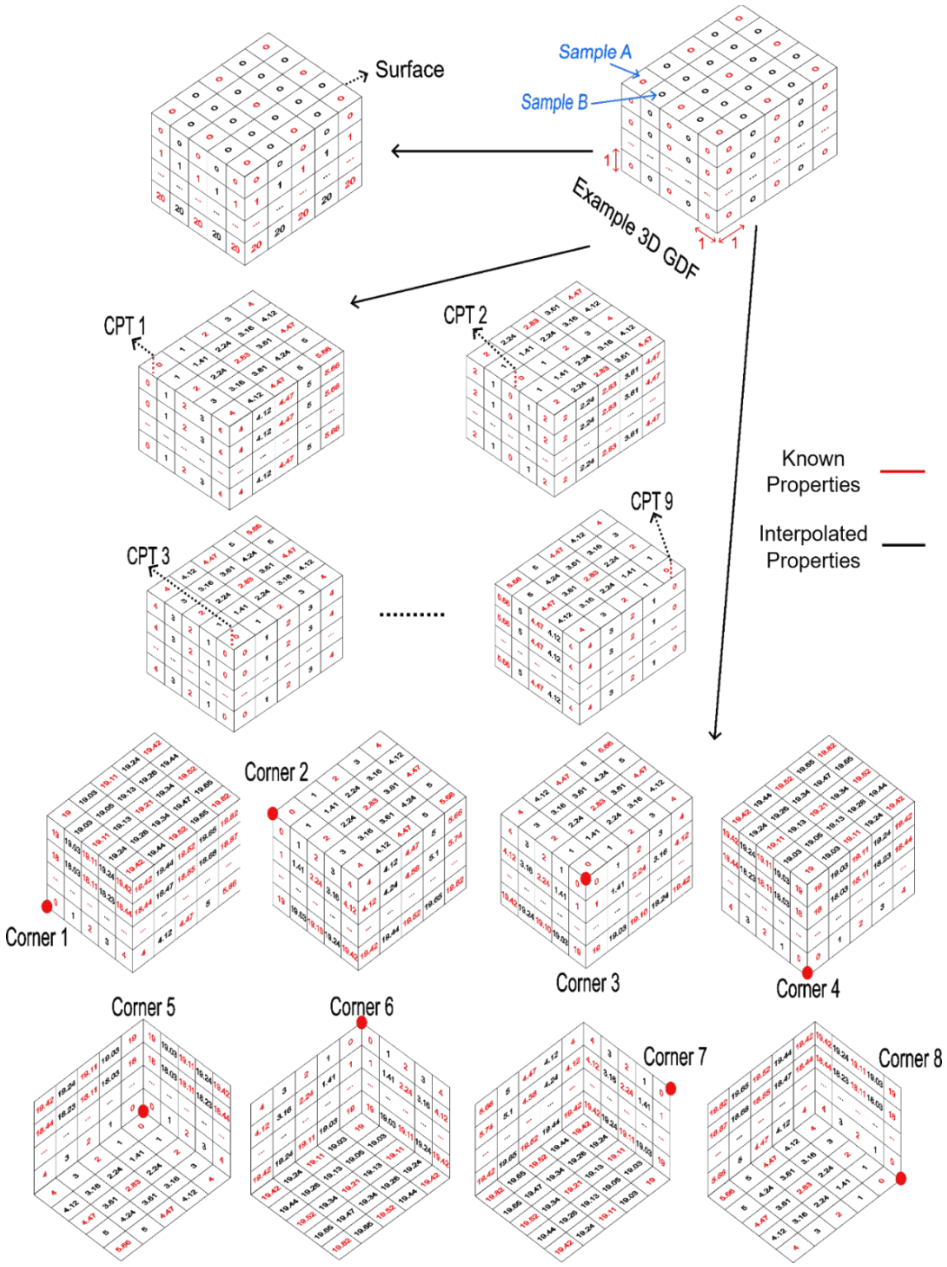


Fig. 1 Illustration of developed 3D-GDF framework for 3-by-3 site investigations.

2.2 Results obtained with proposed framework

To identify the best-performing model for the proposed 3D-GDF framework, five machine-learning algorithms— ET, RF, GB, XGB and KNN were evaluated. Optimal hyperparameters for each learner were obtained through Bayesian optimization, and the performance of all algorithms was evaluated using root-mean-square error (RMSE) and coefficient of determination (R^2). To train and assess the employed ML models, a synthetic three-dimensional random field of corrected cone tip resistance (q_t) was

generated using an open-source Python module named GSTools (Müller et al., 2022). The domain dimensions were 50 m × 50 m horizontally and 20 m vertically, with mesh resolution in z-direction (δ_z) = 0.05 m and mesh in both x and y directions (δ_x) = (δ_y) = 0.5 m. The random field created, based on a Gaussian semi-variogram, provided a statistically controlled representation of spatial variability while allowing reproducible ground-truth values. Paraview (Ahrens et al., 2005) was used to visualize the generated field and the interpolated outputs in three dimensions. Figure 2 illustrates the created three-dimensional random field of synthetic q_t together with the locations of the 6 by 6 CPT used for training, whereas the remaining grid nodes constituted an independent validation set for comparing the performance of each algorithm.

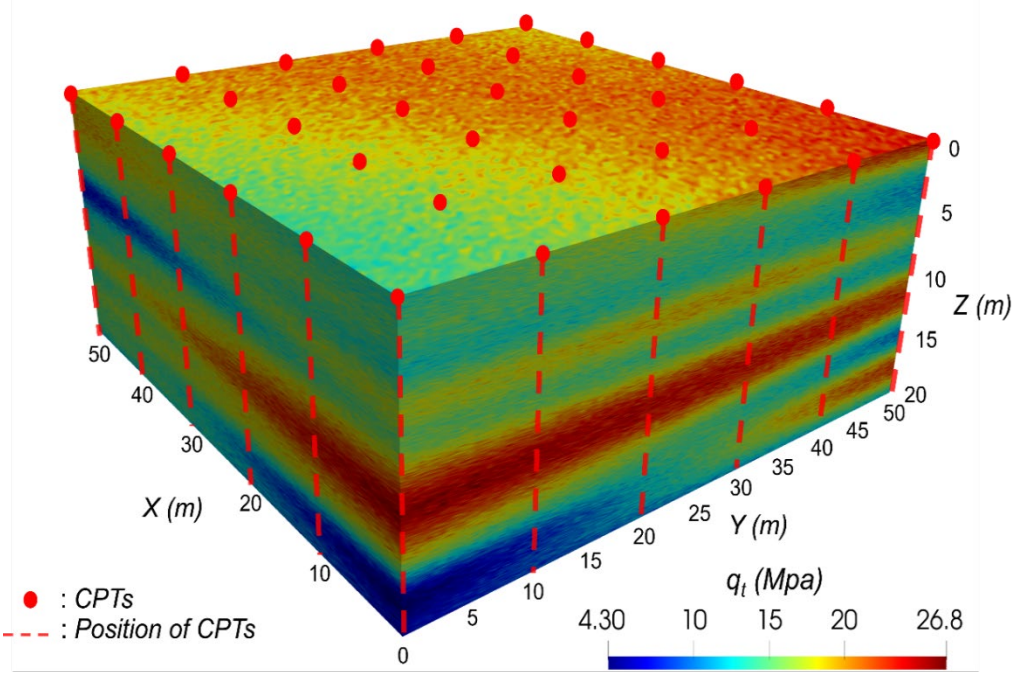


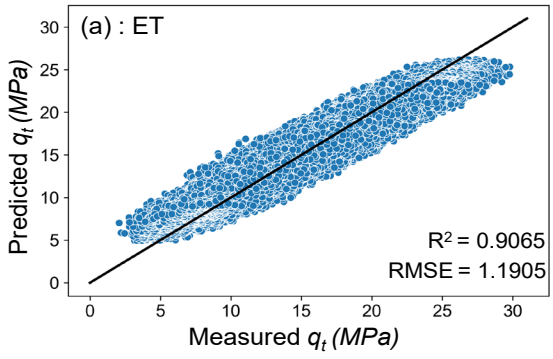
Figure 2. Synthetic q_t field illustration (50 m × 50 m × 20 m, number of CPTs = 36)

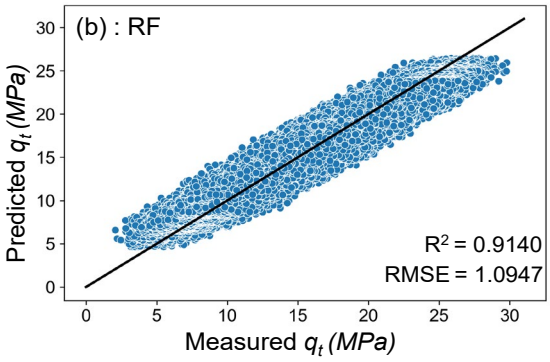
Table 1 summarizes the performance of the 3D-GDF ML-based spatial interpolation framework implemented with ET, RF, GB, XGB and KNN. Across all models, R^2 values on the test dataset lie between 0.899 and 0.914 and the RMSE range from 0.722 to 1.10, indicating generally strong performance of developed models. This is further confirmed through the visualization with scatter plots in Figure 3. Among the five developed ML models, RF achieved the highest accuracy (R^2 = 0.914). While XGB obtained a comparable results of R^2 = 0.912 with shorter runtime but showed overfitting at lower bound values due to the inherent sequential boosting, which can cause bias towards specific features when the dataset is sparsely available (~15 000 training points for 6 by 6 CPTs) compared to a four million nodes prediction grid. ET model, although usually considered to be more efficient than RF, yielded a lower R^2 = 0.907. This can be attributed to the fully random split thresholds of ET that produce less diverse ensembles and higher bias. KNN showed severe overfit, recording a near-zero training RMSE yet the weakest generalization (R^2 = 0.899), while GB performed intermediately. From the

results obtained, RF was determined to be the best-performing ML algorithm, differing from the 2D-GDF study where ET prevailed, demonstrating that the optimal learner is governed by the dimensionality of the data, grid density and value range. The best-performing RF-based 3D-GDF model (3D-GDF-RF) is adopted for subsequent evaluations. The visualization of interpolation using 3D-GDF-RF shown in Figure 4 confirms that the predicted q_t values closely resemble the synthetic q_t field in Figure 2.

Table 1. R^2 and RMSE scores of ML algorithms after hyperparameters tuning.

Machine Learning algorithm		Averaged values after 10 simulations		
		Train	Test	Run time (s)
ET	R^2	0.9354	0.9065	456
	RMSE	0.9156	1.1905	
RF	R^2	0.9490	0.9140	741
	RMSE	0.7219	1.0947	
GB	R^2	0.9439	0.9109	798
	RMSE	0.7950	1.1352	
XGB	R^2	0.9441	0.9129	275
	RMSE	0.7456	1.1090	
KNN	R^2	1.000	0.899	135
	RMSE	$3.4232e^{-10}$	1.2861	





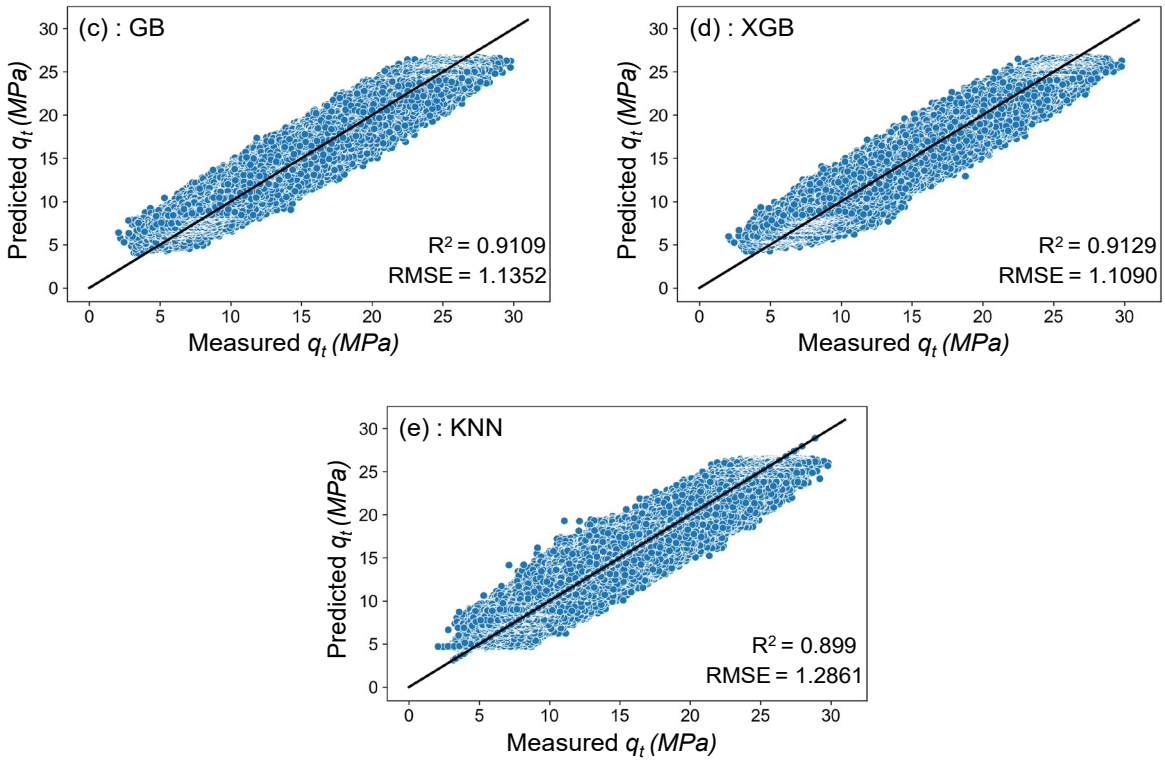


Figure 3. Predicted and measured q_t for test dataset (4,090,601 data).

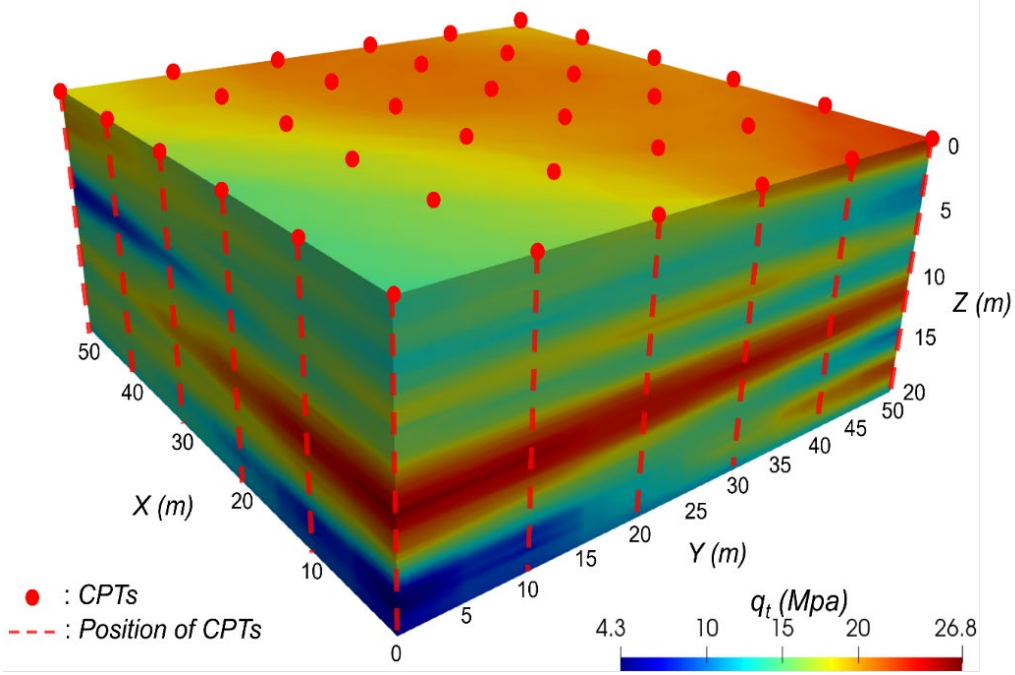


Figure 4. Predicted q_t values obtained with 3D-GDF-RF model.

2.3 Impact of data availability on the proposed framework

Figure 5 presents the R^2 and RMSE values for each ML algorithm, obtained with δ_z = 0.05 m, as a function of CPT spacing. As observed in Figure 5, the accuracy of the RF-GDF model declines as the CPT spacing (δ_x) widens from 5.56 m, with number of CPTs (N_{CPTs}) = 100, to 50 m (N_{CPTs} = 4). The test-set R^2 values fall and RMSE values rise sharply at δ_x = 50 m, whereas only modest R^2 reductions occur between 5 m and 25 m, indicating that reliable interpolation is maintained when investigations are performed at intervals less than 25 m. This spacing aligns with previous studies on CPT site investigations for large-scale projects (Eid et al., 2018). Accordingly, the high prediction accuracy achieved for δ_x < 25 m suggests that the proposed framework can be applied to large-scale projects. Because the given results are obtained with δ_z = 0.05 m, the abovementioned implications are only valid for semi-continuous site investigation data. Therefore, the influence of δ_z values was also evaluated.

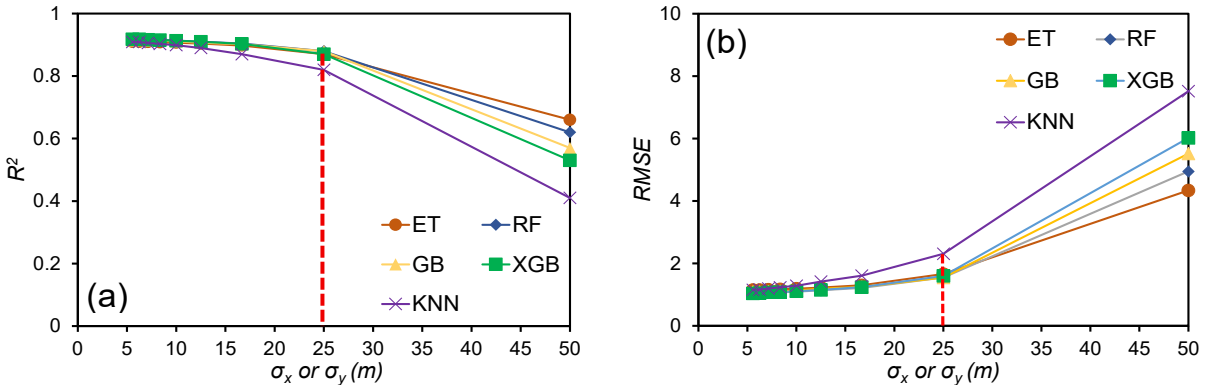


Figure 5. R^2 (a) and RMSE (b) values of five ML-GDF models as a function CPT interval.

Figure 6(a) demonstrates the performance of the RF-GDF model as a function of δ_z, ranging from 0.05 to 5 m (N_{CPTs} = 36), highlighting a clear reduction in R^2 values for both training and test datasets as δ_z increases, indicating the critical role of vertical data resolution for accurate predictions. Higher δ_z intervals also lead to increased uncertainty, as evidenced by the broader range of R^2 values (0.07 to 0.52) at δ_z = 5 m for the test dataset. Nonetheless, the framework still achieves a high R^2 > 0.8 at δ_z = 1.5 m, indicating a practical balance between computational efficiency and accuracy due to significantly reduced data requirements and faster computation compared to δ_z = 0.05 m. Figure 6(b) illustrates the R^2 results for training and testing dataset as a function of N_{CPTs} at δ_z = 4 m. The results indicate that model accuracy improves with increasing N_{CPTs}, although this improvement saturates after N_{CPTs} = 9, implying an optimal site investigation density (i.e., spacing of δ_x) exists to balance accuracy with cost-efficiency. The study also identifies limitations, emphasizing that high-resolution investigations, such as CPT or dilatometer tests, are critical for reliable interpolation, whereas lower-resolution data from SPT might compromise prediction accuracy. Nevertheless, moderate investigation densities at larger vertical intervals (e.g., δ_z = 4 m, N_{CPTs} = 36) can still yield acceptable

results (R^2 ~ 0.4), as illustrated in Figure 6(c), which are comparable to predictions obtained at δ_z < 4 m, suggesting increased horizontal density can partially compensate for low δ_z. The flexibility of the proposed framework in accommodating varying data intervals in x, y, and z directions indicate the strong applicability of the proposed framework for any geotechnical site investigation techniques.

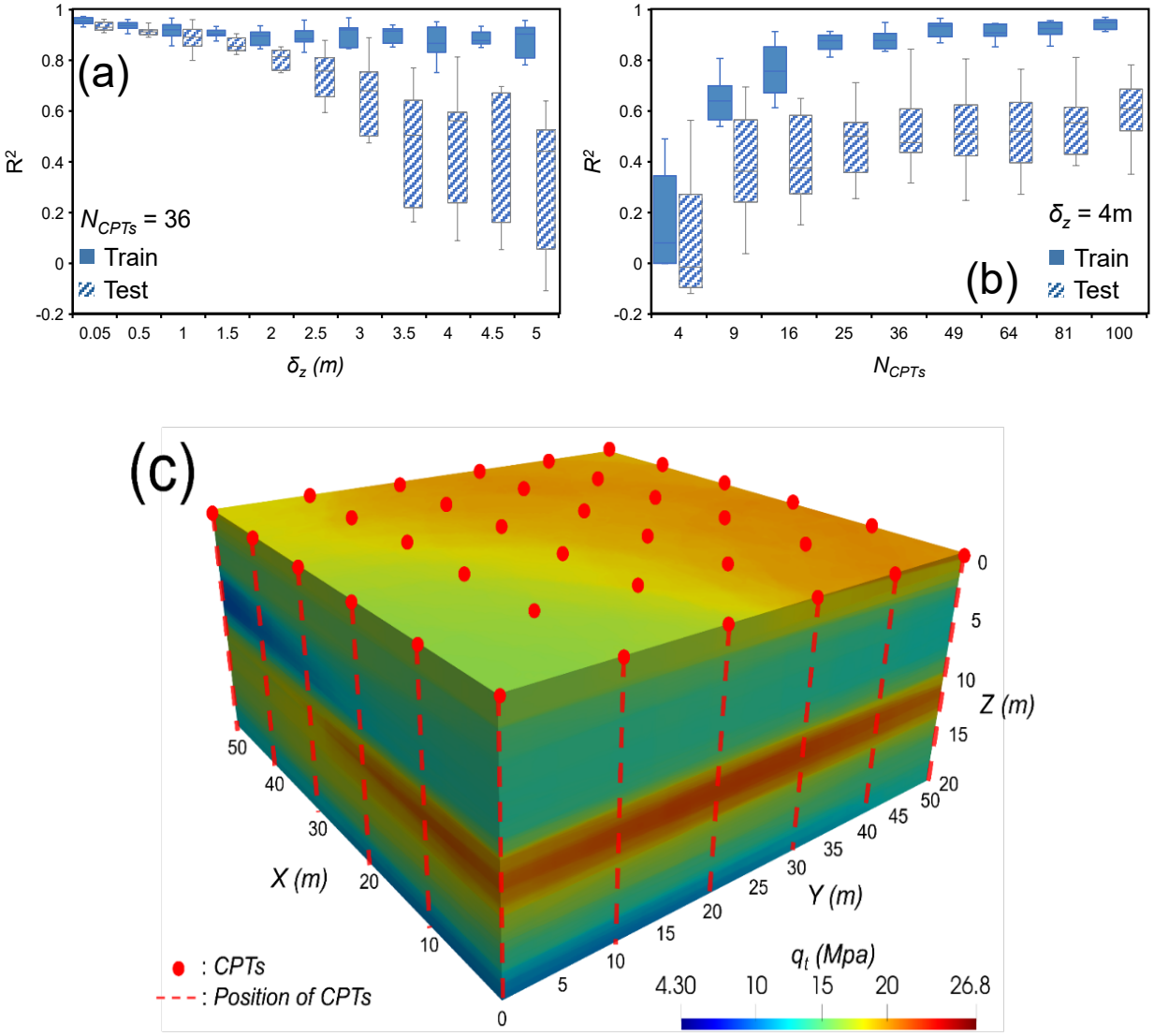


Figure 6. (a) RF-GDF model R^2 values as a function of δ_z at N_{CPTs} = 36, (b) RF-GDF model R^2 values as a function of N_{CPTs} at δ_z = 4 m, (c) predicted q_t field at δ_z = 4 m.

3. CONCLUSIONS

This study proposed an ML-based GDF model for reliable 3D geotechnical interpolation. Among the evaluated ML algorithms, RF demonstrated the best performance and generalization capability. Increased uncertainty at lower N_{CPTs} indicates

the necessity of adequate investigation interval, an N_{CPTs} value of 36 (i.e., CPT spacing of 10 m) provided reliable interpolation with low uncertainty. In addition, comparable q_t predictions between sparse (δ_z = 4 m) and dense (δ_z = 0.05 m) vertical intervals suggest the robustness and applicability of the developed framework with limited site investigation data, emphasizing the practical use of the proposed framework in geotechnical interpolations.

REFERENCES

Ahrens, J., Geveci, B., & Law, C. (2005). ParaView: An End-User Tool for Large-Data Visualization. In *Visualization Handbook* (pp. 717–731). Elsevier. https://doi.org/10.1016/B978-012387582-2/50038-1

Eid, M., Hefny, A., Sorour, T., Zaghloul, Y., & Ezzat, M. (2018). Full-scale well instrumented large diameter bored pile load test in multi layered soil: a case study of damietta port new grain silos project. *Int. J. Curr. Eng. Technol*, *8*(1), 85–98.

Isaaks, E. H., & Srivastava, R. M. (1989). *Applied geostatistics*.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems*, *26*.

Müller, S., Schüler, L., Zech, A., & Heße, F. (2022). GSTools v1.3: a toolbox for geostatistical modelling in Python. *Geoscientific Model Development*, *15*(7), 3161–3182. https://doi.org/10.5194/gmd-15-3161-2022

Xie, J., Huang, J., Zeng, C., Huang, S., & Burton, G. J. (2022). A generic framework for geotechnical subsurface modeling with machine learning. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(5), 1366–1379. https://doi.org/10.1016/j.jrmge.2022.08.001