Digitization of subsurface geological stratigraphy using machine learning and neighborhood aggregation

Yue Hu[1#], *Ze Zhou* Wang[2], and *Xiangfeng* Guo[3]

[1]Institute for Risk and Reliability, Leibniz Universität Hannover, Hannover 30167, Germany
[2]Civil Engineering Division, Department of Engineering, University of Cambridge, Cambridge CB3 0FA, United Kingdom
[3]School of Marine Science and Engineering, South China University of Technology, Guangzhou 511442, China.
[#]Corresponding author: yue.hu@irz.uni-hannover.de

ABSTRACT

In engineering geology and geotechnical engineering, subsurface soils and rocks are natural geomaterials and exhibit inherent variability in stratigraphy due to geological deposition process. Explicit knowledge of subsurface stratigraphy is a critical input for the analysis, design, and construction of geotechnical engineering systems. However, the accurate and reliable modelling of subsurface geological stratigraphy is challenging due to the limited number of available boreholes in practice and the complex nature of soil stratigraphy. This paper presents an innovative machine learning framework built upon the neighborhood aggregation technique for the prediction of digitized subsurface geological stratigraphy. To predict the stratigraphy at a given point of interest, neighborhood aggregation is first performed to intelligently consolidate the stratigraphy information from its neighboring boreholes, resulting in additional features associated with the target location. By combining the extra stratigraphy information with conventional location-specific features, the framework enhances the predictive capabilities of classical machine learning models at a finer scale. The proposed framework is implemented using common machine learning models and is validated using a simulated benchmark 3D example. The results of leave-one-out cross-validation demonstrate that the proposed framework can improve the performance of classical machine learning models, leading to more reasonable stratigraphy transition and associated uncertainty quantification.

Keywords: Geological models; Geotechnical database; Machine learning; Graphical network.

1. Introduction

Modern urban development has increasingly expanded to the underground space due to the increasing demand for land and limited space above ground. To ensure a safe and efficient development of underground projects, explicit knowledge of the spatial distribution of subsurface stratigraphy is essential for the analysis, design, and construction of geotechnical systems. However, subsurface geological stratigraphy is the result of the complex deposition process and loading history which are spatially varying, leading to significant uncertainty in subsurface stratigraphy and soil properties (e.g., Juang et al., 2019; Phoon et al., 2022). In addition, the limited availability of boreholes at a specific site poses further uncertainty to stratigraphy characterization and decision making (e.g., Wang et al., 2022a).

To this end, a series of traditional models/algorithms has been developed in the past decades. Among the existing traditional methods, the random field and the geostatistical models are frequently used for depicting the spatially varying boundaries of different geological units based on assumed trend and autocorrelation functions (e.g., Cardenas, 2023). The Markov chain model is another class of traditional methods that can predict the full spatial distribution of soil stratigraphy cross-sections by utilizing assumed soil transition probability matrices along different directions (e.g., Qi et al., 2016). It is worth noting that the accurate determination of autocorrelation functions or the soil transition probability matrix is often challenging in the presence of sparse boreholes at a specific site (e.g., Wang et al., 2022b).

In addition, traditional methods are mostly applied to local-scale specific sites and may not well handle regional geological modelling involving large number of sites. Within the recent trend of digital transformation of geotechnical engineering, many machine learning (ML) techniques have been adopted for the purpose of geotechnical site characterization and have shown great potential (e.g., Wang et al., 2019; Wu et al., 2021). Generally, extensive boreholes data is a prerequisite to reliably train a machine learning model, by leveraging the available compiled database of many sites. However, the trained machine learning models may still exhibit low definition and generalizability when used between two sparse boreholes at a specific site.

This study presents an innovative machine learning framework built upon the neighborhood aggregation technique for improved prediction of digitized subsurface stratigraphy. In contrast to traditional machine learning models, it can automatically adapt to available borehole data and achieve improved performance at the local scale. A simulated benchmark example is illustrated.

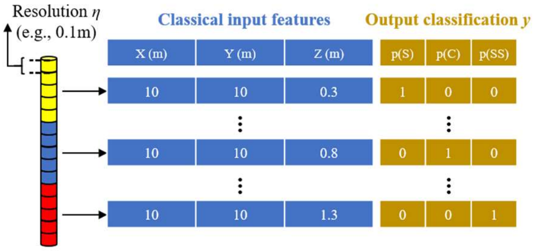
2. Proposed framework

The overall structure of the proposed new machine learning framework includes three components: (1) preprocessing of borehole log; (2) development of additional features using neighborhood aggregation, and (3) machine learning modelling. The three components are described as follows.

2.1. Preprocessing of borehole log

To model the multi-layer pattern of 3D subsurface stratigraphy, layer-wise records of available borehole log are firstly digitized. As shown in Figure 1, the borehole log illustrated on the left contains three soil layers, e.g., a 0.5m-thick sand layer on the top, a 0.5m-thick silty-sand layer on the bottom and a 0.5m-thick clay layer in-between. This borehole has a depth of 1.5m and can be digitized into 15 discrete grid points along the depth with a digitization resolution η=0.1m. The digitization resolution η is the actual spatial extent represented by a discrete grid point in the vertical direction.

To enable 3D subsurface stratigraphy modelling, each discrete grid point is assigned with the 3D spatial coordinates information (e.g., X in meters, Y in meters, and Z in meters) as the classical three-input features, as denoted by the blue tables in Figure 1. The output label of each discrete point is the corresponding soil classification category $\boldsymbol{y}$. For example, the three considered soil classification categories, i.e., sand, clay, and silty-sand, are denoted by 1, 2, and 3, respectively. To eliminate the effects of the order/correlation assigned to these categorical values, a commonly used one-hot encoder is adopted to transform the nominal categorical values to one-hot vectors, as shown by the yellow table in Figure 1.



X (m)	Y (m)	Z (m)	p(S)	p(C)	p(SS)
10	10	0.3	1	0	0
	⋮			⋮	
10	10	0.8	0	1	0
	⋮			⋮	
10	10	1.3	0	0	1

Figure 1. Basic features development for a digitized borehole

2.2. Development of additional new features using neighborhood aggregation

To improve the prediction performance of machine learning models at the local scale, additional features besides the basic coordinate features are developed for digitized boreholes using neighborhood aggregation. This step essentially involves feature engineering in machine learning. Neighborhood aggregation is a technique used in graph network learning (e.g., graph convolution networks), where local neighborhood information of a target node is combined or aggregated to the target node as additional new features for improving prediction accuracy. In the field of graph network learning, it has shown advantages in optimizing classical models (e.g., Schlichtkrull et al., 2018).

Figure 2 illustrates the procedure of developing depth-based additional features for a target borehole by neighborhood aggregation. For example, five available boreholes scatter around a target borehole denoted in grey in the middle. Each neighboring borehole has incorporated different stratification patterns. At each depth level, the soil classification information presented in all neighboring boreholes is extracted for developing additional features of the target borehole. Technically, additional new features at a specific digitized grid point of the target borehole, denoted by $\hat{\boldsymbol{y}}_{\mathrm{T}}$, can be created by an aggregating function f of its neighboring digitized grid points from neighboring boreholes (e.g., Wang et al., 2023):

$$\hat{\boldsymbol{y}}_{\mathrm{T}} = f(\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}) \tag{1}$$

where $\boldsymbol{y}_i$ is the true soil classification (in terms of one-hot vector) of the i-th digitized grid point from neighboring boreholes at the depth level; n is the number of neighboring digitized grid points. Inverse distance weighting (IDW), a commonly used interpolator, is adopted as the aggregation function (e.g., Lu and Wang, 2008). The new features $\hat{\boldsymbol{y}}_{\mathrm{T}}$ are then combined with the classical input features (i.e., X, Y, and Z) to form the new input features. In other words, at each grid point, the three classical input features are appended with the three additional new features corresponding to the probability measure of the occurrence of sand (S), clay (C), and silty-sand (SS) at that grid point, leading to the six-input features tabulated in blue at the bottom of Figure 2. This process iterates until the additional features are developed for all digitized grid points along the target borehole.

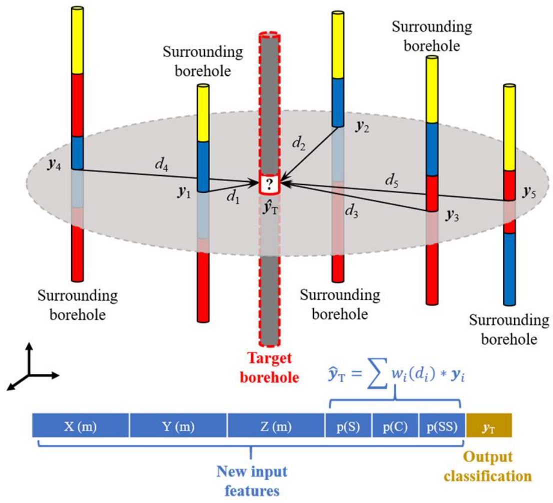


Figure 2. Additional features development for a digitized borehole

2.3. Machine learning modelling

The last step of the proposed framework involves machine learning modelling. Among the many available machine learning models, random forest (RF) model is adopted in this study for illustration. RF is a bagging ensemble learning method that consists of multiple decision trees constructed by randomly selecting subsets of the complete feature space and the training samples (e.g., Ho, 1998). Each decision tree is built independently, and the results from all decision trees are then combined to make the final prediction. The RF prediction is formulated as:

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{f}_{k}(\boldsymbol{x}_{\boldsymbol{i}}) \qquad (2)$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ is the prediction for the input feature $\boldsymbol{x}_{\boldsymbol{i}}$ from the decision tree system; $\boldsymbol{x}_{\boldsymbol{i}}$ is the i-th input vector (e.g., the obtained six input features); $\boldsymbol{f}_{k}$ is the output of the k-th decision tree (e.g., one-hot vector of soil classification). The best estimate of the soil class is by a majority vote from all decision trees. The associated prediction uncertainty can be quantified by the summed probability estimates for soil classification other than the best estimate.

3. Illustration using a benchmark example

In this section, the proposed machine learning framework is illustrated using a 3D stratigraphy benchmark example "S-VG2" proposed by Phoon et al., (2022). As shown in Figure 3, the 3D stratigraphy example is 20 m long × 20 m wide × 10 m deep, along the X, Y, and Z directions, respectively. The model is discretized into 20 × 20 × 100 = 40,000 cells. The actual resolution of each cell is therefore 1 m long × 1 m wide × 0.1 m deep. The configured geometry of this 3D example is likely a representative of the scale of a typical "small" project site. In this example, three soil layers, i.e., sand, clay, and silty-sand, are distributed along the depth direction. To incorporate the spatial variability of soil stratigraphy, the two boundaries between soil layers are inclined to different extents. The boundaries between sand and clay, and clay and silty-sand are respectively defined as (e.g., Phoon et al., 2022):

$$0.05\mathrm{X}-0.05\mathrm{Y}+\mathrm{Z}-2=0 \qquad (3)$$

$$-0.05\mathrm{X}+0.05\mathrm{Y}+\mathrm{Z}-6=0 \qquad (4)$$

The benchmark example "S-VG2" as shown in Figure 3 represents the complete state of "reality". Following the benchmark procedure of data-driven site characterization (DDSC) methods (Phoon et al., 2022), only a subset of boreholes is selected from this 3D example as the training dataset which represents the measured "reality" of the 3D stratigraphy. The "T2" training scenario in Phoon et al., (2022) is adopted and denoted by black triangles in Figure 3. Twelve testing boreholes which are distinct from the training boreholes are denoted by red crosses, with testing borehole number labelled. A 3D visualization of the six training boreholes and 12 testing boreholes is explicitly shown in Figure 4. It reveals that the sand and silty-sand layers tend to be thinner, and the clay layer tend to be thicker towards (X=20, Y=0).

To implement the proposed framework, additional features are developed for the digitized grid points of both training boreholes and testing boreholes. Note that for each of the six training boreholes, additional features are developed using the remaining five training boreholes in the neighborhood. For testing boreholes, all six training boreholes are used for developing additional features. Subsequently, an RF model is trained based on the digitized training boreholes with the new input features. The performance of the RF model is then evaluated at the 12 testing boreholes.

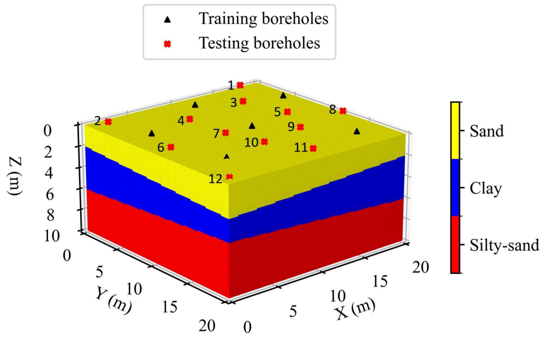


Figure 3. A benchmark 3D soil stratigraphy example

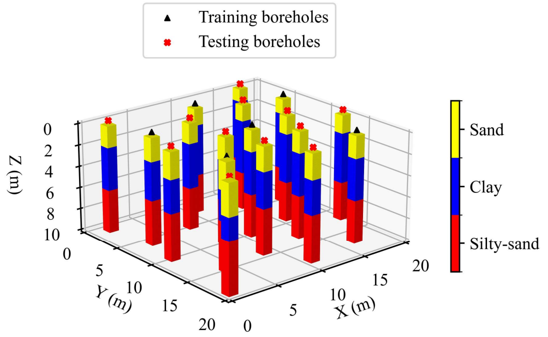


Figure 4. Training and testing boreholes

Table 1. Prediction accuracy summary of RF model at 12 testing boreholes

Borehole #	Classical features	New features
TB1	91%	94%
TB2	93%	93%
TB3	98%	99%
TB4	94%	94
TB5	93%	96%
TB6	95%	97%
TB7	97%	98%
TB8	95%	94%
TB9	99%	100%
TB10	95%	96%
TB11	93%	94%
TB12	85%	90%

The prediction performances of RF with classical input features and new input features at the 12 testing boreholes (e.g., TB1-TB12) are summarized respectively in Table 1. Since the "S-VG2" is a benchmark example exhibiting simply linear stratigraphy patterns, the overall prediction accuracies at these 12 testing boreholes are high, e.g. exceeding 85% for both feature scenarios.

A cone penetration test (CPT)-based benchmark study in Phoon et al., (2022) reported that among the 12 testing boreholes, the best, median, and worst prediction accuracies obtained by GLasso method in this scenario are 0.82, 0.74, and 0.56, respectively. This suggests that direct modelling of the soil stratigraphy, rather than the associated CPT measurement, yields better accuracy in terms of soil stratification.

In addition, as shown in Table 1, the incorporation of new features leads to a slight improvement at 9 out of 12 testing boreholes. In particular, a 5% improvement in accuracy is achieved at TB12. As illustrated in Figure 3, the testing borehole TB12 is located around (X=0, Y=20), which is outside the range of the six training boreholes and presents a quite thin pattern of the clay layer. In other words, extrapolation of TB12 is a challenging task. The detailed prediction performances at TB 12 are shown in Figure 5. It is observed that the actual thickness of the clay layer in the middle is relatively thin, while the best estimates of two feature scenarios both overestimate the thickness of clay layer. The RF with new features slightly rectifies the best estimate.

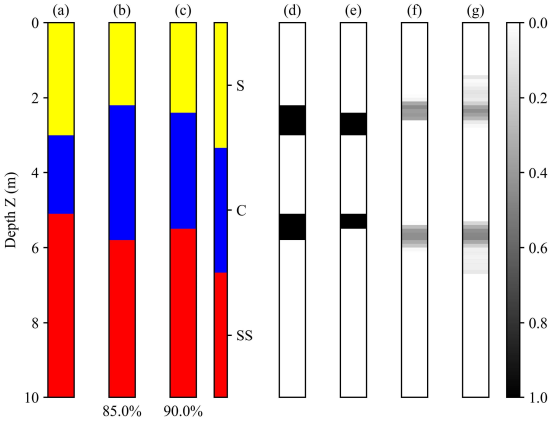


Figure 5. Cross-validation results of testing borehole #381 using random forest algorithm: (a) true borehole; (b) best estimate of RF with classical features; (c) best estimate of RF with new features; (d) prediction error of (b); (e) prediction error of (c); quantified uncertainty of (b); quantified uncertainty of (c)

The proposed machine learning framework is also applicable to complete 3D subsurface stratigraphy modelling. Based on the six training boreholes, additional features are developed for each digitized grid point at all the remaining locations. The trained RF model described above can be re-used to predict the digitized 3D stratigraphy model. The best estimates of 3D soil stratigraphy using RF with classical features and new features are shown in Figure 6 and Figure 7, respectively.

The general distributions of soil stratification are comparable in these two scenarios. However, the stratigraphic boundaries identified in Figure 6 yield unrealistic "step" and noisy patterns. In contrast, the results shown in Figure 7 incorporate less noisy patterns and show smoother transition of stratigraphic boundaries. In addition, the quantified uncertainty of the two feature scenarios are shown in Figure 8 and **Error! Reference source not found.**, respectively. In both feature scenarios, uncertainty bands around the soil stratigraphic boundaries are clearly identified. However, the uncertainty band in **Error! Reference source not found.** appears to be thicker, suggesting possible variability in the stratification boundaries, as shown in Figure 3. The results demonstrate that the developed additional features improve the prediction performance of RF models.

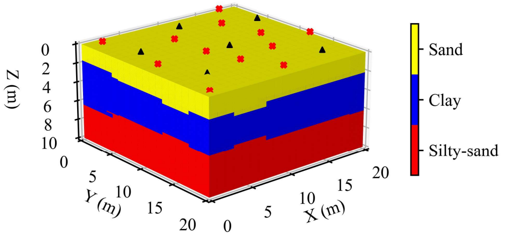


Figure 6. Best estimate of 3D soil stratigraphy prediction using random forest with classical features

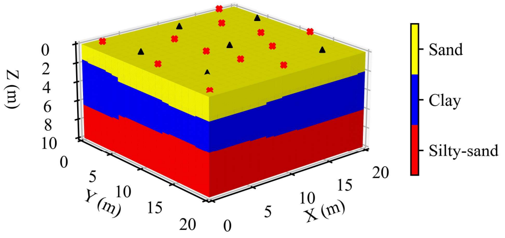


Figure 7. Best estimate of 3D soil stratigraphy prediction using random forest with new features

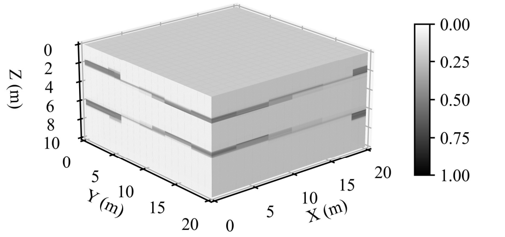


Figure 8. Uncertainty of 3D soil stratigraphy prediction using random forest with classical features

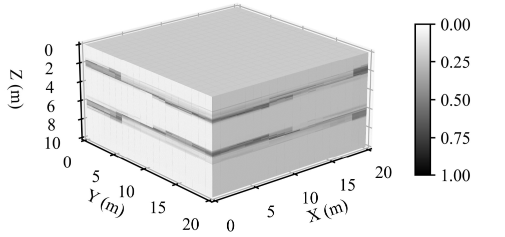


Figure 9. Uncertainty of 3D soil stratigraphy prediction using random forest with new features

4. Conclusions

In this paper, an innovative machine learning framework built upon the neighborhood aggregation technique is presented for the improved prediction of digitized subsurface stratigraphy. The development of additional features using neighborhood aggregation is elaborated. A benchmark example in the literature is used to evaluate the performance of the proposed framework. Results show that direct modelling of soil stratigraphy rather than the associated CPT measurement yields better accuracy in terms of soil stratification. The additional new features lead to improved prediction performance of RF model at the local scale and smoother stratigraphic boundaries, in comparison to the classical input features.

Acknowledgements

The first author is grateful for the financial support provided by the Alexander von Humboldt Foundation, Germany.

References

Cardenas, I.C. 2023. A two-dimensional approach to quantify stratigraphic uncertainty from borehole data using non-homogeneous random fields. *Engineering Geology*, 107001.

Ho, T.K. 1998. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20, 832-844.

Juang, C.H., Zhang, J., Shen, M. and Hu, J., 2019. Probabilistic methods for unified treatment of geotechnical and geological uncertainties in a geotechnical analysis. *Engineering Geology*, 249, pp.148-161.

Lu, G.Y. & Wong, D.W. 2008. An adaptive inverse-distance weighting spatial interpolation technique. Computers & Geosciences, 34, 1044-1055.

Phoon, K.K., Cao, Z.J., Ji, J., Leung, Y.F., Najjar, S., Shuku, T., Tang, C., Yin, Z.Y., Ikumasa, Y. and Ching, J., 2022. Geotechnical uncertainty, modeling, and decision making. *Soils and Foundations*, 62(5), p.101189.

Phoon, K.K., Shuku, T., Ching, J. and Yoshida, I., 2022. Benchmark examples for data-driven site characterisation. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 16(4), pp.599-621.

Qi, X.H., Li, D.Q., Phoon, K.K., Cao, Z.J. and Tang, X.S., 2016. Simulation of geologic uncertainty using coupled Markov chain. *Engineering Geology*, 207, pp.129-140.

Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I. and Welling, M., 2018. Modeling relational data with graph convolutional networks. *In The Semantic Web: 15th International Conference, ESWC 2018*, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15 (pp. 593-607). Springer International Publishing.

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(8), pp.1184-1205.

Wang, Y., Shi, C. and Li, X., 2022a. Machine learning of geological details from borehole logs for development of high-resolution subsurface geological cross-section and geotechnical analysis. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 16(1), pp.2-20.

Wang, Y., Hu, Y. and Phoon, K.K., 2022b. Non-parametric modelling and simulation of spatiotemporally varying geo-data. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), pp.77-97.

Wang, Z.Z., Hu, Y., Guo, X., He, X., Kek, H.Y., Ku, T., Goh, S.H. and Leung, C.F., 2023. Predicting geological interfaces using stacking ensemble learning with multi-scale features. *Canadian Geotechnical Journal*, 60(7), 1036-1054.

Wu, S., Zhang, J.M. and Wang, R., 2021. Machine learning method for CPTu based 3D stratification of New Zealand geotechnical database sites. *Advanced Engineering Informatics*, 50, p.101397.