Data-driven multistage sampling strategy for machine learning of underground digital twins considering stratigraphic uncertainty

Chao Shi[1#], *Yu* Wang[2], and *Viroon* Kamchoom[3]

[1]*Nanyang Technological University, School of Civil and Environmental Engineering, 50 Nanyang Avenue, Singapore*
[2]*City University of Hong Kong, Department of Architecture and Civil Engineering, Tat Chee Avenue, Kowloon, Hong Kong, China*
[3]*King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand*

[#]*Corresponding author: chao.shi@ntu.edu.sg*

ABSTRACT

A sound understanding of subsurface geological conditions is crucial for the digitalisation of underground infrastructure. The building and updating of underground digital twins heavily rely on sparse geotechnical measurements (e.g., boreholes) retrieved from the ground, and an efficient sampling strategy can facilitate the interpretation of subsurface heterogeneities. Geotechnical sampling design can be viewed as a constrained optimization process that aims to obtain as much geological information as possible from a limited number of sampling locations within a given site boundary. In this study, a data-driven intelligent sampling strategy is proposed to optimize borehole locations for a multi-stage site investigation of a three-dimensional (3D) geological domain. The initial sampling plan is determined using weighted centroidal Voronoi tessellation, which assigns uniform sampling densities to zones of different importance. Measurements obtained from the initial stage are combined with prior geological knowledge to build underground digital twins using an image-based stochastic modelling method. Multiple realizations of the geological domain can be developed under the framework of Monte Carlo simulation, and stratigraphic uncertainties associated with multiple random realizations can be quantified using information entropy. The location with the maximum entropy is adaptively selected as the next optimal sampling location. The proposed method is the first sampling strategy that can explicitly consider 3D subsurface stratigraphic variations. The performance of the proposed multi-stage sampling strategy is demonstrated using a simulation example. Results indicate that the proposed method can efficiently identify the optimal sampling locations while accounting for irregular site geometries and 3D subsurface stratigraphic uncertainties.

Keywords: Geotechnical site characterization, smart sampling, site optimization, subsurface heterogeneities.

1. Introduction

The purpose of geotechnical site investigation is to explore subsurface heterogeneities (e.g., stratigraphic distribution) using site-specific data, such as boreholes and cone penetration tests. Due to budget limits and site constraints, only limited boreholes or soil/rock samples are retrieved from the ground. Therefore, how to determine the optimal sampling locations to obtain as much site-specific information as possible is a key challenge. Conventional site sampling strategies are empirical and often involve equal sampling spacing with regular patterns. For example, Eurocode 7-2 (EN-1992-2, 2007) recommends grid patterns with typical sampling spacings of 15 ~ 40 m and 25 ~ 75 m for high-rise buildings and dams, respectively. However, this empirical sampling strategy assigns equal importance to each location and does not consider project-specific requirements. For instance, dense samples should be retrieved from areas that are susceptible to ground movements.

Sampling design and site optimization have been an important topic in different disciplines, such as geosciences (McBratney and Webster 1981), and geotechnical engineering (Wang and Li 2021). Mathematically speaking, site planning can be formulated as a constrained optimization problem that aims to minimize uncertainties and impacts on subsequent engineering design and analysis. In geotechnical engineering domain, many previous studies have attempted to optimize site investigation schemes using either random field theory or machine learning strategies. Zhao and Wang (2019) and Wang and Li (2021) proposed Bayesian Compressive Sensing (BCS) to predict spatially varying soil properties and leveraged the theory of information entropy to optimize a multi-stage sampling process. However, previous studies mainly focused on the characterization of soil property spatial variability, and there is a lack of robust strategies to optimize site investigation will full consideration of three-dimensional subsurface stratigraphic uncertainty.

To address the above-mentioned challenges, this study proposes a smart sampling strategy that can effectively explore uncertainties associated with

subsurface stratigraphy as well as project-specific constraints to optimize site investigation schemes. The strategy can flexibly determine the initial sampling locations, taking full account of project-specific needs. The obtained samples are then integrated with prior geological knowledge for stochastic modelling of subsurface geological domains. Subsequently, the quantified stratigraphic uncertainty is leveraged to specify new sampling locations for the next round of site investigation. The proposed method is the first data-driven smart sampling strategy that explicitly considers 3D stratigraphic uncertainties and irregular site geometries. The performance of the proposed method is demonstrated through a simulation example.

2. Framework of the proposed smart sampling method

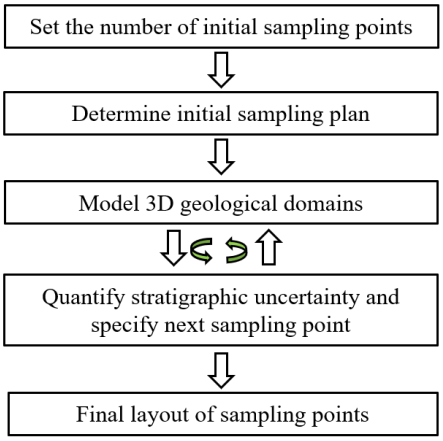


Figure 1. Framework of the smart sampling strategy

Figure 1 shows the framework of the proposed smart sampling strategy. The strategy starts with the determination of site and building boundary. For a given number of sampling points, the initial sampling plan is determined using Weighted Centroidal Voronoi Tessellation (WCVT), a plane partition tool. The obtained site-specific samples are then integrated with a stochastic simulation method called 3D Iterative Convolution eXtreme Gradient Boost (IC-XGBoost3D) for predicting 3D subsurface geological domains. The associated stratigraphic uncertainty can be quantified and leveraged for specifying the optimal next sampling location based on the theory of information entropy. The whole process can be repeated until the planned project budget is reached. In the following subsections, key components of the proposed framework are discussed in detail.

2.1. Voronoi tessellation

Voronoi tessellation is a plane partition algorithm that can divide a plane into a series of regions, i.e., $V(P_1)$, $V(P_2)$,…,$V(P_n)$, and each region can be represented by a single point called "seed". Figure 2 shows a partitioned plane using Voronoi tessellation. In total, the plane has eight Voronoi cells and seeds. Any point within a Voronoi region has a smaller Euclidean distance d to its seed than to any other Voronoi region. There are many algorithms that can be used to create a centroidal Voronoi tessellation (CVT), such as Lloyd algorithm (Lloyd 1982).

For geotechnical engineering applications, it is always preferred to assign more sampling points to areas with higher technical or economic importance. For instance, dense samples should be located within the building boundary as shown in Figure 2, where the ground is more susceptible to settlements due to vertical surcharge imposed by superstructures. To address this concern, the weighted centroidal Voronoi tessellation (WCVT) that assigns different weights (ω) to seeds of different Voronoi regions can be adopted. Mathematically, the dominance region of a weighted seed can be described as follows:

$$V(P_i) = \left\{ \boldsymbol{x} \in \mathbb{R}^2 \middle| \frac{\|x-x_i\|}{\omega_i} < \frac{\|x-x_j\|}{\omega_j} \; for \; j = 1, \dots, N_p, j \neq i \right\} \quad (1)$$

where $\boldsymbol{x}$ denotes the coordinate of a point in space; $\boldsymbol{x}_i$ and ω_i represent the coordinate and weight associated with the i-th seed. As a first approximation, the 2D site can be divided into two discrete zones, i.e., major construction zone Ω_M and ancillary construction zone Ω_A. The ratio for weights associated with Ω_M and Ω_A can be taken to be proportional to the investment or construction budget ratio. The budget ratio (BR) is defined as the ratio of the construction budgets for the areas Ω_M and Ω_A:

$$BR = \frac{Budget \; for \; construction \; in \; major \; area}{Budget \; for \; constuction \; in \; ancillary \; area} = \frac{\omega_M}{\omega_a} \quad (2)$$

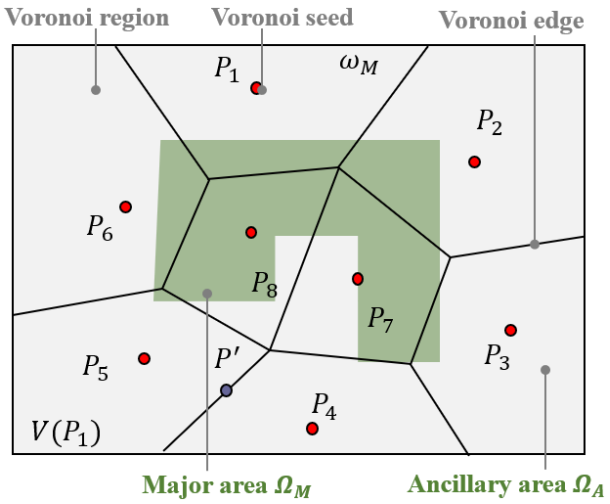


Figure 2. An illustration of Voronoi tessellation

2.2. Machine learning of three-dimensional geological domain

Once site-specific data are retrieved from the ground following the initial sampling plan discussed in subsection 2.1, 3D subsurface geological domains can be developed using a stochastic modelling method, i.e., IC-XGBoost3D (Shi and Wang 2022). IC-XGBoost3D relies on prior geological knowledge reflected in a single training image and site-specific data for stochastic simulations. Figure 3 shows the key modelling procedure of IC-XGBoost3D. As shown in Figure 3a, the single training image and site-specific data are aligned with a 3D geological domain. The training image reflects representative stratigraphic patterns at the site of interest,

and developed geological cross-sections from nearby sites with similar geological settings can readily be taken as training images. Subsequently, the whole 3D geological domain can be divided into a series of 2D simulation slices. The simulation sequence is determined based on the principle that the current simulation slice has the maximum number of site-specific data. Following the simulation sequence, 2D simulation slices are developed. Any previously simulated slice is treated as additional site-specific data. After all the 2D simulation slices have been developed, a 3D geological domain Z can be obtained by assembling all the simulated 2D slices. By changing the random seed to generate multiple random 2D simulation sequences, it is possible to generate multiple 3D geological domains, i.e., Z_1, Z_2,…, Z_n. Detailed implementation procedures can refer to Shi and Wang (2022).

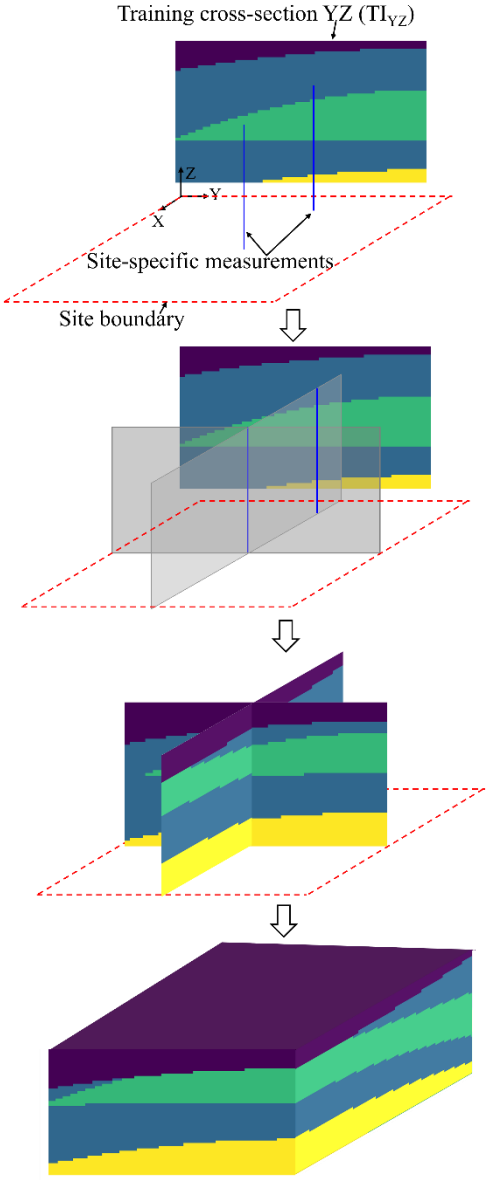


Figure 3. Development of 3D geological domain using IC-XGBoost3D (modified from Shi et al. 2023)

2.3. Uncertainty quantification

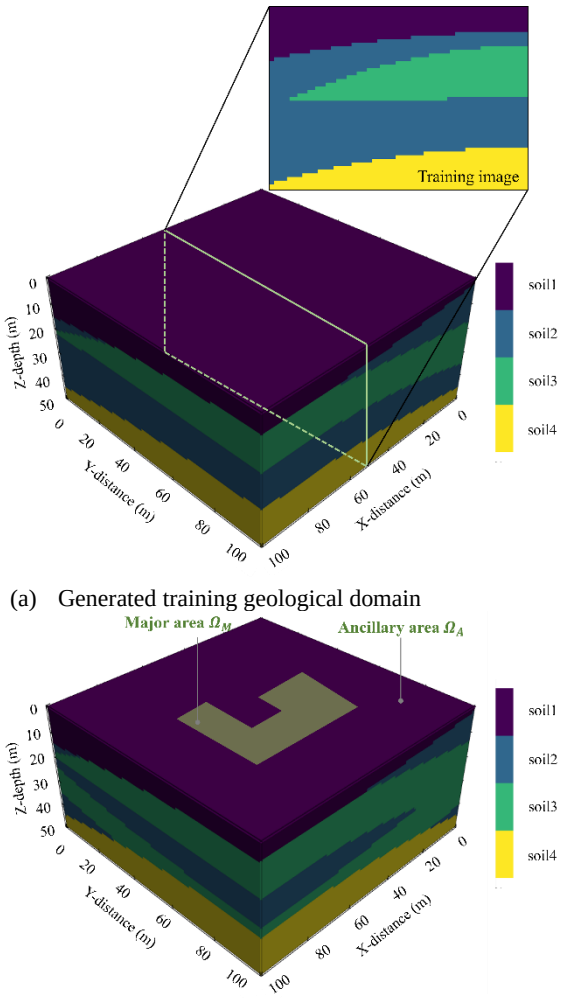
Multiple geological domains can be generated following different random seeds, and the most probable prediction Z_{mp} can be derived by assigning each spatial point with the soil category of the highest occurrence frequency. For illustrative examples where the ground truth geological domain Z_T is available, the prediction accuracy can be calculated as follows:

$$Acc^{3D} = \frac{\sum_{i=1}^{N_X \times N_Y \times N_Z} I[Z_T(x_i^{3D}) = Z_{mp}(x_i^{3D})]}{N_X \times N_Y \times N_Z} \quad (3)$$

where N_X, N_Y, and N_Z stand for the total number of voxels in the X, Y, and Z directions. Meanwhile, the stratigraphic uncertainty associated with Z_{mp} can be quantified using the theory of information entropy. Assuming the occurrence probability of the i-th soil type at $\boldsymbol{x}$ is p_i, the total entropy (H) at a given spatial location is expressed as follows:

$$H(\boldsymbol{x}) = -\sum_i^{N_c}\{p_i \cdot ln\, p_i\} \quad (4)$$

where N_c denotes the total number of soil categories at the site of interest. Areas with a large entropy value denotes a high level of stratigraphic uncertainty.



(a) Generated training geological domain

(b) Ground truth geological domain

Figure 4. Simulated geological domains (modified from Shi et al. 2023)

2.4. Smart determination of next sampling location

Additional sampling locations should be placed in areas with relatively larger entropy values. As geotechnical site investigation always involves vertical line measurements (e.g., boreholes), it is worthwhile to

integrate the calculated entropy values in Eq. (4) along the depth. Once a new measurement is retrieved from the ground, the total entropy at the selected location will reduce to zero. Therefore, the location with the maximum total entropy Δ_{n_b+1} in the 2D plan should be selected as the next sampling point $n_b + 1$:

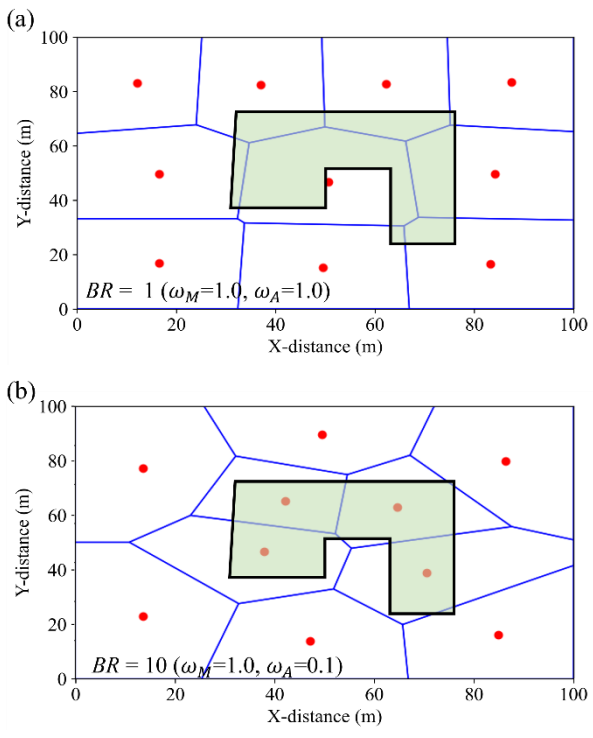
$$\Delta_{n_b+1} = H(\boldsymbol{x}_h^{2D}, n_b) - H(\boldsymbol{x}_h^{2D}, n_b + 1) \quad (5)$$

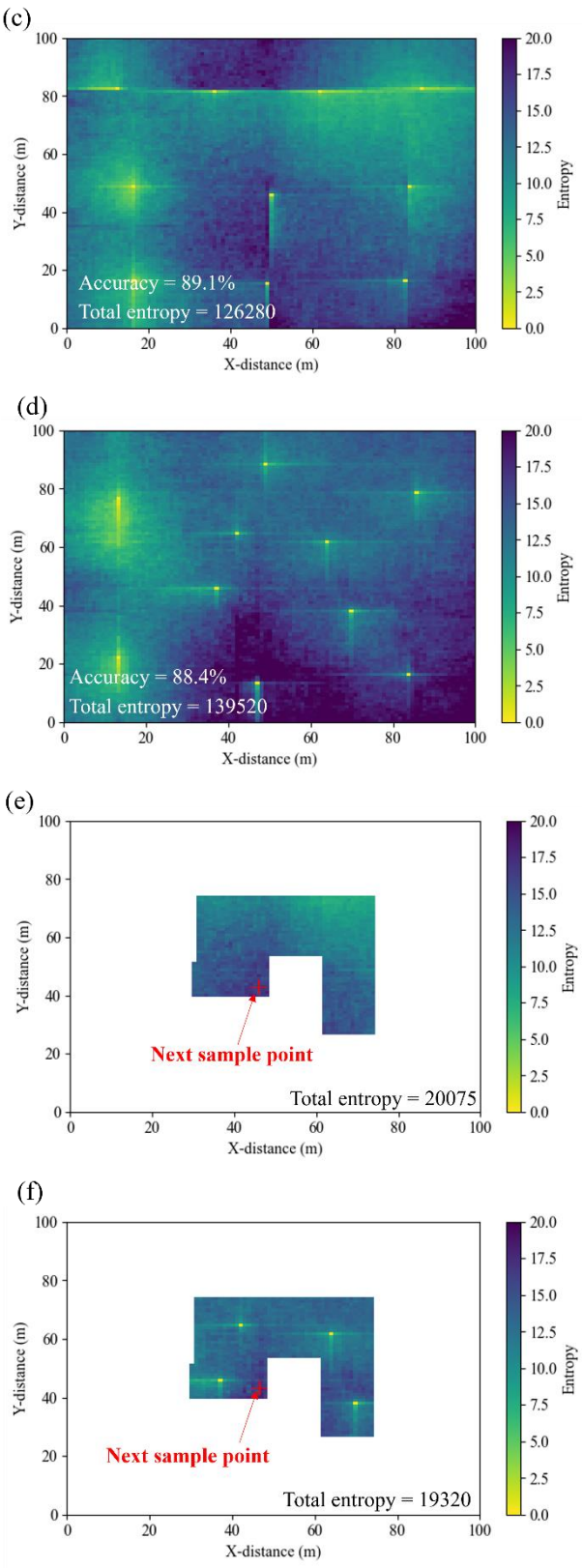
where n_b denotes the number of existing line measurements.

3. Illustrative example

Figure 4 shows the simulated training and ground truth geological domains. The boundaries separating different soil types are taken to follow gaussian distributions. For illustrative purposes, a 2D geological cross-section is taken from the training geological domain (see Figure 4a) at X = 50 as the single training image. Note that the training image shares the similar geological patterns as those of the ground truth geological domain in Figure 4b. As an illustration, the number of boreholes for the initial site investigation is set at 10.

4. Results from the proposed method





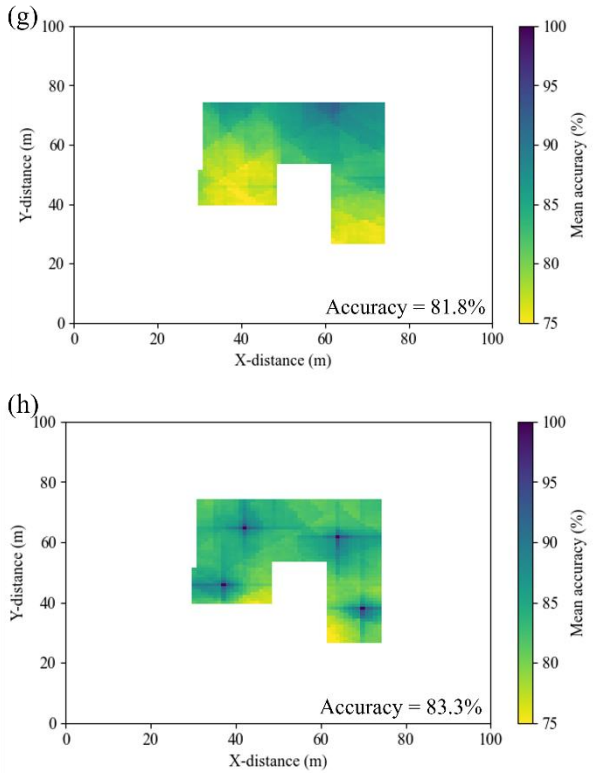


Figure 5. Comparison of different sampling strategies: (a) Sampling plan with near equal spacing at budget ratio (*BR*) = 1.0; (b) Sampling plan with *BR* = 10; (c) Entropy colormap for *BR* = 1; (d) Entropy colormap for *BR* = 10; (e) Entropy colormap for major construction zone with *BR* = 1; (f) Entropy colormap for major construction zone with *BR* = 10; (g) Accuracy colormap for major construction zone with *BR* = 1; (h) Accuracy colormap for major construction zone with *BR* = 10

As an illustration, two budget ratios (*BR*), i.e., 1 and 10, are considered in this study. Figure 5a shows the sampling plan with near equal spacing, which is essentially a special case (i.e., *BR* = 1.0) of the proposed method. As the total number of initial sampling points is 10, the top row consists of 4 points with reduced spacing. In addition, Figure 5b shows the sampling plan with *BR* = 10, and four sampling points are assigned within the major construction zone. Figures 5c and 5d show the total entropy colormaps for *BR* = 1 and 10, respectively. At *BR* = 1, sampling points are distributed uniformly across the entire site. The corresponding accuracy (i.e., 89.1%) is slightly larger (i.e., 88.4%) than that at *BR* = 10, and the total entropy (i.e., 126280) is slightly smaller than that (i.e., 139520) at *BR* = 10. However, when only the major construction zone is considered, the total entropy from *BR* = 10 (see Figure 5e) is smaller than that at *BR* = 1 (see Figure 5f). This is as expected as four sampling points are assigned within the major construction zone at *BR* = 10. As a result, the prediction accuracy (i.e., compare with the ground truth geological domain) at *BR* = 10 (refer to Figure 5h) is about 83.3%, which is slightly larger than 81.8% for *BR* = 1. It is also worth mentioning that the next optimal sampling point from Figures 5e and 5f are essentially coincident with locations of low prediction accuracy as shown in Figures 5g and 5h, which further demonstrates the effectiveness of the proposed method.

5. Summary and conclusion

A data-driven smart sampling strategy is proposed for multi-stage site investigation with full consideration of 3D subsurface stratigraphic uncertainty and irregular site geometries. The strategy enables the flexible determination of initial sampling locations considering project-specific needs during the preliminary stage of site investigation using weighted centroidal Voronoi tessellation. The obtained site-specific data are then integrated with prior geological knowledge for spatial predictions of 3D subsurface geological domains with quantified stratigraphic uncertainty. Subsequently, the quantified uncertainty is adopted to determine the next optimal sampling location based on the principle of maximum entropy reduction. The performance of the proposed method is demonstrated through an illustrative example. Results indicate that the data-driven approach renders efficient sampling within a site with irregular plan geometry while taking full account of the 3D subsurface stratigraphic uncertainty.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grant Council of Hong Kong Special Administrative Region (Project no. CityU 11203322), a grant from the Innovation and Technology Commission of Hong Kong Special Administrative Region (Project No: MHP/099/21), and a grant from Shenzhen Science and Technology Innovation Commission (Shenzhen-Hong Kong-Macau Science and Technology Project (Category C): No: SGDX20210823104002020), China. The research was also supported by the Ministry of Education, Singapore, under its Academic Research Fund (AcRF) Tier 1 Seed Funding Grant (Project no. RS03/23), AcRF regular Tier 1 Grant (Project no. RG69/23), and the Start-Up Grant from Nanyang Technological University. The last author acknowledges the grant (Grant No. FRB66065/0528-RE-KRIS/FF66/53) from King Mongkut's Institute of Technology Ladkrabang (KMITL) and National Science Research and Innovation Fund (NSRF). The financial support is gratefully acknowledged.

References

EN 1997-2, 2007. *Geotechnical Design — Part 2: Ground Investigationand Testing (Eurocode 7-2)*. EN 1997-1: 2007. European Committee for Standardization (CEN), Brussels, Belgium.

Lloyd, S. 1982. "Least squares quantization in PCM." IEEE Trans. Inf. Theory 28 (2), 129 – 137.

McBratney, A.B., Webster, R. 1981. "The design of optimal sampling schemes for local estimation and mapping of regionalized variables — II: Program and examples." Comput. Geosci. 7 (4), 335 – 365. https://doi.org/10.1016/0098-3004(81)90077-7.

Shi, C., Wang, Y. 2022. "Data-driven digital twin construction of subsurface three-dimensional geological domain from training images and limited site-specific boreholes using C-XGBoost3D." Tunn. Undergr. Space Technol. 126, 104493. https://doi.org/10.1016/j.tust.2022.104493.

Shi, C., Wang, Y. and Kamchoom, V. 2023. “Data-driven multi-stage sampling strategy for a three-dimensional geological domain using weighted centroidal voronoi tessellation and IC-XGBoost3D.” Engineering Geology, 325, p.107301. https://doi.org/10.1016/j.enggeo.2023.107301.

Wang, Y. and Li, P. 2021. “Data-driven determination of sample number and efficient sampling locations for geotechnical site investigation of a cross-section using voronoi diagram and bayesian compressive sampling.” Comput. Geotech. 130, 103898. https://doi.org/10.1016/j.compgeo.2020.103898.

Zhao, T. and Wang, Y. 2019. “Determination of efficient sampling locations in geotechnical site characterization using information entropy and Bayesian compressive sampling.” Can. Geotech. J. 56 (11), 1622 – 1637. https://doi.org/10.1139/cgj-2018-0286.