A machine learning paradigm for subsurface stratigraphy from sparse data

Chao Shi[1#]

[1]*Nanyang Technological University, School of Civil and Environmental Engineering, 50 Nanyang Avenue, Singapore*

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

ABSTRACT

Subsurface stratigraphy is an indispensable component of geotechnical site characterization and primarily deals with the interpretation of geological interfaces from site-specific measurements, such as boreholes. Traditional geological profiling methods often rely on engineering judgement for manual drawing or entirely depend on parametric models for interpolation. Both approaches face challenges when dealing with limited geo-data. To effectively address the dilemma, a new machine learning paradigm is proposed in this study to combine valuable prior geological knowledge and sparse site-specific measurements for data-driven predictions of both two-dimensional geological cross-sections and three-dimensional geological domains. The valuable prior knowledge is quantitatively represented as training images, which are compiled and stored in a training image database that is further enriched and augmented by employing deep generative models. Subsequently, the optimal training images that are compatible with the available site-specific data are adaptively selected for onward stochastic predictions under the framework of non-parametric Bayesian analysis. The method has been successfully applied to tackle geological profiling challenges in Hong Kong. The proposed framework is demonstrated to be capable of not only predicting the most probable geological patterns but also effectively quantifying associated stratigraphic uncertainty. The framework holds great potential of revolutionizing current engineering practices.

Keywords: Geotechnical site characterization, ensemble learning, stratigraphic uncertainty, training image database.

1. Introduction

Interpreting subsurface stratigraphy from sparse site-specific data (e.g., borehole logs) is a must for every geotechnical project, and it is also a basic task of geotechnical site characterization. Over the past several years, the determination and modelling of spatially varying geotechnical properties (e.g., Young's modulus, undrained shear strength) have been studied extensively. There has been much less attention paid to automatic modelling and development of subsurface two-dimensional (2D) geological cross-sections as well as three-dimensional (3D) geological domains from site-specific data. In engineering practice, hand digitalization using linear interpolation techniques to connect the same stratigraphic interfaces between adjacent boreholes is still the prevalent strategy for subsurface stratigraphy. This simplified practice is effective to deal with simple geology and may encounter difficulty for challenging grounds, such as interbedded soil layers, and inclined-fold strata. It is also widely acknowledged that the delineation and interpretation of complex geological structures from limited site-specific data can be effectively supplemented by engineering judgement of experienced geologists. However, the valuable prior geological knowledge has not been quantitatively leveraged to predict subsurface stratigraphy due to a lack of effective methods to integrate site-specific measurements with prior knowledge for data-driven predictions of subsurface stratigraphic distributions. It is imperative to have an effective tool to explicitly incorporate prior geological knowledge for stratigraphic modelling and quantification of associated stratigraphic uncertainty.

Although there are several advanced stochastic simulation tools dedicated to developing subsurface stratigraphy, these methods mainly apply to simple stratigraphic patterns, e.g., depositional strata. For example, kriging (e.g., Nobre and Sykes 1992), Markov random field (MRF) (e.g., Gong et al. 2020, Yan et al. 2023), and coupled Markov random chains (CMC) (e.g., Deng et al. 2020, Elfeki 2005, Li et al. 2019, Qi et al. 2016) have been developed for stratigraphic modelling. These models are parametric and require the explicit specification of parametric functions or calibration of site-specific parameters. For example, MRF methods often require the prior determination of initial configuration, and CMC approaches need stationary transition probabilities for sequential modelling of soil/rock types.

The recent rapid development in computer vision-based techniques provides new windows to address the classical stratigraphic challenge. There has been a surge in the application of machine learning algorithms to predict spatial distributions of stratigraphic boundaries. For example, deep learning techniques, such as generative adversarial networks (GANs) and Convolutional Neural Networks (CNN), have gained popularity in geological modelling (Mosser et al. 2017, Laloy et al. 2018, Zhang et al. 2021, Tang et al. 2021, Chen et al. 2023, Wang et al. 2024, Lyu et al. 2024).

However, the performance of deep learning algorithms normally requires a large amount of training images with abundant site-specific measurements, which are often unavailable in geotechnical site characterization.

To explicitly solve the abovementioned challenges, Shi and Wang (2021a, 2021b) proposed to represent valuable prior geological information at the site of interest in a single training image and then developed a single image-based machine learning algorithm for conditional predictions of subsurface stratigraphy based on limited site-specific measurements. The proposed paradigm effectively leverages prior geological knowledge and overcomes the sparse data challenge, representing a physics-informed approach for stratigraphic modelling. This study comprehensively reviews the key components of the developed machine learning paradigm for subsurface stratigraphy, including theory formulation, compilation of a domain-specific training image database, 2D and 3D geological modelling, as well as the typical applications to real engineering projects. It is worth mentioning that the proposed paradigm can not only accurately estimate the most probable subsurface geological model but also renders the explicit quantification of associated stratigraphic uncertainty in a data-driven manner.

2. Machine learning of subsurface stratigraphy

Fig. 1 shows the framework of the proposed machine learning paradigm for data-driven predictions of subsurface stratigraphy. The framework essentially aligns with the essence of nonparametric Bayesian analysis, which combines the flexibility of machine learning with uncertainty quantification for inference of subsurface geological cross-sections. One key recipe for the proposed approach is training images, which represent prior geological knowledge at the site of interest. Qualified training images can either be borrowed from nearby sites with similar geological settings or synthesized using generative models. Once qualified training images are collected, stochastic simulations of 2D and 3D subsurface stratigraphy can be carried out conditioning on sparse site-specific data and training images selected from the compiled training image database. If multiple qualified training images are available, diverse stratigraphic patterns may be extracted from multiple training images for ensemble learning of subsurface stratigraphy. The framework not only allows the estimation of the most probable prediction, but also facilitates the quantification of associated stratigraphic uncertainty. More specifically, the proposed machine learning paradigm explicitly solves the data sparsity challenge associated with the application of conventional machine learning to geotechnical site characterization. On one hand, the proposed image-based machine learning methods for 2D and 3D stratigraphic modelling require a minimum of one training image for training. On the other hand, an iterative and sequential modelling approach is implemented, which effectively overcomes the challenge associated with limited site-specific data. Key components of the proposed machine learning paradigm are discussed in detail in the following sections.

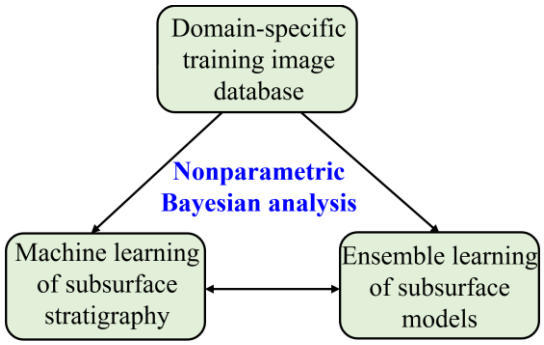


Figure 1. Framework of the machine learning paradigm

2.1. Non-parametric Bayesian analysis

Mathematically speaking, the development of subsurface geological cross-sections, *X*, from sparse data, *M*, and training images, *TIs*, can be formulated as a Bayesian prediction problem, aiming to maximize the posterior probability $P(X|\,M)$:

$$P(X|\,M) = \sum_{i=1}^{n_s} P\left(X\middle|\, TI^{(i)}, M\right) \cdot P\left(TI^{(i)}|M\right) \qquad (1)$$

where $P\left(X\middle|\, TI^{(i)}, M\right)$ is the likelihood term representing the probability of observing X for a given set of site-specific data and a single training image $TI^{(i)}$, which represents the *i*-th image sample drawn from a compiled training image database; $P\left(TI^{(i)}|M\right)$ is the conditional probability reflecting the compatibility of the selected training image sample with site-specific data; n_s denotes the total number of training images available for a particular site. In practice, the likelihood term can be estimated using image-based stochastic simulation methods, such as multiple point statistics (MPS) and Iterative Convolutional eXtreme Gradient Boost (IC-XGBoost) algorithm (Shi and Wang 2021b). In addition, a training image with high occurrence probability $P\left(TI^{(i)}|M\right)$ should be selected for forward development of geological cross-sections. Ideally, a qualified *TI* should be sampled from the underlying data generating process p_{data}. Unfortunately, the direct approximation of p_{data} is often prohibitive. However, it is possible to collect multiple qualified training image samples for ensemble learning of subsurface stratigraphy. Eq. (1) essentially represents a bagged estimate of the posterior probability $P(X|\,M)$ by combining outputs of multiple stratigraphic modelling models, and model average has been found to be an effective strategy to minimize prediction errors and improve model generalization performance (Hastie et al. 2009).

2.2. Domain-specific Training image database

The significance of training images has been emphasized for the proposed machine learning paradigm. A training image can be viewed as a prior ensemble of local geological knowledge and experience (e.g., stratigraphic connectivity between different geological domains). More importantly, a training image is a numerical representation of believed spatial stratigraphic

heterogeneities at the site of interest and should reflect the major repetitive stratigraphic relationships and structures (Mariethoz and Caers 2014). However, the acquisition of qualified training images in practice may be challenging, particularly for young engineers with insufficient prior knowledge of local geology. According to Heim (1990), the stratigraphic patterns of different natural deposits (e.g., marine deposit and alluvial strata) may be categorized based on their geological origins. In general, qualified training images for a particular site may be readily available from nearby sites with similar geological settings or from conceptual models developed by experienced geologists. It is also worth mentioning that a single training image only represents a possible geological scenario or stratigraphic configuration, which may not exhaust all the potential stratigraphic features or fully characterize the underlying data-generating process (i.e., p_{data}). A combination of multiple qualified training images constitutes a wider prior knowledge model and enables a comprehensive appraisal of subsurface stratigraphy with quantified uncertainty. As pointed out in Section 2.1, a training image database is valuable as it can serve as effective supplements to overcome the intrinsic data sparsity challenge associated with subsurface stratigraphy.

As shown in Figure 2, potential training images can be obtained from one of the following four sources (Shi and Wang 2023): (a) Geological cross-sections can be borrowed from nearby sites or past stages of the current project. Those developed cross-sections reflect the required level of accuracy for practical engineering design. Figure 2a shows a geological cross-section collected from a recent reclamation project, and the section details key depositional relationships between different soil layers; (b) Conceptual geological models developed from regional geological maps or experienced geologists can also serve as potential training images. Figure 2b shows a conceptual weathering profile for decomposed granite in tropical and subtropical areas, such as Hong Kong and Singapore. There is a general trend for the level of decomposition to reduce with the increase in depth; (c) Process-based numerical models or laboratory tests can be carried out to generate or simulate realistic geological patterns that pertain to a particular geological setting. Figure 2c shows a simulated alluvial fan constructed from 600,000 years of sedimentary process using a forward geological modelling method; (d) training images can also be synthesized using generative models, such as generative adversarial networks (GANs). Generative models have been widely used in computer science to augment training image database by redistribution of image patterns in available training data to synthesize diverse image features. Recently, Shi and Wang (2024) developed a GAN-based model to generate multiple plausible image samples conditioning on a single training image sample. From the design point of view, geological cross-sections obtained from previous projects are considered credible as they were directly interpreted from site-specific measurements and have been adopted for practical engineering design and analysis. Shi and Wang (2023) have also developed a training image database for weathered granite and tuff slopes in Hong Kong by compiling slope stratigraphy from well-documented slope projects. The database is categorized based on geological origins, locations, and application scenarios. Once domain-specific training images are established, it is possible to develop 2D geological cross-sections and 3D geological domains.

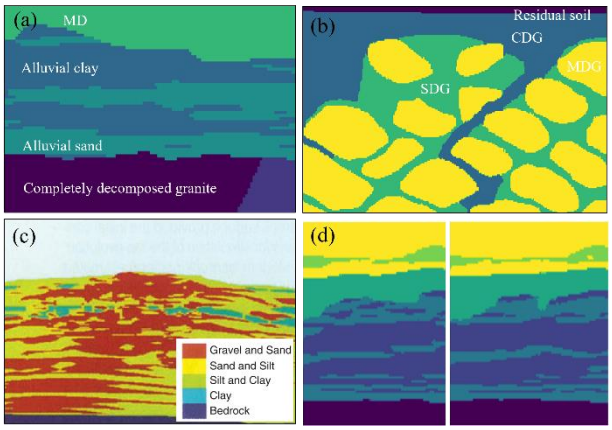


Figure 2. Potential sources of training images to construct domain-specific database.

2.3. Development of two-dimensional geological cross-section using IC-XGBoost2D

The likelihood term in Eq. (1) can be estimated using stochastic simulation tools. Mathematically speaking, the estimation of the probability term requires establishing the mapping relationship $f(\boldsymbol{x};\theta)$ between spatial coordinates $\boldsymbol{x}$ with corresponding soil categories within the target geological cross-section X.

$$X = f(\boldsymbol{x};\theta) + \varepsilon, \boldsymbol{x} \in \mathbb{R}^{\boldsymbol{2}}, \boldsymbol{s.t.} f(\boldsymbol{x_M};\theta) = y_M \qquad (2)$$

where θ denotes parameters to be estimated, and ε is a random noise; $M = (\boldsymbol{x_M}, y_M)$ represents site-specific measurements. Numerical methods are available in literature to estimate the mapping relationship, and this study adopts a single image-based stochastic simulation method called IC-XGBoost2D, which was developed by Shi and Wang (2021b). The method is capable of developing 2D geological cross-sections from sparse site-specific data and a single training image reflecting prior geological knowledge. Fig. 3 shows the basic architecture of the IC-XGBoost algorithm. The method consists of two parts, namely, training and prediction. The training part aims to extract 2D stratigraphic patterns from the single training image, and the prediction part targets to develop a 2D geological cross-section conditioning on sparse site-specific data.

The training begins with the determination of feature extractor or a 3 × 3 grid template with three distant columns. The template is then transferred to scan the single training image to exhaust the potential stratigraphic statistics via convolution with a Laplace filter, resulting in a feature map. Grid templates with different spacings are adopted to construct multiple feature maps corresponding to multi-scale stratigraphic patterns. As only cells with non-zero convolved values reflect important stratigraphic interfaces, those cells are extracted for further processing via a series of operations, i.e., non-zero pooling and dropout before feeding into a

XGBoost algorithm for multi-class classification. The pre-trained model is then adopted for forward spatial predictions in a multi-scale manner. The spatial prediction of the 1st grid involves the prediction of soil types in between two adjacent boreholes following a random simulation sequence. Soil types predicted in the previous grid scale are treated as known site-specific data for the next round of spatial prediction. The whole process repeats until all the unknown cells are predicted, thereby completing a geological cross-section.

It is also worthwhile to point out that the IC-XGBoost algorithm is a stochastic simulation method. The uncertainty mainly originates from random simulation paths for predictions associated with each grid level as well as prediction uncertainty of XGBboost. It is possible to quantify the associated stratigraphic uncertainty under the framework of Monte Carlo simulation.

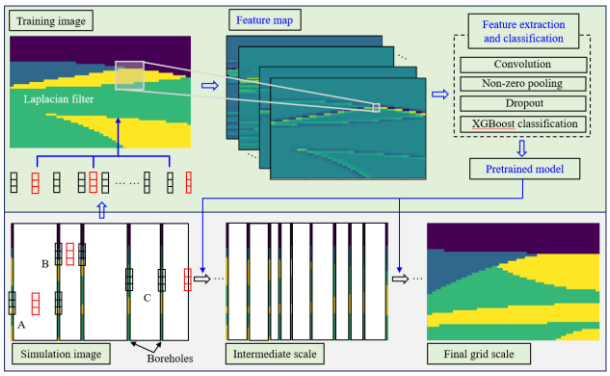
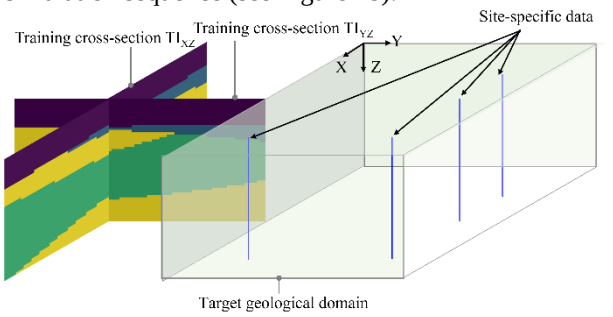


Figure 3. Basic architecture of IC-XGBoost algorithm to develop 2D geological cross-sections.

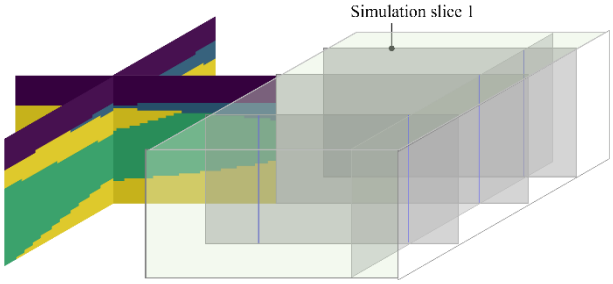
2.4. Construction of three-dimensional geological domain using IC-XGBoost3D

For most geotechnical problems, it is sufficient to develop 2D geological cross-sections for conservative engineering design and analysis. However, an accurate evaluation of engineering risks often requires a thorough appreciation of subsurface 3D stratigraphic heterogeneities. Figure 4 shows the structure of the 3D IC-XGBoost algorithm (Shi and Wang 2022), which is an extension of the 2D IC-XGBoost. The essence of the IC-XGBoost3D is sequential simulations of 2D geological cross-sections using IC-XGBoost2D before combining multiple predicted 2D cross-sections, yielding a 3D geological domain. To capture the potential spatial stratigraphic heterogeneities, a pair of perpendicular training images may be required, and each image reflects the typical stratigraphic pattens in the respective direction. It is also possible to utilize a single training image for spatial predictions of subsurface 3D geological domains based on the assumption that the geology at the site of interest is isotropic and can be learned from a representative single training image.

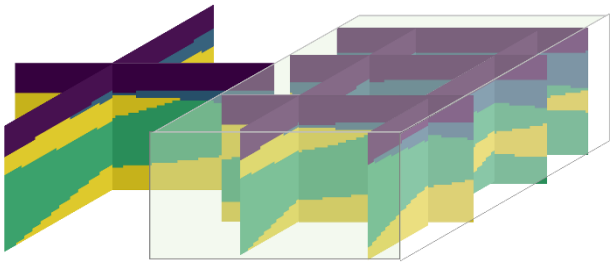
Figure 4a shows the discretization of target geological domain into a series of 2D simulation vertical slices as well as the alignment of site-specific boreholes. The next step is to determine the simulation sequence. The selection of the next simulation slice is based on the principle that the current 2D vertical slice to be predicted should have the maximum number of site-specific data. If more than two candidate vertical slices contain the same number of site-specific boreholes, a vertical slice is randomly drawn from the collection for forward spatial predictions. The delineation of soil stratigraphic patterns along a selected simulation slice is carried out using IC-XGBoost2D conditioning on site-specific data and the training image in parallel. Figure 4c shows the sequential development of 2D geological cross-sections. Once a 2D simulation slice is predicted, it is combined with original borehole logs as a new set of site-specific data. The whole process repeats until all the 2D vertical slices are predicted. Subsequently, all the predicted 2D slices are combined to yield a complete 3D geological domain. Notably, both the 2D and 3D IC-XGBoost algorithm are purely data-driven and do not require the explicit specification of distributions of any hyperparameters. The 3D IC-XGBoost algorithm inherits two major sources of uncertainty, i.e., random 2D simulation paths and classification uncertainty, along with the random simulation sequence (see Figure 4b).



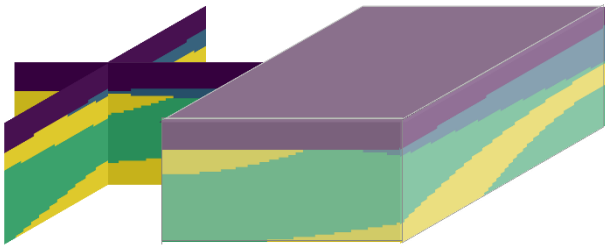
(a) Training image and site-specific data



(b) Determination of simulation sequence



(c) Sequential prediction of 2D simulation slices



(d) Reconstruction of a 3D geological realization

Figure 4. Framework for data-driven prediction of 3D geological domain

2.5. Ensemble learning of geological cross-sections

The performance of the proposed image-based stochastic methods relies on qualified training images and site-specific data. When site-specific measurements are too limited, the overall prediction is primarily governed by training images. On the other hand, when sufficient site-specific data are available, the prediction performance will be very much dependent on measurements rather than training images. Therefore, it is imperative to select training images with relatively larger occurrence probability $P(TI|M)$ for forward development of geological cross-sections. There are several methods that can be used to select appropriate training images for stochastic simulations.

For a given geological cross-section or image, the pixels around stratigraphic interfaces are more informative than those within the main bodies of soil layers, and it is possible to accurately recognize objects in an image with fragments of edges (Shotton et al. 2008). Therefore, it is reasonable to evaluate the similarity between a training image and site-specific data by comparing stratigraphic statistics on the boundaries derived from TI and site-specific data. For example, Mood (1940) and Boisvert et al. (2007) proposed to rank compatibility of a TI with site-specific data through comparison of distribution of runs (Mood 1940) and multiple-point density function (MPDF) of wells. These methods mainly focused on the 1D vertical borehole logs and does not account for the spatial stratigraphic variations across multiple boreholes in the horizontal direction. Therefore, Shi and Wang (2021c) designed an edge orientation detector to extract edge gradients in both horizontal and vertical directions for compatibility evaluation based on an edge orientation histogram (EOH).

When the training image database is structured, it is also possible to categorize and classify training images and site-specific data based on different criteria. Shi and Wang (2023) pointed out that geological origins, location proximity, and application scenarios are three possible criteria. Knowledge of the origin of soil deposits is of paramount importance to understand the nature of the deposits. Heim (1990) have summarized seven modes of origins or 16 typical types of deposits. In addition, location proximity is also an important indicator. Training images that are in close proximity to site-specific data are deemed to share similar local depositional environments (Earle 2015). Last but not least, geological cross-sections are more relevant if they are developed for the similar application scenarios (e.g., slope stability) as different application scenarios might focus on the accurate delineation of different stratigraphic patterns. For example, thin weak seams may be more crucial for slope stability analysis but can be ignored for reclamation projects.

It is also worth mentioning that each candidate training image only represents a specific geological configuration under a given geological origin and application scenario and can be viewed as a basis feature for the subsurface system (Scheidt et al. 2016). The combination of multiple training images can be taken as an "orthogonal decomposition" of the subsurface system and renders a comprehensive appraisal of subsurface geological variations. Therefore, it is beneficial to adopt multiple training images for subsurface stratigraphy, which essentially aligns with the essence of ensemble learning. Ensemble learning is a common technique to improve prediction performance by making use of multiple single models (i.e., prior geological models). By ranking the compatibility of candidate training images $\{TI^{(i)}|\, i = 1, 2, \ldots, n_t\}$ with site-specific data M, it is possible to evaluate the occurrence probability $P(TI|M)$ and estimate the posterior probability $P(X|M)$.

2.6. Uncertainty quantification

For the 2D and 3D IC-XGBoost algorithm, it is possible to generate multiple plausible geological realizations or domains by changing the random seed. The most probable prediction (*MPP*) can be derived by assigning each spatial point with the soil category of the highest occurrence frequency. *MPP* is taken as the final result of stochastic simulations. The number of stochastic realizations, N_r, is determined when the percentage change in *MPP* does not vary significantly with every k = 10 additional realizations. The threshold for the percentage change is taken to be 0.1% by default. As a rule of thumb, 100 realizations are often considered sufficient to yield a stable prediction result. For ensemble learning, the final *MPP* at each spatial point can be derived from all the generated geological realizations that are conditioned on n_s selected training images $\{TI^{(i)}|\, i = 1, 2, \ldots, n_s\}$ as follows:

$$MPP = \frac{1}{m_1+m_2+\cdots+m_{n_s}} mode\{\{Z_1^{TI^{(1)}}, \ldots, Z_{m_1}^{TI^{(1)}}\} \cup \ldots \cup \{Z_1^{TI^{(i)}}, \ldots, Z_{m_i}^{TI^{(i)}}\} \cup \ldots \cup \{Z_1^{TI^{(n_s)}}, \ldots, Z_{m_{n_s}}^{TI^{(n_s)}}\}\} \quad (3)$$

where m_i denotes the total number of stochastic realizations conditioned on the i-th training image (i.e., $TI^{(i)}$). The value of m_i can be taken to be proportional to the occurrence probability $P(TI|M)$ in Eq. (1).

It is also worthwhile to quantify the deviation of multiple geological realizations from the most probable prediction. More specifically, the stratigraphic uncertainty associated with *MPP* can be quantified using the theory of information entropy, and the entropy H at each spatial coordinate of a geological domain can be estimated as follows:

$$H = -\sum_{i=1}^{N_c} p_i \cdot log p_i \quad (4)$$

where N_c denotes the number of soil types at the site of interest; p_i is the occurrence probability of the i-th soil type among N_r stochastic realizations. In the following illustrative example, where the ground truth geological model Z_{gt} is available, it is straightforward to measure the prediction accuracy *Acc* by comparing Z_{gt} with *MPP*:

$$Acc = \frac{\sum_j I(Z_{gt}^j = MPP^j)}{N_p} \quad (5)$$

where N_p denotes the total number of discretized points in a geological cross-section or geological domain; I is an indicator function that has a value of one when the condition within the parenthesis is true, or zero otherwise;

Z_{gt}^{j} and MPP^{j} represent the ground truth and *MPP* soil type at the *j*-th spatial point within a 2D geological cross-section or a 3D geological domain.

In practice, the ground truth geological model is often unavailable. The accuracy measure in Eq. (5) only applies to simulated examples and is used for validation purposes. Alternatively, it may be possible to calculate the prediction accuracy using leave-one-out cross-section (LOOCV), which is a commonly used strategy in statistics to measure the prediction performance of a statistical model. In this study, each one of available borehole logs N_B may be iteratively removed from the training dataset and reserved for validation, and the final prediction performance is taken as the average of N_B predictions:

$$Acc^{CV} = \frac{1}{N_B}\sum_{k=1}^{N_B} Acc^{k} \quad (6)$$

where Acc^{k} represents the prediction accuracy when the *k*-th borehole is held out for validation.

3. Illustrative example

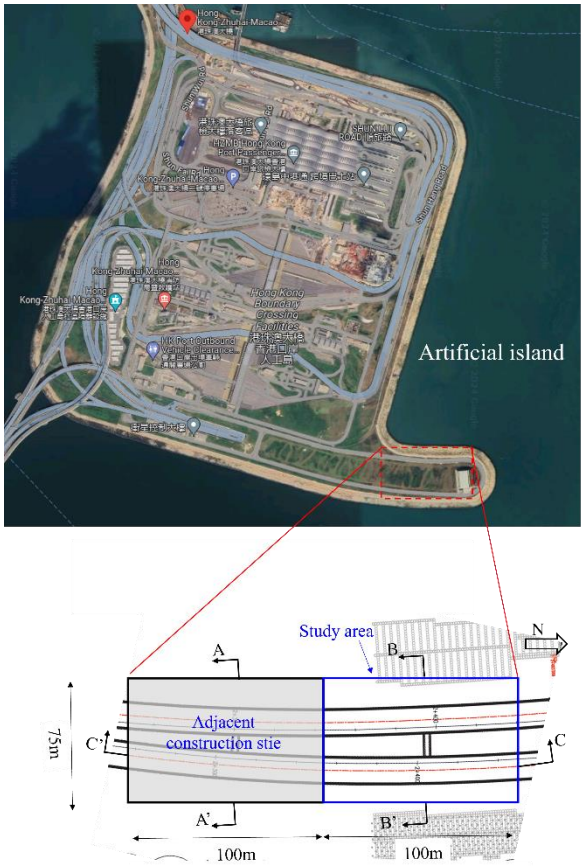


Figure 5. Location plan of the reclamation site in Hong Kong

In this study, site-specific measurements and geological cross-sections collected from a recent reclamation project in Hong Kong are used to demonstrate the performance of the proposed machine learning paradigm. Fig. 5 shows the location plan of the site of interest. The artificial island lies to the east of the existing Hong Kong International Airport and was reclaimed from the sea. The tunnelling site locates in the northeast corner of the artificial island. A comprehensive site investigation campaign was carried out to delineate subsurface geology. In total, more than 100 in-situ measurements, i.e., Cone Penetration Tests (CPTs) and boreholes, were conducted at different construction stages to decipher subsurface stratigraphic distributions and monitor the consolidation process of the fine-grained materials. The spatial distribution of those interbedded fine-grained materials has a significant influence on the long-term ground settlement as well as the serviceability of future superstructures. The study site has a plan dimension of 100m (long) × 75 m (width). Figure 6 shows a perspective view of 37 line measurements (i.e., CPTs and boreholes) at the site of interest. From the revealed geology, the site mainly comprises six soil types, namely, Fill, Disturbed Marine Deposit (DMD), Alluvial Sand (Alls), Alluvial Clay (Allc), and Completely Decomposed Granite (CDG) or better.

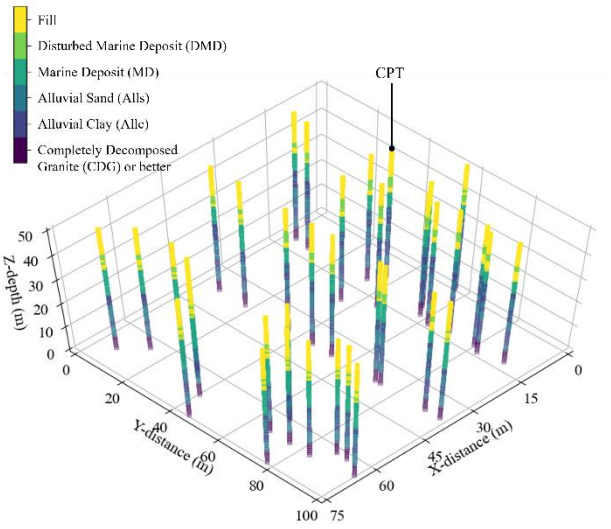
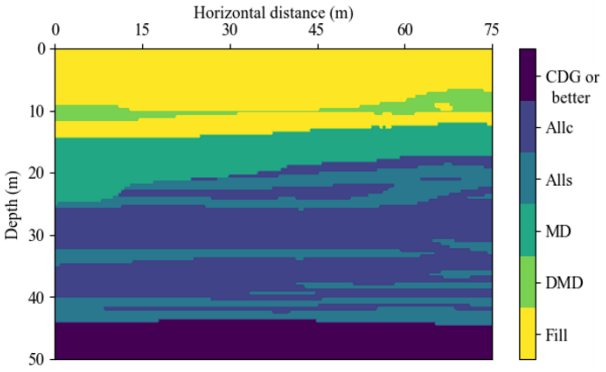


Figure 6. Site-specific measurements within the study area

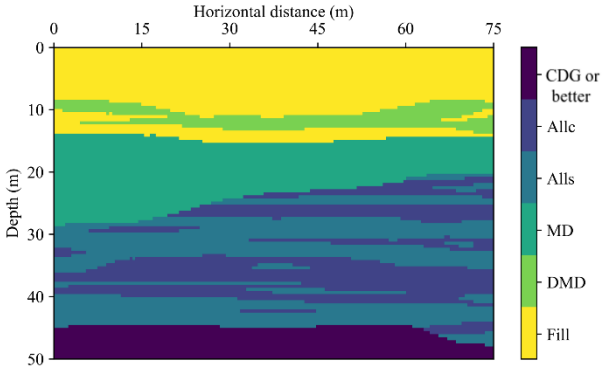
In practice, it is typical to develop 2D geological cross-sections for engineering design and analysis. The interpretation of subsurface stratigraphic distribution from sparse measurements has still largely relied on personal experience and judgement of engineers. To facilitate subsequent tunnelling design, several crucial 2D cross-sections were developed. Figure 7 shows three selected geological cross-sections along A-A', B-B', and C-C' in Figure 5. Note that the stratigraphic boundary of Fill, DMD, and MD are predominantly horizontal, and it is relatively straightforward to determine their boundaries using conventional line interpolation practice or parametric models. The key challenge lies in the delineation of stratigraphic relationships of interbedded Alls and Allc.

Two scenarios are specifically considered in this study. The first scenario (i.e., case A) involves the prediction of a 3D geological domain based on the site-specific data in Figure 6 and training images (i.e., Section A-A' and the left half of Section C-C') collected from the adjacent construction site in Figure. 5. In addition, another scenario (i.e., case B) is dedicated to investigating the ensemble learning performance of the proposed method. More specifically, Section A-A' is taken as the training image, and six line measurements

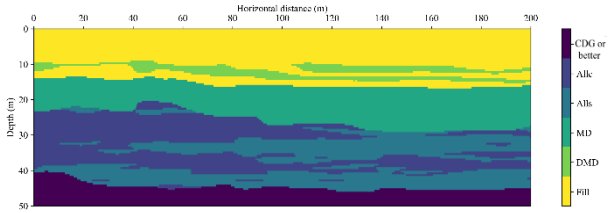
are extracted from the longitudinal cross-section C-C' as site-specific data. The aim is to predict the longitudinal geological cross-section based on a small single training image and sparse site-specific data.



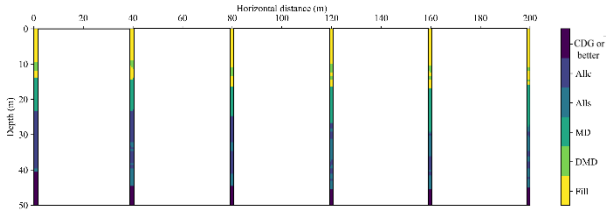
(a) Transverse geological cross-section A-A'



(b) Transverse geological cross-section B-B'



(c) Longitudinal geological cross-section C-C'



(d) Site-specific data extracted from section C-C'

Figure 7. Geological cross-sections developed by experienced geologists.

4. Results from the proposed method

4.1. Case A: data-driven prediction of 3D geological domain

Following the simulation procedures in Section 2.4, 3D stochastic simulations were carried out. The only input required is site-specific measurements and a pair of training images reflecting the potential spatial stratigraphic anisotropy at the site of interest. The target geological domain has a dimension of X = 100, Y = 100, and Z = 50. The time required to generate a 3D geological domain was about 40min using a laptop with Intel Core i7-4790 CPU @ 3.6 GHz and 8.00 GB RAM. In total, 100 geological realizations were generated. Figure 8 shows the most probable 3D geological domain derived from 100 realizations. Note that the stratigraphic connectivity and sequence can reasonably be captured. The stratigraphic boundaries of Fill, DMD, and MD are relatively simple and smooth as they are predominantly horizontal. In comparison, the stratigraphic interfaces of interbedded Alls/Allc are more complicated.

Figure 9 shows the stratigraphic uncertainty corresponding to the most probable geological domain. Areas with high entropy values mainly cluster around the predicted soil layer boundaries, indicating high stratigraphic uncertainty. The entropy values diminish to zero within the main body of each soil layer. Bands with large entropy values mainly concentrate around Alls/Allc. In other words, more site-specific data are required to have an in-depth understanding of stratigraphic connectivity between Alls and Allc. It is also worthwhile to point out that land reclamation is a man-made process, multiple rounds of site investigation campaigns are normally required in order to accurately monitor the consolidation status for construction planning. The method proposed in this study provides a data-driven tool to automatically build and update 3D geological domains from sparse data and prior training images.

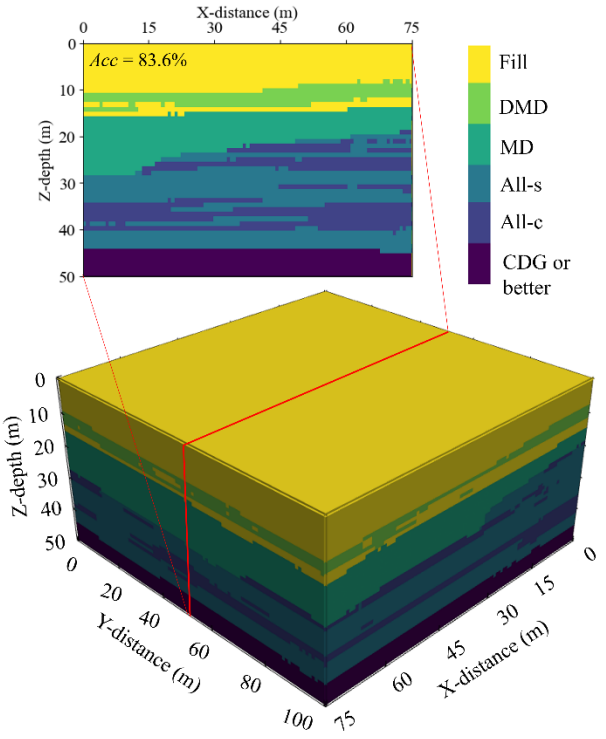


Figure 8. Predicted most probable geological domain.

It is also worthwhile to compare the predictions of 2D geological cross-sections with those interpreted by practicing engineers. Figure 8 also shows the 2D geological cross-section predicted by the proposed IC-XGBoost algorithm at Y = 50m. Similarly, it is straightforward to determine the stratigraphic boundaries of Fill, MD, and CDG or better. Despite the large stratigraphic variations of Alls/Allc, the overall

prediction accuracy is about 83.6%. This is encouraging as the ground truth 2D geological cross-sections is normally interpreted by projecting adjacent line measurements onto the target 2D cross-section for stratigraphic modelling, requiring a significant number of measurements. Figure 9 also shows the entropy colormap corresponding to the 2D *MPP* in Figure 8. The stratigraphic uncertainty rapidly reduces to zero at locations close to the borehole (i.e., Y = 50m).

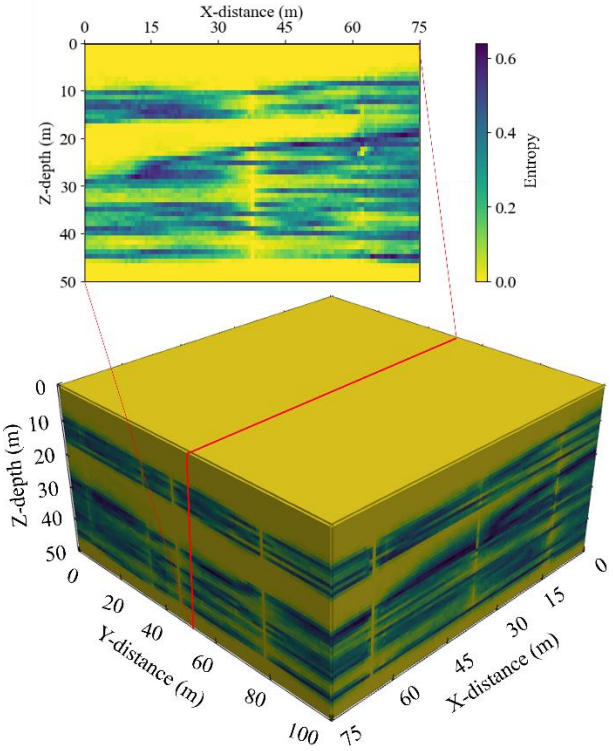


Figure 9. Stratigraphic uncertainty associated with the most probable geological domain.

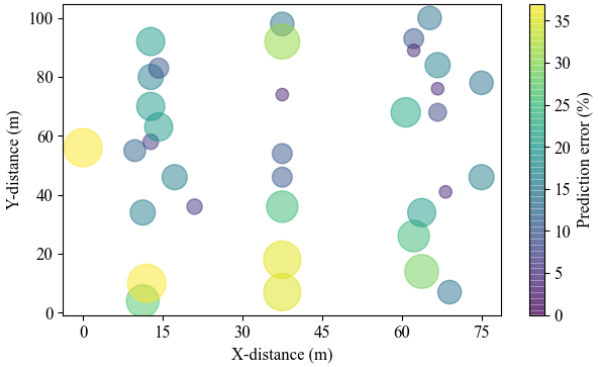


Figure 10. Scatter plot of prediction errors.

LOOCV was further performed to demonstrate the prediction performance of the proposed algorithm. Each one of the 37 site-specific measurements was iteratively removed from the training dataset and served as validation dataset. In total, 37 rounds of LOOCV were carried out. Figure 10 shows the scatter plot of prediction errors. The center of each bubble denotes the location of the reserved measurement. Bubbles with a larger diameter denote a higher prediction error. The prediction error ranges between 0% and 35% with the average of 23%. The points with largest prediction errors mainly locate close to the boundary. This is expected as the interpretation of stratigraphic profiles for points close to boundaries normally involves extrapolation, which is prone to prediction errors.

4.2. Case B: ensemble learning of 2D geological cross-section

The two perpendicular training images in case A were directly taken from a nearby site with the similar geological settings. For case B, the geological cross-section along A-A' in Figure 5 was taken as the training image, and six boreholes extracted from the long longitudinal geological cross-section C-C' were taken as site-specific data. Although the execution of the proposed machine learning algorithm does not require the training image to have the same size as that of the target cross-section, the training cross-section has a total horizontal length of 75m, which is much shorter than that of the longitudinal cross-section in Figure 7d. As a result, the short-range stratigraphic patterns reflected in Figure 7b may not be representative of those in the target cross-section, which is primarily governed by long-range stratigraphic connectivity. However, it is possible to construct a domain-specific training image database based on a single training image using generative adversarial networks (GANs). The details of GAN for generating multiple plausible training images can refer to Lyu et al. (2024) and Shi and Wang (2024). In total, 50 elongated random image samples were synthesized. Each of the 50 training image samples was iteratively combined with site-specific measurements for development of subsurface geological cross-sections. The compatibility of training images with site-specific data can be ranked using the edge orientation histogram (EOH) proposed by Shi and Wang (2023) or computed total entropy values using Eq. (4). Figure 11 shows two elongated random image samples generated by GAN. As a first approximation, top 5 image samples that are most compatible with site-specific data are selected for ensemble learning. The conditional probability $P(TI|M)$ can be taken to be proportional to the EOH or the total entropy values. Results indicate that there is minimal difference in the conditional probability for this particular example and may be considered to share approximately the same weight.

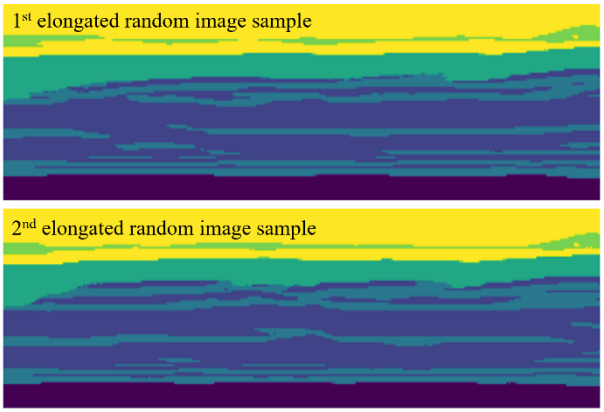


Figure 11. Random image samples generated by GAN.

Figure 12a shows the developed 2D geological cross-section conditioning on the original single training image

in Figure 7a and site-specific data. Although the training image mainly reflects short-range stratigraphic patterns, the predicted stratigraphic boundaries reasonably replicate the key spatial connectivity in Figure 7c with a prediction accuracy of 84.3%. The key difference between the predicted cross-section and that interpreted by engineers lies within the boundary of DMD. The predicted stratigraphic boundaries of DMD in Figure 11a is discontinuous and lack horizonal connectivity. In comparison, Figure 12c shows the ensemble learning results following Eq. (3). The prediction accuracy improves slightly from 84.3% to 86.7%. Although the increment is not so significant, the stratigraphic connectivity has significantly strengthened. For example, DMD is no longer isolated and extents continuously in the horizontal direction. In addition, the stratigraphic boundaries become smoother with much less patchy patterns compared to that in Figure 12a. This can be explained by the fact that the newly generated random image samples have enhanced horizontal stratigraphic connectivity compared to the original training image, thereby leading to enhanced prediction performance and informative stratigraphic uncertainty (e.g., concentrated entropy bands in Figure 12d).

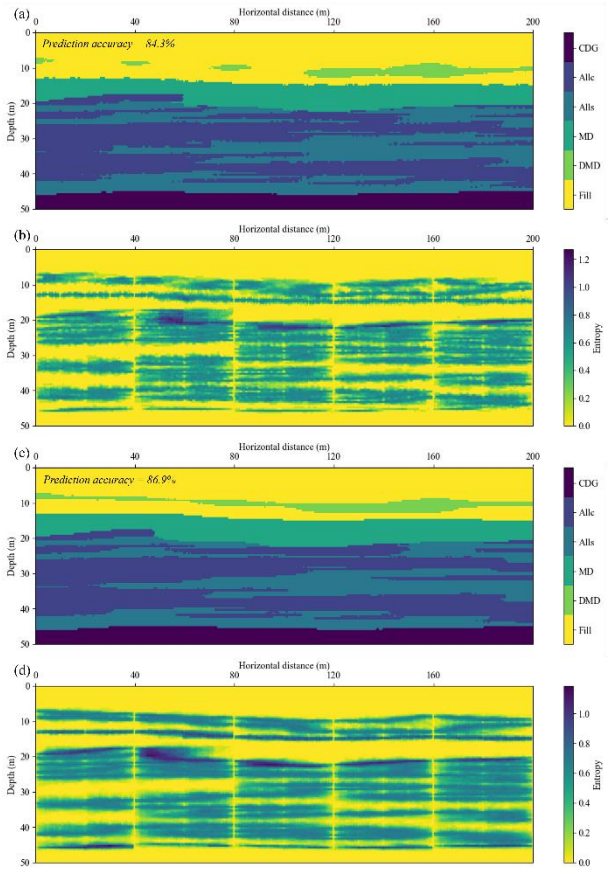


Figure 12. Ensemble learning results: (a) Most probable prediction associated with a single training image; (b) Stratigraphic uncertainty associated with the prediction from a single training image; (c) Most probable prediction derived from multiple training images; (d) Stratigraphic uncertainty associated with the prediction from multiple training images.

5. Summary and conclusion

Delineation of subsurface stratigraphic distributions is a key task of geotechnical site characterization. Traditional stratigraphic modelling methods either rely on oversimplified linear interpolation practice or complicated parametric models for subsurface stratigraphy. Both strategies may encounter significant challenges when only limited site-specific data are available. On the other hand, valuable prior geological knowledge has been implicitly embedded in the traditional stochastic simulation methods but has not been explicitly quantified and leveraged. In this study, a machine learning paradigm is proposed to automatically build and update subsurface stratigraphy from sparse site-specific data. The framework leverages valuable prior geological knowledge and quantitatively represent it as training images. Subsequently, image-based machine learning algorithms modified from conventional CNN structures are developed to predict subsurface 2D geological cross-sections and 3D geological domains from sparse data and one or two training images. As a single training image only reflects a specific scenario or configuration for a particular geological setting, it is possible to construct and build a domain-specific training image database to exhaust potential stratigraphic connectivity. The collected or generated training images can be adaptively ranked and selected for ensemble learning of subsurface stratigraphic distributions. The performance of the proposed machine learning framework is demonstrated via real examples collected from a recent reclamation project in Hong Kong. Results indicate that the proposed image-based stochastic methods can not only accurately predict 2D and 3D subsurface stratigraphic distributions from limited site-specific data but also allow the quantitative evaluation of associated stratigraphic uncertainty. It is also found that ensemble learning can help enhance prediction performance of the proposed stochastic simulation methods.

Acknowledgements

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 financial support is gratefully acknowledged.

References

Boisvert, J. B., Pyrcz, M. J. and Deutsch, C. V. (2007). "Multiple-point statistics for training image selection." Natural Resources Research, 16, pp.313-321. https://10.1007/s11053-008-9058-9.

Chen, W., Ding, J., Wang, T., Connolly, D. P., and Wan, X. (2023). "Soil property recovery from incomplete in-situ geotechnical test data using a hybrid deep generative framework." Engineering Geology, 326, 107332. https://doi.org/10.1016/j.enggeo.2023.107332.

Deng, Z. P., Jiang, S. H., Niu, J. T., Pan, M., and Liu, L. L. (2020). "Stratigraphic uncertainty characterization using generalized coupled Markov chain." Bulletin of

Engineering Geology and the Environment, 79, 5061-5078. https://doi.org/10.1007/s10064-020-01883-y.

Earle, S. (2015). *Physical Geology*. Bccampus.

Elfeki, A. M. M., and Dekking, F. M. (2005). "Modelling subsurface heterogeneity by coupled markov chains: directional dependency, walther's law and entropy." Geotechnical & Geological Engineering, 23(6), 721–756. https://10.1007/s10706-004-2899-z.

Gong, W., Zhao, C., Juang, C. H., Tang, H., Wang, H., and Hu, X. (2020). "Stratigraphic uncertainty modelling with random field approach." Computers and Geotechnics, 125, 103681. https://doi.org/10.1016/j.compgeo.2020.103681.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Heim, G. E. (1990). "Knowledge of the Origin of Soil Deposits isof Primary Importance to Understanding the Nature of theDeposit." Bulletin of the Association of EngineeringGeologists27 (1): 109–112.

Laloy, E., Hérault, R., Jacques, D., and Linde, N. (2018). "Training-image based geostatistical inversion using a spatial generative adversarial neural network." Water Resources Research, 54(1), 381-406. https://doi.org/10.1002/2017WR022148.

Li, J., Cai, Y., Li, X., and Zhang, L. (2019). "Simulating realistic geological stratigraphy using direction-dependent coupled Markov chain model." Computers and Geotechnics, 115, 103147. https://10.1016/j.compgeo.2019.103147.

Lyu, B., Wang, Y. and Shi, C. (2024). "Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge." Computers and Geotechnics, 170, p.106336. https://doi.org/10.1016/j.compgeo.2024.106336.

Mariethoz, G., and Caers, J. (2014). *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons.

Mood, A.M., (1940). "The distribution theory of runs." Ann. Math. Stat. 11 (4), 367–392. https://10.1214/aoms/1177731825.

Nobre, M. M., and Sykes, J. F. (1992). "Application of Bayesian kriging to subsurface characterization." Canadian geotechnical journal, 29(4), 589-598. https://doi.org/10.1139/t92-066.

Mosser, L., Dubrule, O., and Blunt, M. J. (2017). "Reconstruction of three-dimensional porous media using generative adversarial neural networks." Physical Review E, 96(4), 043309. https://doi.org/10.1016/j.jappgeo.2023.105042.

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, 129–140. https://doi.org/10.1016/j.enggeo.2016.04.017.

Scheidt, C., A. M. Fernandes, C. Paola, and J. Caers. (2016). "Quantifying Natural Delta Variability using a Multiple-point Geostatistics Prior Uncertainty Model." Journal of Geophysical Research: Earth Surface121 (10): 1800–1818. https://doi.org/10.1002/2016JF003922.

Shi, C., and Wang, Y. (2021a). "Nonparametric and data-driven interpolation of subsurface soil stratigraphy from limited data using multiple point statistics." Canadian Geotechnical Journal, 58(2), 261-280. https://doi.org/10.1139/cgj-2019-0843.

Shi, C. and Wang, Y. (2021b). "Development of subsurface geological cross-section from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost." Journal of Geotechnical and Geoenvironmental Engineering, 147(9), p.04021082. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002583.

Shi, C., and Wang, Y. (2021c). "Training image selection for development of subsurface geological cross-section by conditional simulations." Engineering Geology, 295, 106415. https://doi.org/10.1016/j.enggeo.2021.106415.

Shi, C. and Wang, Y. (2022). "Data-driven construction of Three-dimensional subsurface geological models from limited Site-specific boreholes and prior geological knowledge for underground digital twin." Tunnelling and underground space technology, 126, p.104493. https://doi.org/10.1016/j.tust.2022.104493.

Shi, C., and Wang, Y. (2023). "Development of training image database for subsurface stratigraphy." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 17(1), 23-40. https://doi.org/10.1080/17499518.2023.2169942.

Shi, C., and Wang, Y. (2024). "Stochastic modelling of subsurface stratigraphy from sparse data and augmented training images." the 18th World Conference of the Associated Research Centres for the Urban Underground Space, Singapore, 01st – 04th November 2023.

Shotton, J., Blake, A. and Cipolla, R. (2008). "Multiscale categorical object recognition using contour fragments." IEEE transactions on pattern analysis and machine intelligence, 30(7), 1270-1281. https://10.1109/TPAMI.2007.70772.

Tang, M., Liu, Y., and Durlofsky, L. J. (2021). "Deep-learning-based surrogate flow modeling and geological parameterization for data assimilation in 3D subsurface flow." Computer Methods in Applied Mechanics and Engineering, 376, 113636. https://doi.org/10.1016/j.cma.2020.113636.

Wang, Y., Lyu, B., Shi, C. and Hu, Y. (2024). "Non-parametric simulation of random field samples from incomplete measurements using generative adversarial networks." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 18(1), pp.60-84. https://doi.org/10.1080/17499518.2023.2222383.

Yan, W., Shen, P., Zhou, W.-H., and Ma, G. (2023). "A rigorous random field-based framework for 3D stratigraphic uncertainty modelling." Engineering Geology, 323, 107235. https://doi.org/10.1016/j.enggeo.2023.107235.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., and Ding, X. (2021). "Application of deep learning algorithms in geotechnical engineering: a short critical review." Artificial Intelligence Review, 1-41. https://doi.org/10.1007/s10462-021-09967-1.