The following publication Liu H, Zhang N, Yin Z (2025), "Probabilistic stratigraphic modelling from sparse boreholes based on deep learning". Geotechnique, Vol. 75 No. 11 pp. 1457–1469 is published by Emerald and is available at https://doi.org/10.1680/jgeot.24.00998.

Probabilistic stratigraphic modeling from sparse boreholes based on deep learning

Hong-Chi LIU[1], Ning ZHANG[2], Zhen-Yu YIN[3]

Abstract:

The safety and efficiency of geotechnical design and construction heavily rely on the stratigraphic model from site investigation. However, inherent intricate stratigraphic variations and sparse borehole data introduce uncertainty and pose challenges for subsurface stratigraphy modeling. This paper proposes a hybrid neural network of Pixel Bi-directional Long Short-Term Memory (Bi-LSTM) with dense conditional random field (CRF) for probabilistic stratigraphic modeling using limited site-specific boreholes. The proposed method provides a powerful and useful tool for effectively capturing complex spatial dependencies and probabilistic evaluation of subsurface stratigraphic uncertainty. Given the hierarchical and heterogeneous characteristics of layered soils, a novel soft-boundary label relaxation (Soft-BLR) technique is developed to vectorize stratigraphic variables. Within the framework of the proposed hybrid neural network, the Pixel Bi-LSTM is combined with the Monte Carlo dropout to efficiently approximate the complex stratigraphic distribution. By defining a linear combination of Gaussian kernels, the dense conditional random field is established in the predicted soil profile to further minimize uncertainty around stratigraphic boundaries. Furthermore, this model is illustrated by the synthetic case and applied to two real cases from Hong Kong. The proposed method can not only derive a more reasonable stratigraphic model but also map the spatial uncertainty.

Keywords: Sparse boreholes; Stratigraphic uncertainty; Label relaxation; Pixel Bi-LSTM; Dense CRF

Notations:

$\mathcal{G}$ is the discretized subsurface domain of interest with $\boldsymbol{N_z}$ columns and $\boldsymbol{N_x}$ rows

$p \in \boldsymbol{P}$ is the discretized cell represented by the ordinates (i, j)

$\boldsymbol{S}$ is the set of soil categories (e.g., sand, clay)

$Z(\cdot)$ is the real geological function mapping

$f(\cdot)$ is the real probability distribution of soil categories

[1]Ph.D. Student, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China. Email: hongchi.liu@connect.polyu.hk

[2]Research Assistant Professor, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China. Email: ning-cee.zhang@polyu.edu.hk (corresponding author)

[3]Professor, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China. Email: zhenyu.yin@polyu.edu.hk (corresponding author)

$d(\cdot)$ is the predicted probability distribution of soil categories

t is the traditional one-hot value

t^* is the proposed multi-hot-like value

$\odot$ is the element-wise multiplication

σ is the sigmoid function

$\boldsymbol{i}_t, \boldsymbol{o}_t, \boldsymbol{f}_t$ are state values of the input gate, output gate, and forget state at step t in LSTM cell

$\boldsymbol{c}_t, \tilde{\boldsymbol{c}}_t, \boldsymbol{h}_t$ are the memory value, candidate value, and output value at step t in LSTM cell

$w \in \boldsymbol{W}, \boldsymbol{b}$ are learnable weight and bias matrices

$\boldsymbol{\omega}$ is the potential combination of model parameters

$\boldsymbol{N}$ is the sampled spatial neighborhood

$\mathcal{B}(\rho)$ is the Bernoulli distribution with the dropout ratio ρ

λ, α are a scale factor and a positive constant value in the Soft-BLR

ψ_u, ψ_p are the individual unary and pairwise cliques at each cell

$k^{(l)}$ is the Gaussian kernel with a linear weight $\omega^{(l)}$

$\theta_\alpha, \theta_\beta, \theta_\gamma$ represent impact degrees in the dense CRF

$I(\cdot)$ is the indicator function of the Iverson bracket

Dp measures the dispersion between the predicted and the target values

1 Introduction

In geotechnical practice, stratigraphic modeling plays a pivotal role in improving the understanding of subsurface engineering characteristics. A reasonable subsurface geological profile greatly affects the reliability of geotechnical design and the safety of subsequent construction (Jaksa et al., 2005). For example, due to the underestimated existence of a weak clay seam, the Shum Wan Road landslide in HK released about 26,000 m³ of mass and caused five injuries after days of rainstorms (Chen & Lee, 2004). Based on 89 geotechnical projects in the USA, site explorations in more than 85% of cases are severely inadequate for site characterization (National Research Council, 1984). However, inherent intricate stratigraphic variations and sparse site-specific boreholes pose a challenge to the accurate delineation of geological conditions. The estimated stratigraphic profile inevitably involves uncertainties from the natural spatial variability, incomplete knowledge from sparse boreholes, measurement errors, and transformation errors (Phoon & Kulhawy, 1999a). Particularly for a small number of borehole samples, the statistical uncertainty can exhibit a greater magnitude compared to other uncertainty sources (Phoon & Kulhawy, 1999b; Wang & Zhao, 2017).

To represent the spatial correlations over sparse boreholes, many theoretical autocorrelation functions (ACFs)

have been investigated to deliver different spatial patterns (D.-Q. Li et al., 2015). After selecting an appropriate theoretical ACF, one practical challenge is the demand for intensive data to calibrate different random fields (Cami et al., 2020). A Monte Carlo (MC) simulation-based method was developed to approximate the complex stratigraphic variations through stochastic stratigraphic configurations (Gong et al., 2020). To constrain the stratigraphic uncertainty modeling, some Markov-based approaches (e.g., coupled Markov chain, Markov random field) were developed (Qi et al., 2016; Zhao et al., 2021; Wei & Wang, 2022; J.-Z. Zhang et al., 2022). However, the stationary assumption of a Markov process conflicts with the non-stationary nature of geological data (Wang & Zhao, 2017; Jamshidi Chenari et al., 2018). It is also difficult to compute the sophisticated probability transition matrix for geotechnical practitioners, particularly when the Markov process is extended to the three-dimensional real world. More recently, some researchers have been dedicated to developing data-driven methods for extracting spatial patterns automatically. For example, multiple points statistics were applied to infer the most probable stratigraphic configuration with additional information from similar construction sites (Shi & Wang, 2021b). Further, a novel IG-XGBoost model was proposed for conducting multiscale interpolation iteratively, regarding the stratigraphic configuration as a multi-classification problem (Shi & Wang, 2021a).

With outstanding capabilities in feature extraction and non-linear fitting, deep learning algorithms such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and physics-informed neural networks (PINNs) are increasingly being embraced in geotechnical disciplines (Ninić et al., 2017; Zhang et al., 2021; Cheng et al., 2023; Zhang et al., 2023). However, the direct application of these models to the geotechnical task of stratigraphic delineation faces the difficulty of few sample training, which can lead to overfitting on the sparse available data (Samui & Sitharam, 2010; Zhou et al., 2019). This issue is particularly pronounced when using conventional one-hot encoding (Zhou et al., 2019; Gao et al., 2021; Hou et al., 2022). In such cases, the well-trained model may fall into a "label bias" problem (Torralba & Efros, 2011), where it tends to assign overly confident probabilities close to 0 or 1 to categorical values at unseen locations. Therefore, an ingenious data labeling method is urgently needed to reflect the spatial variability of geomaterials before neural network training. Additionally, it is crucial to consider the hierarchical relationship between different strata, which is a distinctive characteristic of layered soils. Combining these geological features could significantly enhance the accuracy and reliability of deep learning models in geotechnical applications.

To address the above difficulties, a hybrid neural network for probabilistic stratigraphic modeling is proposed in this paper. For better characterizing subsurface stratigraphic features, the Soft-Boundary label relaxation (Soft-

BLR) technique is developed to vectorize categorical random variables as multi-hot values, instead of one-hot values. The Pixel Bi-LSTM is integrated with the MC-dropout regularization to efficiently capture pixel-level spatial patterns. After obtaining the predicted stratigraphic profile, a dense conditional random field (CRF) is established to eliminate isolated outliers around boundaries and output a more reasonable profile. In addition, the dispersion map is provided to quantitatively show how stratigraphic uncertainty varies in space, offering valuable insights for better decision-making in the reliability-focused geotechnical design.

2 Framework of Proposed Hybrid Neural Network

In this study, the continuous subsurface cross-section is discretized into a set of cells $\boldsymbol{P}$ with the same size. Each $p \in \boldsymbol{P}$ is represented by a single pair of grid coordinates (i, j), where $i \in \boldsymbol{N_z}$ is the column number and $j \in \boldsymbol{N_x}$ is the row number. Characterized by sparse spatial distributions, site investigations reveal the set of soil categories $\boldsymbol{S}$ (e.g., $\{S_k = fill, S_l = sand, \dots, S_t = clay\}$) at drilling locations. Let $Z(p) = S_k$ denote the soil category S_k at the borehole cell p. Stratigraphic delineation is aimed at inferring the soil categories for unsampled cells from limited site-specific boreholes.

2.1 Probabilistic Pixel Bi-LSTM

The real geological function $Z(\cdot)$ commonly exhibits non-stationary behavior and discrete heterogeneity in geotechnical practice (Qi et al., 2016), suggesting complex deposit histories and tectonic activities. Within the framework of the proposed approach, the probabilistic Bi-directional Long Short-Term Memory (Bi-LSTM) algorithm is developed to model a probability distribution function $f(\cdot)$ for approximating $Z(\cdot)$ using site-specific borehole data. Thus, the categorical index of $Z(p)$ is converted into the likelihood distribution of $f(p)$.

As shown in Fig. 1, the vanilla LSTM unit is composed of input, output, and forget gates, as well as a memory cell. Such architecture allows LSTM to store long-range historical information. The typical calculation process of an LSTM cell with the input data p_t at step t can be parameterized as follows:

$$
\begin{aligned}
&\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f \chi_t + b_f\right), \\
&\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \chi_t + b_i), \\
&\tilde{\boldsymbol{c}}_t = tanh(\boldsymbol{W}_{\tilde{c}} \chi_t + b_c), \qquad \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \chi_t + b_o) \\
&\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \qquad \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right)
\end{aligned}
\tag{1}
$$

Where $\odot$ is the element-wise multiplication and σ is the sigmoid function. $\chi_t = [x_t \quad h_{t-1}]^T$. $\boldsymbol{i}_t, \boldsymbol{o}_t, \boldsymbol{f}_t$ stand for the state value of the input gate, output gate, and forget state respectively. $\boldsymbol{c}_t$ represents the memory value of the

cell while $\tilde{\boldsymbol{c}}_t$ is the candidate state value, $\boldsymbol{h}_t$ is the output of the Bi-LSTM cell. $\boldsymbol{W}$ and $\boldsymbol{b}$ are the corresponding weight and bias matrices at each of the aforementioned gates. It has been indicated that the contextual constraints between two nodes in the Markov model are found to decay exponentially with distance but decay like a power law in LSTM (Lin & Tegmark, 2017). Inspiringly, LSTM provides an alternative approach to embedding scaling spatial dependencies into stratigraphic modeling.

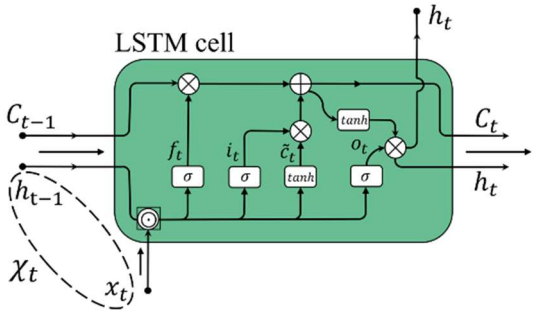


Fig. 1 Schematic illustration of the typical LSTM cell

Ideally, the soil type at a certain location should be estimated based on all borehole cells. However, the computation process can become expensive and tedious as some known points are too far away. As shown in Fig. 2, a cross-shaped neighborhood sampling system is designed for the proposed Pixel Bi-LSTM to capture spatial dependencies around the cell of interest efficiently. For example, if given the soil type at the borehole cell E, the proposed model is trained based on a more comprehensive $P(B|p_A, p_B, p_C, p_D, p_E)$ encompassing information from the surrounding neighborhood instead of the simple conditional probability expression $P(B|p_E)$. Then, the sampled cross-shaped neighborhood is split into horizontal and vertical adjacent matrices to pass through different Bi-LSTM branches simultaneously. Finally, the predicted probabilistic distribution of soil categories is obtained by concatenating the outputs of the Bi-LSTM after the dense layer. Therefore, the estimated value of d_E can be factorized as:

$$d_E = \sum_{s \in \boldsymbol{S_{all}}} \left(\underbrace{w_h P(s|p_A, p_E, p_c)}_{Horizontal\ branch} + \underbrace{w_v P(s|p_B, p_E, p_D)}_{Vertical\ bran} \right) \tag{2}$$

where $\boldsymbol{S_{all}}$ is the set of soil events appearing in boreholes. w_h, w_v are the weight matrices that need to be learned.

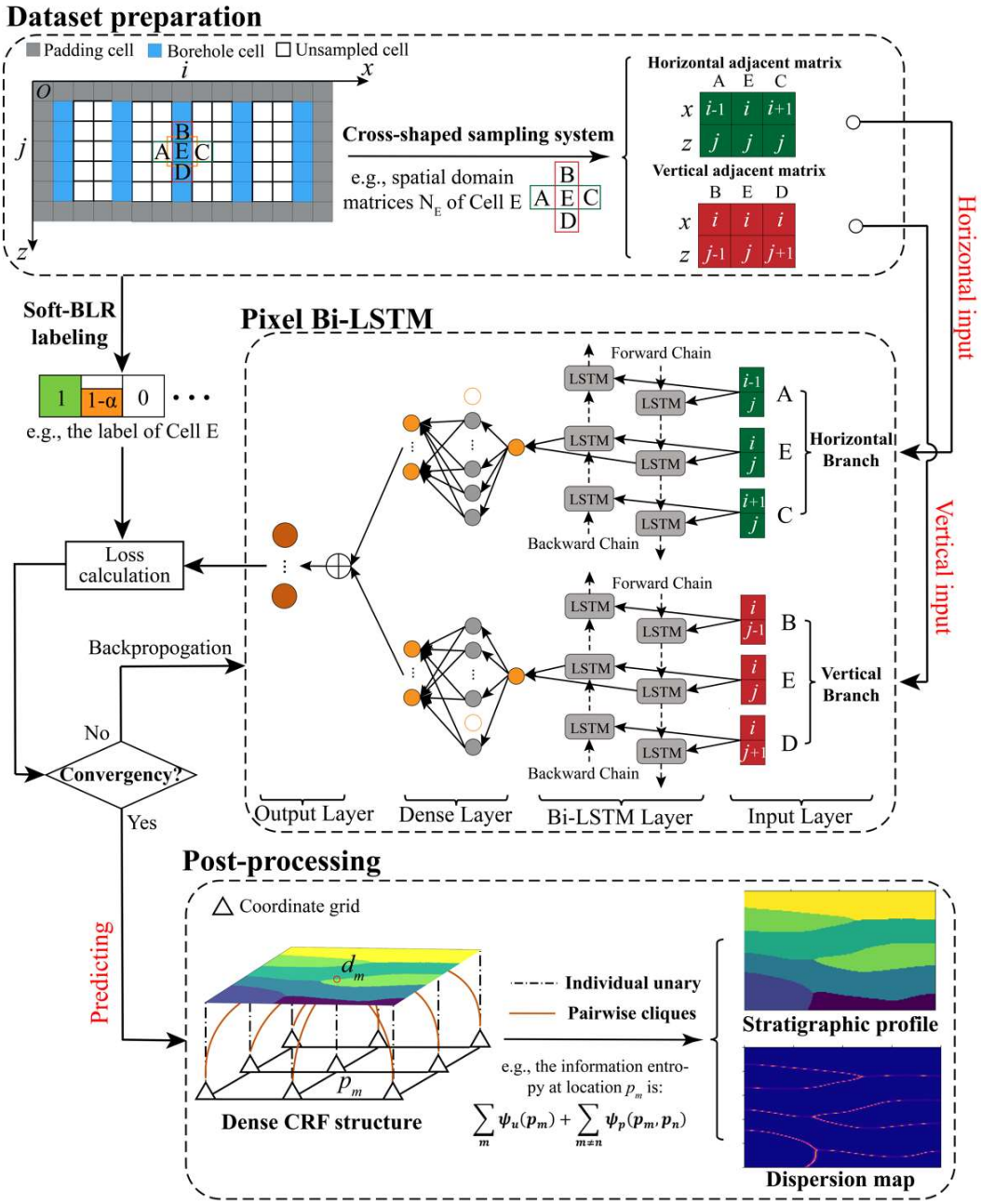


Fig. 2 Schematic illustration of the probabilistic Pixel Bi-LSTM

The direct inference of d_E with spatial domain matrices may lead to a biased result affected by the specific transformation and stochastic noise associated with the training dataset. Thus, Monte Carlo dropout (MC-dropout) where network neurons are randomly masked during the approximation phase (Gal & Ghahramani, 2016) is implemented. The original d_E derived from Eq. (2) is further converted into the simulation result conditioned on the stochastic combination of potential model parameters $\boldsymbol{\omega}$. In this manner, the predicted result $\boldsymbol{d}$ can be rewritten as:

$$\boldsymbol{d} = f(\boldsymbol{\omega}, \boldsymbol{N_E}) \tag{3}$$

where the spatial domain matrices $\boldsymbol{N_E} = \{p_A, p_B, p_C, p_D, p_E\}$ represent the cross-shaped sampling neighborhood.

Fig. 3 illustrates the application of MC-dropout in the Bi-LSTM and dense layers for variational Bayesian approximation (Damianou & Lawrence, 2013). For the MC-dropout in the Bi-LSTM layer, the Bi-LSTM cell in Eq. (1) is rewritten as:

$$\begin{aligned} \boldsymbol{f}_t &= \sigma\big(\boldsymbol{W}_f(\chi_t \odot \delta_t) + b_f\big), \\ \boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i(\chi_t \odot \delta_t) + b_i), \\ \tilde{\boldsymbol{c}}_t &= tanh(\boldsymbol{W}_{\tilde{c}}(\chi_t \odot \delta_t) + b_c), \qquad \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o(\chi_t \odot \delta_t) + b_o) \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \qquad \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \end{aligned} \tag{4}$$

where $\delta_t \sim \mathcal{B}(\rho)$ is the random mask vector with the dropout ratio of ρ.

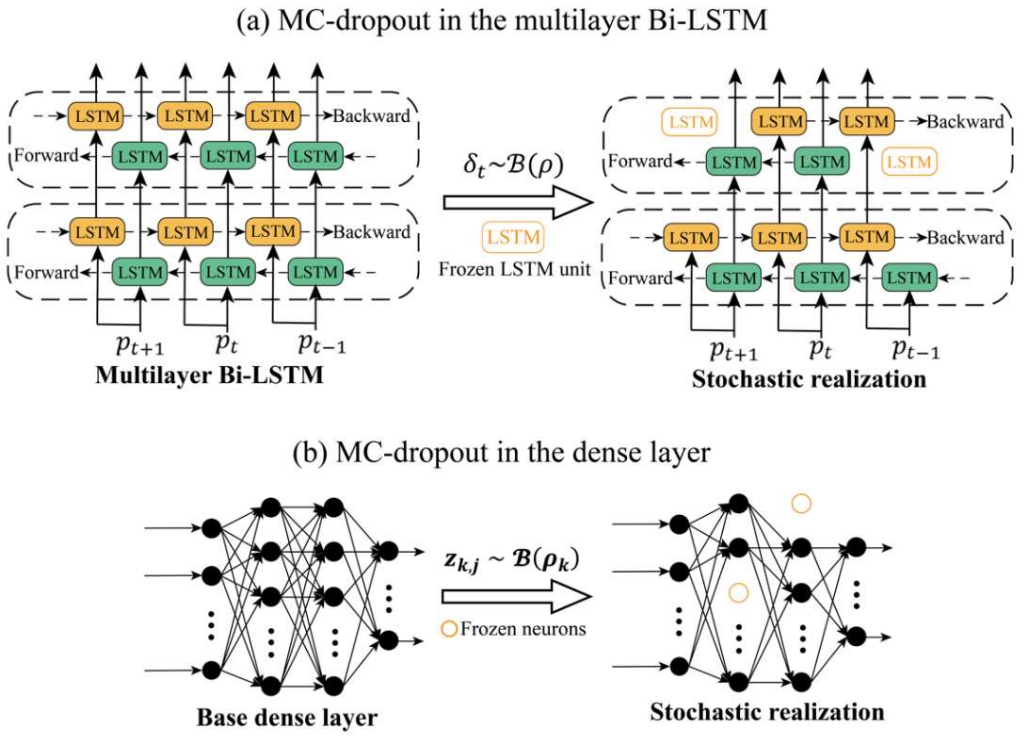


Fig. 3 Illustration of variational Bayesian approximation via MC-dropout

For the MC-dropout in the dense layer (see Fig. 3 (b)), $\boldsymbol{d}$ in Eq. (3) can be parametrized based on the Bayes' theorem:

$$p(\boldsymbol{d}|\boldsymbol{D}, \boldsymbol{\omega}) = \int p(\boldsymbol{d}|\boldsymbol{D}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{D})\, d\boldsymbol{\omega} \tag{5}$$

Where $\boldsymbol{D}$ is denoted as the outputs of the Bi-LSTM layer and $p(\boldsymbol{\omega}|\boldsymbol{D})$ is the posterior distribution of weights. It is computationally intractable to approximate the likelihood in Eq. (5) which is equivalent to deploying an ensemble of an infinite number of models. The inferential process is simplified to model the probability distribution over $p(\boldsymbol{\omega}|\boldsymbol{D})$ with a simpler distribution $q(\boldsymbol{\omega})$. $q(\boldsymbol{\omega})$ is the distribution over matrices whose columns

are randomly set to zero. For the k^{th} $(k \geq 2)$ layer being dropped out (Gal & Ghahramani, 2016), $q(\boldsymbol{\omega})$ is defined as:

$$\boldsymbol{W}_k = \boldsymbol{M}_k \cdot diag\left(\left[z_{k,j}\right]_{j=1}^{K_i}\right) \quad (6)$$
$$z_{k,j} \sim \mathcal{B}(\rho_k) \quad \forall k \in \{1, \ldots, D\}, \forall j \in \{1, \ldots, K_{k-1}\}$$

where ρ_k and matrices $\boldsymbol{M}_k$ are the given variational parameters. The binary variable $z_{k,j} = 0$ corresponds to the unit j in the layer $(k-1)^{th}$ being dropped out as an input to the k^{th} layer.

The Adam optimization with L2 regularization technique is chosen to optimally update network parameters $\boldsymbol{\omega}$ at each epoch. Regarding the target categorical value is $\boldsymbol{t}$, the cross-entropy loss can be calculated as:

$$\boldsymbol{\mathcal{L}} = -\sum_{\boldsymbol{s}}^{\boldsymbol{S_{all}}} \boldsymbol{t} log(\boldsymbol{d}) + \frac{\lambda_{L2}}{2} \|\boldsymbol{w}\|_2 \quad (7)$$

where $\|\boldsymbol{w}\|_2 = (w_1)^2 + (w_2)^2 + \cdots$ is the L2 regularization term and increasing λ_{L2} value strengthens the regularization effect over model parameters.

Before model training, the input cross-shaped adjacency matrices are normalized, and the labels are preprocessed using the Soft-BLR technique to better capture uncertainties at the soil interfaces. To rigorously evaluate the model's performance and ensure strong generalization, K-Fold cross-validation is employed, ensuring the model is tested on different subsets of the sampled data. The training phase stops either when validation loss worsens or the maximum number of epochs is reached, preventing overfitting. During the prediction phase, the model iteratively processes the cross-shaped neighborhood around each unknown cell. Note that the proposed model is adaptable to both regular and irregular borehole spacing, enhancing its practical flexibility.

2.2 Soft-Boundary Label Relaxation

Converting the raw borehole data into a format that the algorithm can recognize and learn from, data labeling is a crucial step for model training. Compared with numerical encoding, the one-hot encoding technique performs better in representing categorical features (Seger, 2018). In this way, geological random variables, e.g., Soil A and Soil B, can be defined as multi-dimensional binary vectors (the vector with only one element containing the value 1, while all others contain the value 0):

$$t_{\boldsymbol{a}} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, t_{\boldsymbol{b}} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad (8)$$

where the dimension of $t_{\boldsymbol{a}}, t_{\boldsymbol{b}}$ corresponds to the size of $\boldsymbol{S}$.

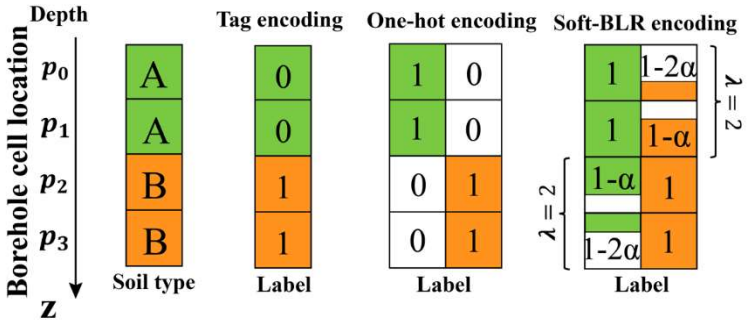


Fig. 4 Comparison between different encoding methods for labeling borehole cells around the stratigraphic boundary

Compared with the continuous stratigraphic variation expressed by boreholes, the pixel-based stratigraphic configuration can be considered a coarse representation. For example, if generating a soil profile with a vertical resolution of 0.2 m, the interlayer soil with 1.5m thickness would be discretized into cells with integer numbers of 7 or 8. When labeling these pixel cells, it surpasses the capability of Eq. (8) to store the inhomogeneous stratigraphic information of borehole data. Particularly, it has been identified that the misalignment or ambiguity of the pixel-based model mainly clusters around inferred stratum boundaries (Shi & Wang, 2023). Moreover, the conventional one-hot encoding technique cannot take into account the chronostratigraphic order of soil classes in the layered profile. Therefore, a novel label encoding technique "Soft-boundary label relaxation" (Furitsu et al., 2021) is developed in this paper to accept multiple soil classes at a borehole cell and consider a wider range of geological changes. The relaxed class label of a borehole cell contains the binary state of its neighboring borehole cell within an arbitrary distance. The comparison between different label encoding methods is illustrated in Fig. 4. To introduce the wider stratigraphic boundary, several conditions should be recognized: (1) the likelihood of being Soil A at the borehole cell p_1 must be larger than that being Soil B; (2) the difference in the likelihood of soil types must increase when the cell location is further away from the border. These can be formulated as:

$$P(A|p_1) > P(B|p_1) \tag{9}$$

$$P(A|p_0) - P(B|p_0) > P(A|p_1) - P(B|p_1) \tag{10}$$

Eq. (10) can be rewritten as:

$$P(B|p_1) - P(B|p_0) > P(A|p_1) - P(A|p_0) \tag{11}$$

Since the random variable observed at p_0 and p_1 are the same as being Soil A, we assume the difference between $P(A|p_1)$ and $P(A|p_0)$ to be 0. Then, the difference in the left side of Eq. (11) should be larger than 0 and is regarded as a positive constant α to represent the possibility of Class B decreasing linearly with the distance from the stratigraphic boundary. Hence, a wider label boundary can be constructed by defining the scale factor λ

of being another soil:

$$P(A|p_1) - P(A|p_0) = 0 \tag{12}$$

$$P(B|p_1) - P(B|p_0) = \alpha > 0 \tag{13}$$

$$Distance(p_o) \leq \lambda, Distance(p_1) \leq \lambda \tag{14}$$

Where $Distance(\cdot)$ is the function measuring the chessboard distance between the cell and the border. To ensure multi-hot-like values remain positive, the weight of being another soil is $max\{0, 1 - \lambda\alpha\}$. The new soil labels $t_a^*,\ t_b^*$ at $p_0,\ p_1$ can be denoted as:

$$t_a^* = \begin{pmatrix} 1 \\ 1-\alpha \\ \vdots \\ 0 \end{pmatrix}, t_b^* = \begin{pmatrix} 1-\alpha \\ 1 \\ \vdots \\ 0 \end{pmatrix} \tag{15}$$

Under the framework of the proposed hybrid neural network, the Soft-BLR technique is used to store anisotropic information at each borehole cell, embed the hierarchical order of soil variables, and further affect the alignment of unsampled cells around stratigraphic boundaries. Soft-BLR acts as a flexible constraint allowing for ambiguous stratigraphic boundaries that are related to the chessboard distance from the border.

2.3 Dense Conditional Random Field

However, the probabilistic Pixel Bi-LSTM could not fully consider pixel-level dependencies among inferred soil types in latent variable modeling, which might underestimate spatial dependencies. A novel conditional random field (CRF) is developed for post-processing categorical predictions and estimating a more reliable stratigraphic profile. Compared with basic random fields, CRF integrates the investigation information at the given site to constrain the probabilistic site characterization. Current CRFs are composed of unary potentials on individual cells and pairwise potentials on the local neighborhood (X. Y. Li et al., 2016; Liu et al., 2017; Xie et al., 2023). Restricted by the sparsity of boreholes, these CRFs have to focus on capturing local geological settings and might overestimate the stratigraphic variations. Additionally, the resulting adjacency CRF structures have inherent limitations in modeling long-range connections. Instead of directly modeling on sparse boreholes, we construct CRF structures based on the complete soil profile inferred by the probabilistic Pixel Bi-LSTM. Therefore, a dense CRF can be defined by incorporating prior geological knowledge at borehole locations, along with additional information from unsampled cells. Let p_m denote the coordinate vector of cell m and d_m be the corresponding multi-hot-like value. Given the complete graph $\mathcal{G}$ over the discretized cross-section domain of size $N_x \times N_z$, the Gibbs energy consisting of individual unary and pairwise cliques at each cell can be combined as (Krähenbühl & Koltun, 2011):

$$E(p) = \sum_{m} \psi_u(p_m) + \sum_{m \neq n} \psi_p(p_m, p_n) \tag{16}$$

where $m,\ n$ range from 1 to $N_x \times N_z$. The computation of $\psi_u(p_m)$ has been performed by the probabilistic Pixel Bi-LSTM that produces the multi-hot-like value d_m. For highly efficient inference of the pairwise edge potential, $\psi_p(p_m, p_n)$ is written as a linear combination of Gaussian kernels:

$$\psi_p(p_m, p_n) = \mu(p_m, p_n) \underbrace{\sum_{l=1}^{K} \omega^{(l)} k^{(l)}(f_m, f_n)}_{k(f_m, f_n)} \tag{17}$$

where f_m, f_n are extracted feature vectors. The label compatibility function μ is determined by the Potts model (Koller & Friedman, 2009), penalizing neighboring cells with different events. $k^{(l)}(f_m, f_n) = \exp\left(-\frac{1}{2}(f_m - f_n)^T \Lambda^{(l)}(f_m - f_n)\right)$ is the Gaussian kernel with a linear weight $\omega^{(l)}$. To specify, $k(f_m, f_n)$ is characterized by probabilistic stratigraphic events d_m, d_n and location information p_m, p_n:

$$k(f_m, f_n) = \omega^{(1)} \underbrace{\exp\left(-\frac{|p_m - p_n|^2}{2\theta_\alpha^2} - \frac{|d_m - d_n|^2}{2\theta_\beta^2}\right)}_{appearance\ kernel} + \omega^{(2)} \underbrace{\exp\left(-\frac{|p_m - p_n|^2}{2\theta_\gamma^2}\right)}_{smoothness\ kernel} \tag{18}$$

Reflecting desirable geological settings, the appearance kernel is the empirical term to regulate the probability transition over soil categories marginally. $\theta_\alpha, \theta_\beta, \theta_\gamma$ represent respective impact degrees. The smoothness kernel aims to modify misalignments and remove isolated outliers. To evaluate the performance of the hybrid neural network, we use the global accuracy (GA) for the known subsurface cross-section (Shi & Wang, 2021a) and the borehole accuracy (BA) for sparse boreholes only. It should be noted that the selected borehole for calculating BA isn't included in the model training. Thus, the discrepancy between the predicted most probable stratigraphic events and the ground truth can be measured as follows:

$$Global\ Accuracy\ (GA) = \frac{\sum_{m=1}^{N_x \times N_z} I(s_m = \hat{s}_m)}{N_x \times N_z} \tag{19}$$

$$Borehole\ Accuracy\ (BA) = \frac{\sum_{m=1}^{N_z} I(s_m = \hat{s}_m)}{N_z} \tag{20}$$

where s_m represents the most likely stratigraphic event among d_m, and $\hat{y}_m$ is denoted as the true categorical value in cell m. $I(\cdot)$ is the indicator function of the Iverson bracket countering how many predictions are the same as the ground truth.

Since the information entropy $\boldsymbol{d}$ over soil categories is obtained through the probabilistic Pixel Bi-LSTM, the uncertainty of the most probable stratigraphic event at each cell $\boldsymbol{p}$ can be quantified. By adopting the Soft-BLR technique, the predicted information entropy reflects both the incomplete knowledge from sparse boreholes

and the transformation uncertainty inherent in geological modeling. Visualizing $\boldsymbol{d}$ provides a clear overview of the dispersion $Dp(p_m)$, capturing the uncertainty associated with these factors. The resulting dispersion map is depicted by calculating each $Dp(p_m)$ as follows:

$$D_p(p_m) = 1 - d_m^s, \qquad 0 \le D_p(p_m) \le 1 \tag{21}$$

where d_m^s represents the probability of the most probable soil category s among the predicted multi-hot-like value d_m at cell $D_p(p_m)$.

3 Illustration

In this section, a parametric analysis is conducted based on a synthetic multilayered subsurface profile to evaluate the impact of three key aspects on the performance of the proposed method. The synthetic case includes six soil types, labeled 'Soil 1' to 'Soil 6', and is manually generated using various sinusoidal and cosinusoidal boundary curves (see Fig. 5 (a)). These curves divide the subsurface cross-section into a grid of 200 × 125 pixels in the horizontal and vertical directions, respectively. To address the challenge of sparse survey data in practical engineering, five evenly distributed virtual boreholes (up to 2.5% of the total), labeled 'BH1' to 'BH5' (see Fig. 5 (b)), are used as known geological conditions to infer the remaining unsampled portions. The soil class distribution of the sampled dataset at each borehole location is presented in Fig. 5 (c). Note that the proposed model is purely data-driven and infers the most probable subsurface profile based solely on site-specific data without requiring additional data from other projects. Despite the robust mathematical foundation, the proposed model is fully programmed in Python to facilitate automatic stratigraphic delineation, making it accessible and user-friendly for geotechnical engineers. As depicted in Fig. 2, the modeling process is streamlined to allow for application in practical scenarios, eliminating the need for extensive computational expertise. Based on the grid search method (see Appendix), the recommended hyperparameter configuration for the neural network is provided in Table 1. It can be observed that the model's performance initially improves and then stabilizes as the model size, batch size, and learning rate are increased. As shown in Fig. 5 (d), The decreasing trends in both training and validation loss, converging to similar values over epochs (early stop) indicate that the model has thoroughly captured the underlying patterns and generalizes well to unseen data. Moreover, the total training and inference time remains under 6 seconds, suggesting that this model is an efficient tool for site characterization in engineering practice.

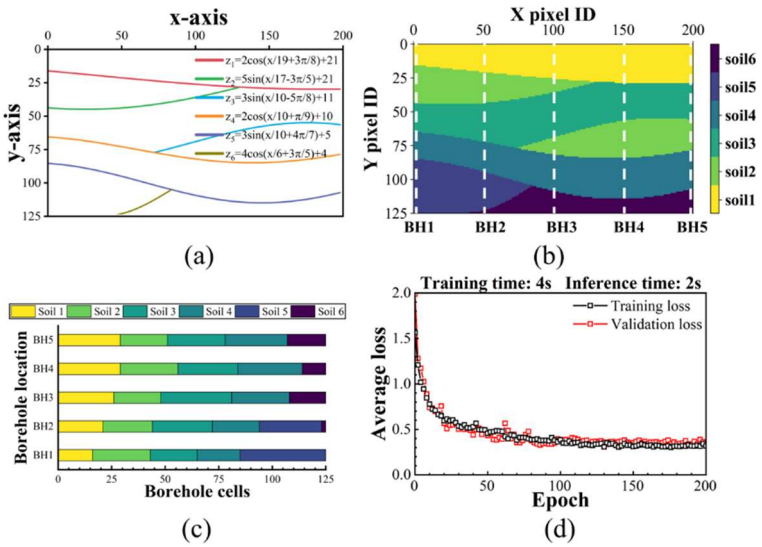


Fig. 5 (a) Synthetic soil interfaces; (b) Colormap of the synthetic cross-section with five virtual boreholes; (c) Soil distributions of known boreholes; (d) Example loss curve plot of the proposed neural network

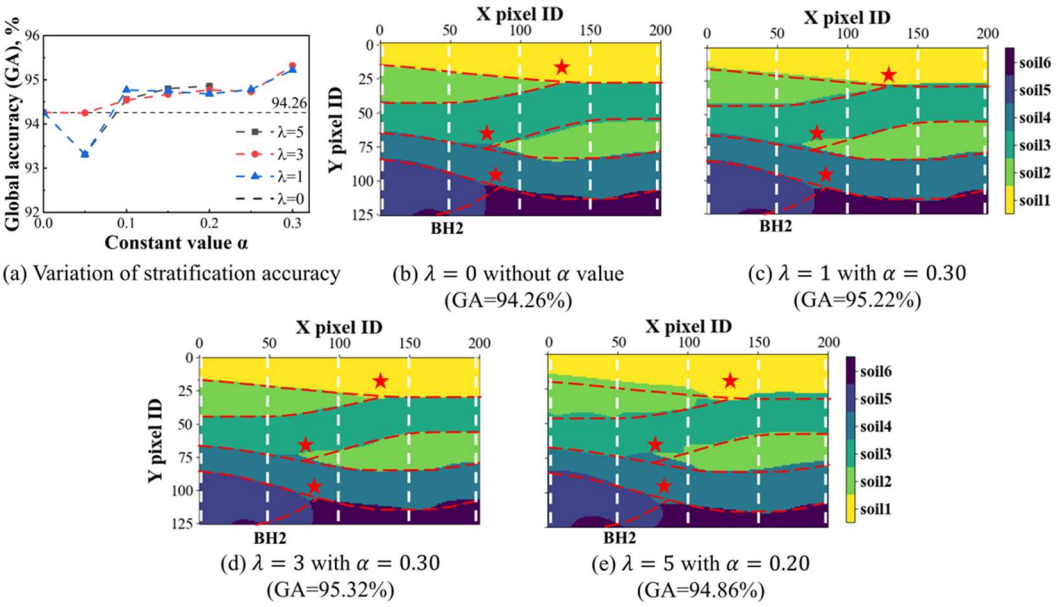
Table 1 A summary of hyperparameters used in the illustrative example

Hyperparameter	Value
Bi-LSTM Layers	2
Hidden Size	64
Learning rate	1e-3
Batch size	64
Epochs	200

3.1 Effects of Soft-BLR

Various constraints are applied to soil interfaces by adjusting the scale factor $\lambda = 0, 1, 3, 5$ and the likelihood variable α from 0.05 to 0.30. It is important to note that when $\lambda = 0$, oft-BLR corresponds to conventional one-hot encoding without an α value. Additionally, when $\lambda = 5$, the α value should not exceed 0.2 to prevent negative values in the multi-hot-like vector. Across different λ values (as shown in Fig. 6 (a)), the accuracy of the interpreted geological cross-sections improves with increasing α, then stabilizes. Compared with $\lambda = 0$ in Fig. 6 (b), the global accuracy (GA) defined in Eq. (19) shows an improvement of up to 1.06% when $\lambda = 1$ and $\alpha = 0.15$ in Fig. 6 (d). However, predicted stratigraphic boundaries would appear twisting rather than smooth curves when larger λ, α (see in Fig. 6 (e)). This phenomenon can be understood as how rapidly the likelihood of being another soil type decays away from soil interfaces. This is consistent with ACFs which indicate distant spatial dependencies. Furthermore, intersection points (red stars in Fig. 6 (b)-(e)) of multiple boundary curves located in unsampled cells could be affected by the pair parameter values in Soft-BLR. For instance, the presence of an anomaly "Soil 6" area on the right side of the BH2 bottom successfully shows geological-order characteristics.

As seen in Fig. 6 (c), the poor prediction performance near BH2 may be attributed to potential bias caused by the limited amount of "Soil 6" data (<2% of the total cells at BH2). This proves that the role of Soft-BLR can alleviate the label bias problem in the training dataset and store sufficient information reflecting subsurface sequential stratigraphy. In this paper, $\lambda = 3$ and $\alpha = 0.30$ are adopted to impose appropriate flexible soil interfaces.



(a) Variation of stratification accuracy

(b) $\lambda = 0$ without α value (GA=94.26%)

(c) $\lambda = 1$ with $\alpha = 0.30$ (GA=95.22%)

(d) $\lambda = 3$ with $\alpha = 0.30$ (GA=95.32%)

(e) $\lambda = 5$ with $\alpha = 0.20$ (GA=94.86%)

Fig. 6 Effect of flexible Soft-BLR constraints

3.2 Effects of MC-dropout

By randomly dropping neurons (see Fig. 3), an ensemble of stratigraphic configurations can be generated. As shown in Fig. 7 (a), the global accuracy trend of the predicted stratigraphic model remains consistent across different dropout ratios $\rho = \{0.1, 0.3, 0.5, 0.7\}$, with accuracy sharply rising initially before stabilizing after 40 realizations. As the dropout ratio increases (illustrated in Fig. 7 (b)), the accuracy of the most probable configuration shows a slight improvement before sharply decreasing at $\rho = 0.7$. This indicates that a moderate dropout rate helps the model learn general spatial dependencies, while a high dropout rate (>0.5) leads to underfitting. Therefore, $\rho = 0.3$ is recommended to enhance generalization and reduce overfitting. To further clarify the simulation process at $\rho = 0.3$, the initial realization and the most probable stratigraphic profile after 20 realizations with corresponding dispersion maps are disclosed in Fig. 7 (c) - (f). The dispersion map, derived from Eq. (21), measures statistical discrepancies in soil interfaces across multiple realizations. In the map, brighter yellow areas indicate greater uncertainty and the width of these regions represents the range of uncertainty. Compared to Fig. 6 (d), the Pixel Bi-LSTM with MC-dropout successfully eliminates the isolated area on the left side of the BH2 bottom. Luckily, the bottom-left region of the dispersion map in Fig. 5 (f) can highlight the

uncertainty arising from incomplete knowledge of "Soil 6" around BH2.

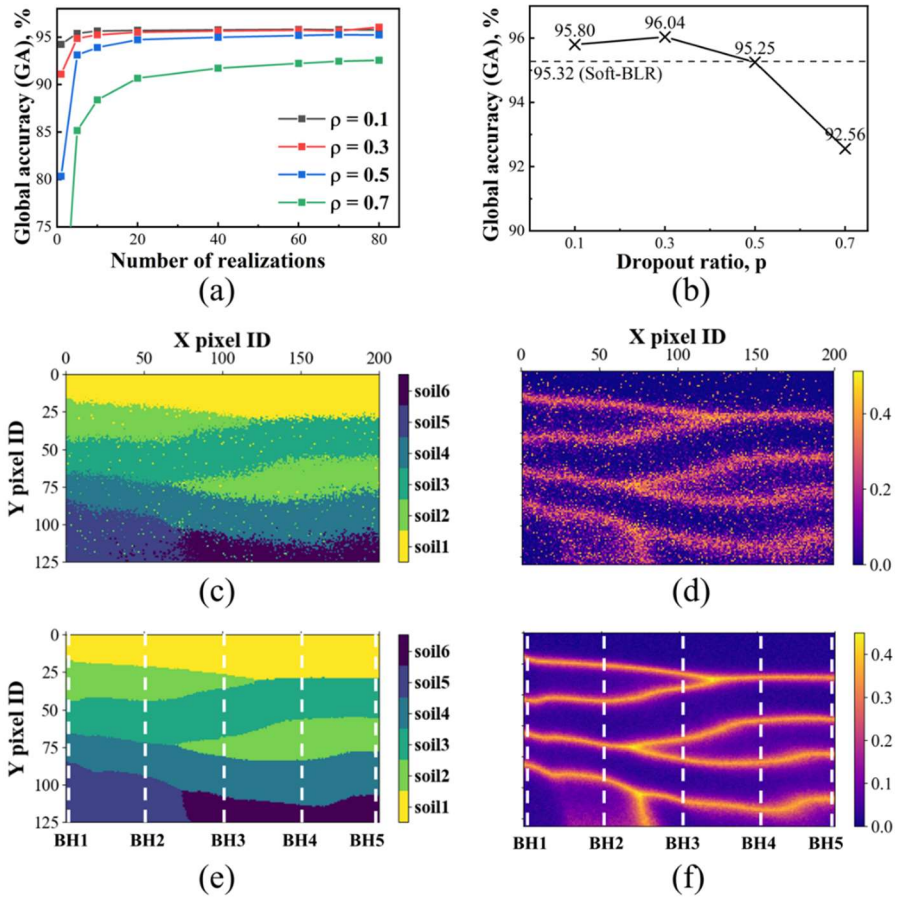


Fig. 7 Variation of prediction accuracy: (a) with realization number; (b) with different dropout ratios. One realization: (c) stratigraphic profile; (d) dispersion map. The most probable result: (e) stratigraphic profile; (f) dispersion map.

3.3 Effects of dense CRF

Nevertheless, the soil interfaces of the obtained stratigraphic profile in Fig. 7 (e) remain rough with some clustered noises. This section describes how the dense CRF technique strengthens the extracted spatial dependencies and refines smoother soil interfaces. In reference to Eq. (18), $\theta_\alpha, \theta_\beta, \theta_\gamma$ describe the strength of similarities between nearby spatial cells and $\theta_\alpha, \theta_\beta, \theta_\gamma$ define the principal eigenvector of the geological bodies. Various dense CRF structures (listed in Table 2) are modeled to clarify the roles of field kernel and spatial kernel play in stratigraphic delineation.

Table 2 Various dense CRF structures

Scenarios	$\omega^{(1)}$	$\omega^{(2)}$	θ_α	θ_β	θ_γ
(a)	6	1	(1, 6)	(1, 10)	(1, 6)

(b)	6	1	(6, 1)	(10, 1)	(6, 1)
(c)	5	3	(3, 3)	(1, 1)	(3, 3)
(d)	3	5	(3, 3)	(1, 1)	(3, 3)
(e)	5	3	(2, 5)	(1, 14)	(3, 1)

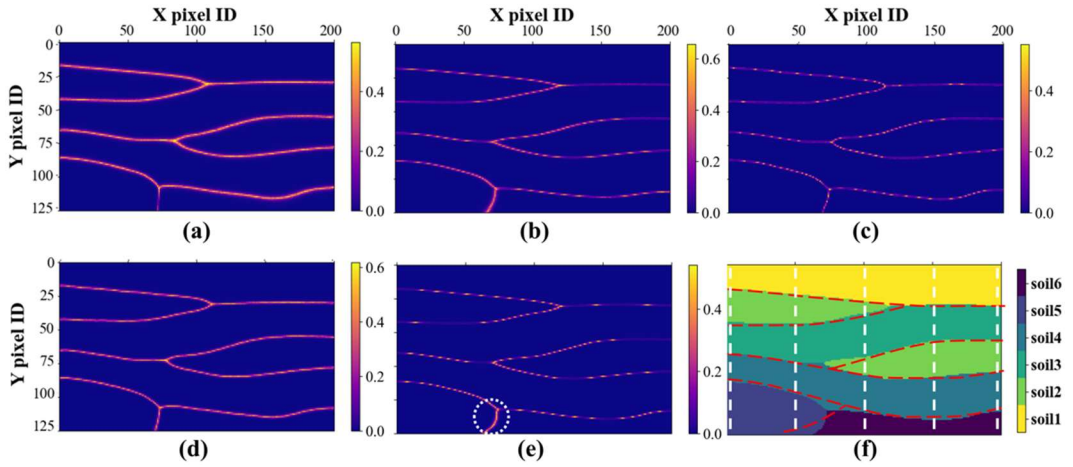


Fig. 8 Dispersion maps with (a) horizontally dominant spatial pattern; (b) vertically dominant spatial pattern; (c) strongly heterogeneous spatial pattern; (d) weakly heterogeneous spatial pattern; (e) one possible spatial pattern for this synthetic case with (f) corresponding stratigraphic profile (GA=96.12%).

In scenarios (a) and (b) using a larger appearance kernel, the boundary lines in the dispersion map show apparent directional features, with a vertically dominant spatial pattern in Fig. 8 (a) and a horizontally dominant spatial pattern in Fig. 8 (b). As shown in Fig. 8 (c) and (d), variations in the degree of geological anomaly across different soil layers can also be simulated by adjusting the parameters $\omega^{(1)}$, $\omega^{(2)}$ and θ_β which regulate the geological features of spatial heterogeneity. For this studied case, scenario (e) presents a CRF structure accounting for vertically dominant spatial patterns and distinct heterogeneity between strata. After applying the dense CRF, as shown in Fig. 8 (e) and (f), the soil interfaces are noticeably smoothed, improving prediction accuracy to 96.12%. Given the high uncertainty in the circled area of Fig. 8 (e), another experiment is conducted by virtually drilling an additional borehole to test whether prediction errors are consistent with the dispersion map. As illustrated in Fig. 9, the model's prediction accuracy further increased to 96.96% after including the additional borehole BH6 at the midpoint of BH2 and BH3. Based on the synthetic case, multiple examples with varying numbers of boreholes were generated to further evaluate the performance of the proposed method (see Fig. 10). Overall, the dense CRF technique proves effective not only in generating more accurate stratigraphic cross-sections but also in offering practical insights for optimizing site exploration strategies.

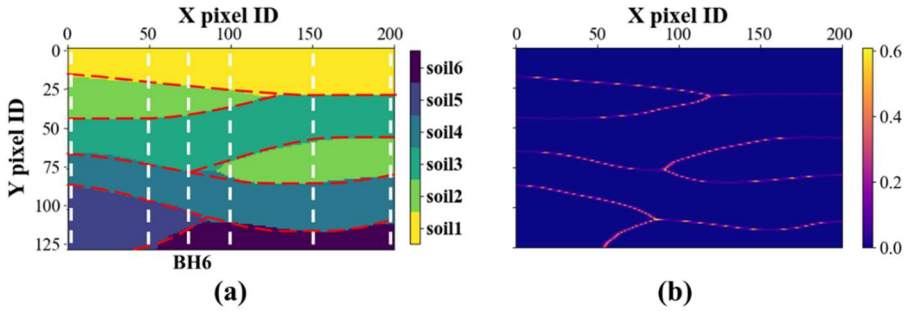


Fig. 9 (a) The most probable stratigraphic profile with (b) the corresponding dispersion map via six measurements.

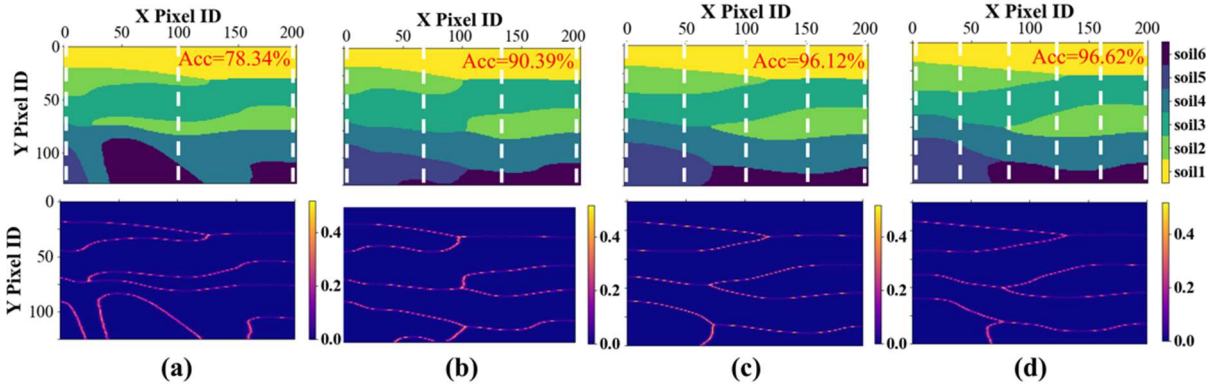


Fig. 10 Stratification results of the synthetic case: (a) via three measurements; (b) via four measurements; (c) via five measurements; (d) via six measurements.

4 Applications

In geotechnical engineering practice, accurate stratigraphic modeling based on sparse observation data is critical for numerous applications, such as settlement analysis and slope support design. Hence, the proposed hybrid neural network is applied to two different real-world borehole projects from Hong Kong: (a) data extracted from a reclamation construction site, and (b) data acquired from a cut slope project.

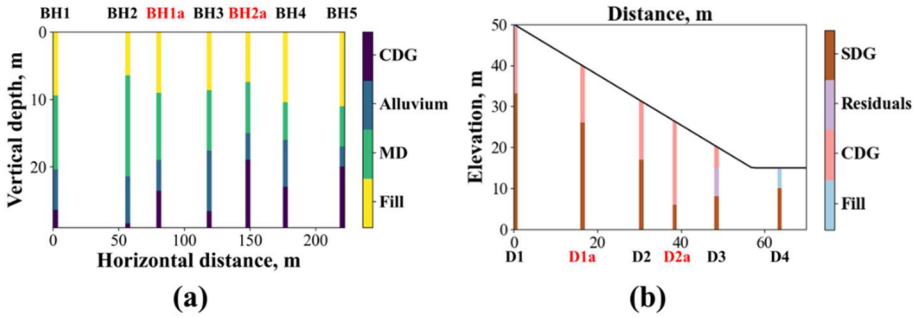


Fig. 11 (a) Stratigraphic details of collected boreholes in Hong Kong; (b) Cross-section of slope subsurface stratigraphy from collected boreholes

For the reclamation construction project, seven boreholes (labeled BH1-BH5 and BH1a-BH2a) were drilled

to investigate the subsurface geological conditions of a 29 m × 222 m cross-section site. Fig. 11 (a) presents the stratigraphic details of these boreholes, revealing soil layers consisting of Fill, Marine Deposit (MD), Alluvium, and Completely Decomposed Granite (CDG). Characterized by low strength and high compressibility, MD poses a risk of differential settlement during reclamation, which can jeopardize construction safety. Thus, deriving reliable stratigraphic boundaries from limited borehole data is crucial before foundation design and numerical analysis. In this project, the thickness of MD varies between 5.6 m in BH4 and 15 m in BH2. The vertical and horizontal resolutions of the target subsurface domain are set to 0.2 m and 1 m, respectively. To evaluate the proposed method, its performance is compared with other approaches, such as linear interpolation, ordinary kriging (OK), and Markov random field (MRF) (Wei & Wang, 2022). Additionally, four borehole schemes are designed to test the reliability of each method in interpreting geological characteristics from sparse borehole data (see Fig. 12). Note that the white dashed lines in Fig. 12 (a)-(d) represent the boreholes used for training these models. Since the complete subsurface profile is unknown, BA defined in Eq. (20) is used to assess the discrepancy between predicted and actual stratigraphy at the locations of rest untrained boreholes.

Table 3 Prediction results corresponding to different borehole schemes in the reclamation construction site

Model	Boreholes	Scheme 1	Scheme 2	Scheme 3	Scheme 5	Average time
Linear interpolation	Rest BH1-BH5	0.78	0.77	0.88		<1s
	BH1a, BH2a	0.71	0.78	0.69	0.69	
Kriging	Rest BH1-BH5	0.62	0.65	0.89		17s
	BH1a, BH2a	0.48	0.54	0.65	0.70	
MRF	Rest BH1-BH5	0.77	0.74	0.92		927s
	BH1a, BH2a	0.68	0.81	0.66	0.69	
Proposed model	Rest BH1-BH5	0.80	0.87	0.93		16s
	BH1a, BH2a	0.77	0.68	0.69	0.72	

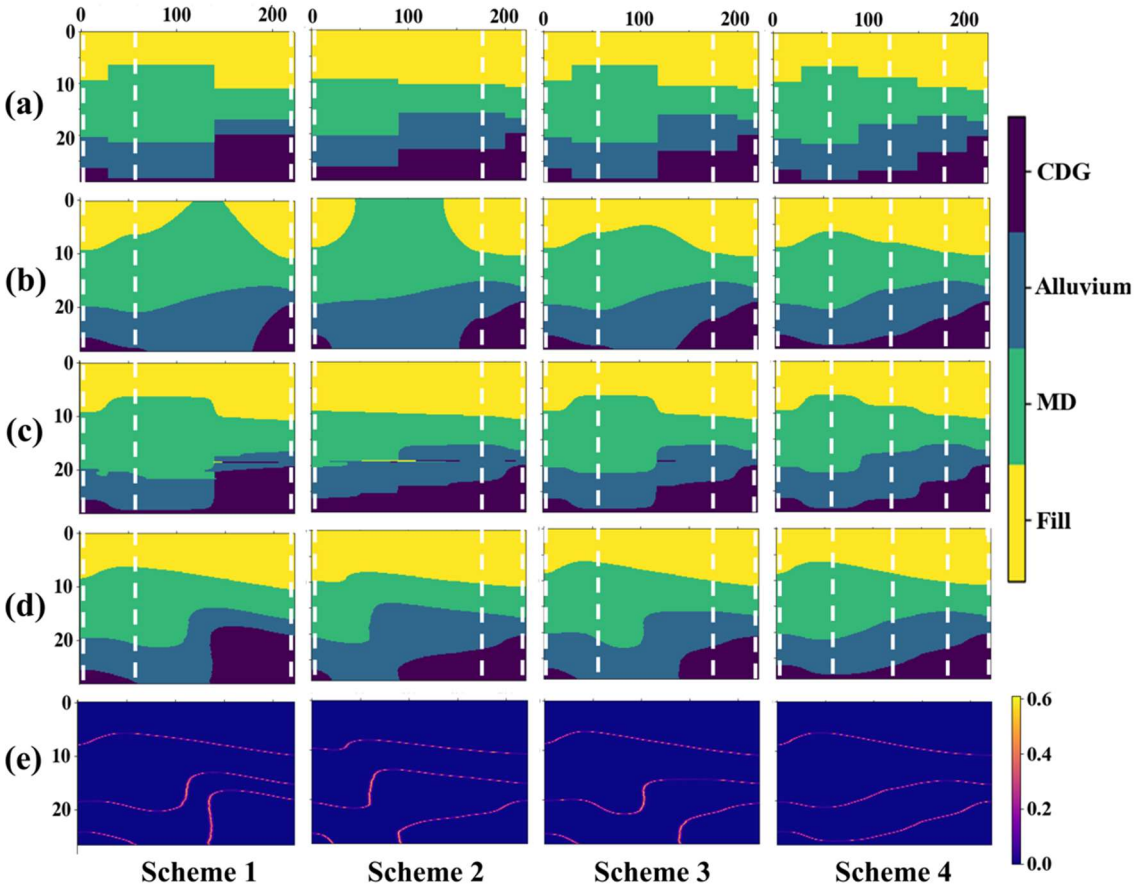


Fig. 12 Stratification results of the reclamation construction case: (a) via the linear interpolation; (b) via the OK; (c) via the MRF; (d) via the proposed model with (e) corresponding dispersion maps.

Table 3 summarizes the BA of each model across different borehole schemes. Except for the performance of MRF in Scheme 2 on BH1a-BH2a, the proposed model achieves a higher accuracy than other methods. However, in Schemes 4 and 5, where more training boreholes are included, the BA of MRF drops significantly from 81% to 66% and 69%, respectively. Leveraging the advantage of GPU parallel computing, the proposed model's inference process takes an average of just 18 seconds, significantly faster than the 927 seconds required by the MRF model. Although linear interpolation completes stratigraphic delineation in under 1 second, its poor performance in Fig. 12 (a) makes it unsuitable for generating high-resolution profiles from sparse data. In contrast, the MRF model often introduces unrealistic anomalies within the geological bodies, particularly in Schemes 1-3 (see Fig. 12 (c)). In contrast, both the proposed model and the OK method produce smoother soil interfaces from limited borehole data, which is advantageous for meshing in numerical simulations. However, OK tends to overestimate the thickness of MD in Schemes 1 and 2 (Fig. 12 (b)), indicating it fails to capture the stratigraphic features (e.g., the fill layer overlays MD). As depicted in Fig. 12 (e), the proposed model's dispersion maps highlight that interpolation uncertainty is clustered around the interfaces between MD, Alluvium, and CDG, consistent with the

observed interpolation errors. The proposed probabilistic model provides an effective tool for geotechnical engineers, enabling the creation of reasonable soil profiles and supporting reliability-based foundation design.

The second practical case involves a cut slope near a residential estate, investigated using six boreholes, including the initial survey (D1-D4) and supplementary drillings (D1a-D2a). As shown in Fig. 11 (b), the slope has an elevation of 35.0 m with an inclination angle of nearly 32°, and the stratigraphy consists of Fill, CDG, Residual Soil, and Slightly Decomposed Granite (SDG). Notably, Residual Soil, found only in borehole D3 with a thickness of 7.5 m, is mainly composed of sandy silt/clay and should be considered a weak clay seam. In Hong Kong, where heavy rainfall is common, the presence of this weak seam at the slope toe can potentially trigger shallow slope failure. Therefore, designing lateral support for this case requires precise characterization of the subsurface weak seam based on limited field data. Using different dense CRF structures outlined in Table 2, Fig. 13 shows various stratification results based on the four initial boreholes. When boreholes D1a and D2a are used to evaluate the model's performance, the borehole accuracy of the stratigraphic cross-sections ranges from 88% to 90%. However, the identified scope of the weak seam varies slightly in Fig. 13 (a). Fig. 13 (b) also reveals significant uncertainties along the vertical soil boundaries in all dispersion maps under different CRF structures, indicating a need for additional borehole data in these areas. The interpreted soil interfaces between CDG, Residual Soil, and SDG show strong heterogeneity in the dispersion maps. With the inclusion of borehole data from D1a and D2a, the model's understanding of the subsurface domain improves, as shown in Fig. 14 using Scenario 5's CRF settings. Fig. 14 (b) confirms the vertically dominant morphological characteristics of the soil boundaries, aligning with the interpolation errors around the weak seam in Fig. 13. Additional boreholes are recommended between D3 and D4 for more accurate stratification. Given that the most probable result in Fig. 14 (a) indicates about 3.0% silt/clay content, permanent support structures like shotcrete facing with soil nails are proposed for future design work. This case demonstrates the model's capability to produce an ensemble of possible soil profiles for characterizing the weak seam. The geostatistical uncertainty depicted in the dense CRF structure's dispersion map highlights the model's potential for guiding fieldwork and improving geotechnical understanding.

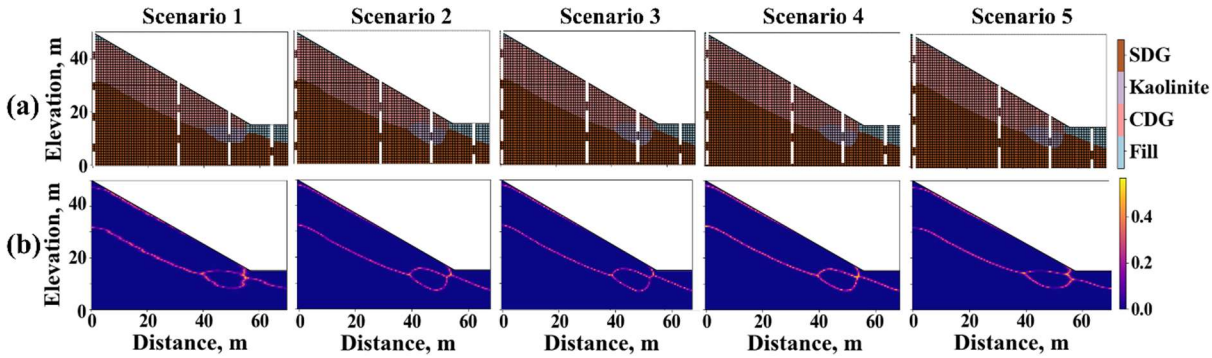


Fig. 13 Stratification results of the cut slope case under different dense CRF structures via four measurements: (a) most probable configuration; (b) dispersion map.

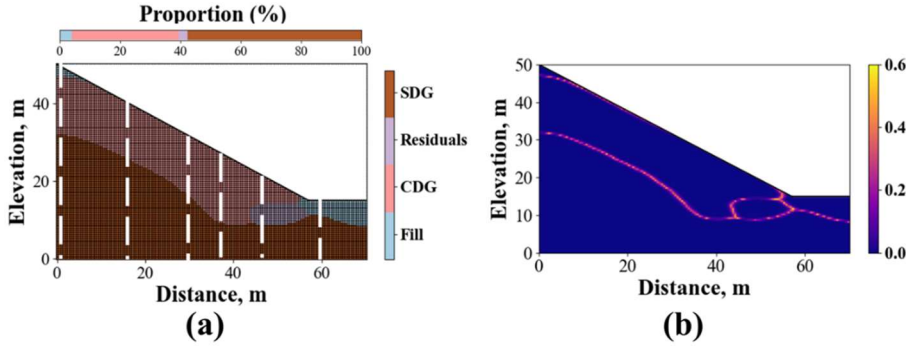


Fig. 14 Stratification results of the cut slope case via six measurements: (a) most probable configuration; (b) dispersion map.

5 Conclusions

A novel hybrid neural network is proposed for probabilistic subsurface stratigraphic modeling that considers the label bias problem, geostatistical uncertainty, and different spatial distribution patterns. The proposed approach is illustrated by the synthetic case with sparse boreholes and applied to two real cases in Hong Kong. Within the proposed hybrid neural network framework, complex stratigraphic patterns can be extracted purely from sparse site-specific measurements and the dispersion map of the most probable stratigraphic cross-section aligns with the precision of stratigraphic characterization. The proposed coordinate-based hybrid model can be easily extended for 3D stratigraphic modeling by adding a Bi-LSTM branch to capture spatial dependencies from the third dimension. Due to the word limit, this paper focuses on demonstrating the robust reliability of the proposed hybrid neural network in comparison with existing methods. Future research will concentrate on applying this hybrid neural network to 3D stratigraphic modeling, facilitating reliability-based geotechnical analysis and design.

Acknowledgments

This research was financially supported by the Research Grants Council (RGC) of the Hong Kong Special Administrative Region Government (HKSARG) of China (Grant no. 15220221, 15227923, 15229223, 15220423).

Appendix:

Table. The grid search for determining suggested hyperparameters

Bi-LSTM Layers	Hidden Size	Learning Rate	Batch size	Elapsed Time	Accuracy
1	64	1e-2	64	2s	89.00%
2	64	1e-2	64	3s	95.14%
3	64	1e-2	64	3s	94.03%
2	16	1e-2	64	3s	89.74%
2	32	1e-2	64	3s	94.32%
2	128	1e-2	64	3s	94.64%
2	64	1e-1	64	3s	85.14%
2	64	1e-3	64	3s	93.11%
2	64	1e-2	32	3s	95.04%
2	64	1e-2	128	2s	94.38%

References:

Cami, B., Javankhoshdel, S., Phoon, K.-K., & Ching, J. (2020). Scale of Fluctuation for Spatially Varying Soils: Estimation Methods and Values. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *6*(4), 03120002. https://doi.org/10.1061/AJRUA6.0001083

Chen, H., & Lee, C. F. (2004). Geohazards of slope mass movement and its prevention in Hong Kong. *Engineering Geology*, *76*(1), 3–25. https://doi.org/10.1016/j.enggeo.2004.06.003

Cheng, Z., Wang, J., & Xiong, W. (2023). A machine learning-based strategy for experimentally estimating force chains of granular materials using X-ray micro-tomography. *Géotechnique*, 1–13. https://doi.org/10.1680/jgeot.21.00281

Damianou, A., & Lawrence, N. D. (2013). Deep Gaussian Processes. In C. M. Carvalho & P. Ravikumar (Eds.),

456 *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (Vol. 31,
457 pp. 207–215). PMLR. https://proceedings.mlr.press/v31/damianou13a.html
458 Furitsu, Y., Deguchi, D., Kawanishi, Y., Ide, I., Murase, H., Mukojima, H., & Nagamine, N. (2021). Soft-Boundary
459 Label Relaxation with class placement constraints for semantic segmentation of the railway environment.
460 *Pattern Recognition Letters*, *150*, 258–264. https://doi.org/10.1016/j.patrec.2021.07.014
461 Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in
462 Deep Learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of The 33rd International*
463 *Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR.
464 https://proceedings.mlr.press/v48/gal16.html
465 Gao, B., Wang, R., Lin, C., Guo, X., Liu, B., & Zhang, W. (2021). TBM penetration rate prediction based on the
466 long short-term memory neural network. *Underground Space*, *6*(6), 718–731.
467 https://doi.org/10.1016/j.undsp.2020.01.003
468 Gong, W., Zhao, C., Juang, C. H., Tang, H., Wang, H., & Hu, X. (2020). Stratigraphic uncertainty modelling with
469 random field approach. *Computers and Geotechnics*, *125*, 103681.
470 https://doi.org/10.1016/j.compgeo.2020.103681
471 Hou, S., Liu, Y., & Yang, Q. (2022). Real-time prediction of rock mass classification based on TBM operation big
472 data and stacking technique of ensemble learning. *Journal of Rock Mechanics and Geotechnical*
473 *Engineering*, *14*(1), 123–143. https://doi.org/10.1016/j.jrmge.2021.05.004
474 Jaksa, M. B., Goldsworthy, J. S., Fenton, G. A., Kaggwa, W. S., Griffiths, D. V., Kuo, Y. L., & Poulos, H. G.
475 (2005). Towards reliable and effective site investigations. *Géotechnique*, *55*(2), 109–121.
476 https://doi.org/10.1680/geot.2005.55.2.109
477 Jamshidi Chenari, R., Kamyab Farahbakhsh, H., Heidarie Golafzani, S., & Eslami, A. (2018). Non-stationary
478 realisation of CPT data: Considering lithological and inherent heterogeneity. *Georisk: Assessment and*
479 *Management of Risk for Engineered Systems and Geohazards*, *12*(4), 265–278.
480 Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
481 https://books.google.co.in/books?id=7dzpHCHzNQ4C
482 Krähenbühl, P., & Koltun, V. (2011). Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials.
483 *Advances in Neural Information Processing Systems*, *24*.
484 https://proceedings.neurips.cc/paper/2011/hash/beda24c1e1b46055dff2c39c98fd6fc1-Abstract.html
485 Li, D.-Q., Jiang, S.-H., Cao, Z.-J., Zhou, W., Zhou, C.-B., & Zhang, L.-M. (2015). A multiple response-surface

method for slope reliability analysis considering spatial variability of soil properties. *Engineering Geology*, *187*, 60–72. https://doi.org/10.1016/j.enggeo.2014.12.003

Li, X. Y., Zhang, L. M., & Li, J. H. (2016). Using Conditioned Random Field to Characterize the Variability of Geologic Profiles. *Journal of Geotechnical and Geoenvironmental Engineering*, *142*(4), 04015096. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001428

Lin, H. W., & Tegmark, M. (2017). Critical Behavior in Physics and Probabilistic Formal Languages. *Entropy*, *19*(7). https://doi.org/10.3390/e19070299

Liu, L.-L., Cheng, Y.-M., & Zhang, S.-H. (2017). Conditional random field reliability analysis of a cohesion-frictional slope. *Computers and Geotechnics*, *82*, 173–186. https://doi.org/10.1016/j.compgeo.2016.10.014

National Research Council. (1984). *Geotechnical Site Investigations for Underground Projects: Volume 1*. The National Academies Press. https://doi.org/10.17226/919

Ninić, J., Freitag, S., & Meschke, G. (2017). A hybrid finite element and surrogate modelling approach for simulation and monitoring supported TBM steering. *Tunnelling and Underground Space Technology*, *63*, 12–28. https://doi.org/10.1016/j.tust.2016.12.004

Phoon, K.-K., & Kulhawy, F. H. (1999a). Characterization of geotechnical variability. *Canadian Geotechnical Journal*, *36*(4), 612–624. https://doi.org/10.1139/t99-038

Phoon, K.-K., & Kulhawy, F. H. (1999b). Evaluation of geotechnical property variability. *Canadian Geotechnical Journal*, *36*(4), 625–639. https://doi.org/10.1139/t99-039

Qi, X.-H., Li, D.-Q., Phoon, K.-K., Cao, Z.-J., & 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

Samui, P., & Sitharam, T. G. (2010). Site Characterization Model Using Artificial Neural Network and Kriging. *International Journal of Geomechanics*, *10*(5), 171–180. https://doi.org/10.1061/(ASCE)1532-3641(2010)10:5(171)

Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: Binary versus one-hot and feature hashing*. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-237426

Shi, C., & Wang, Y. (2021a). 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), 04021082.

https://doi.org/10.1061/(ASCE)GT.1943-5606.0002583

Shi, C., & Wang, Y. (2021b). 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., & Wang, Y. (2023). Data-driven spatio-temporal analysis of consolidation for rapid reclamation. *Géotechnique*, 1–21. https://doi.org/10.1680/jgeot.22.00016

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

Wang, Y., & Zhao, T. (2017). Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique*, *67*(6), 523–536. https://doi.org/10.1680/jgeot.16.P.143

Wei, X., & Wang, H. (2022). Stochastic stratigraphic modeling using Bayesian machine learning. *Engineering Geology*, *307*, 106789. https://doi.org/10.1016/j.enggeo.2022.106789

Xie, P., Chen, K., Skibniewski, M. J., Wang, J., & Luo, H. (2023). Parametric geological model update and probabilistic analysis of shield tunnel excavation: A borehole-based conditional random fields approach. *Computers and Geotechnics*, *157*, 105349. https://doi.org/10.1016/j.compgeo.2023.105349

Zhang, J., Phoon, K. K., Zhang, D., Huang, H., & Tang, C. (2021). Deep learning-based evaluation of factor of safety with confidence interval for tunnel deformation in spatially variable soil. *Journal of Rock Mechanics and Geotechnical Engineering*, *13*(6), 1358–1367. https://doi.org/10.1016/j.jrmge.2021.09.001

Zhang, J.-Z., Liu, Z.-Q., Zhang, D.-M., Huang, H.-W., Phoon, K.-K., & Xue, Y.-D. (2022). Improved coupled Markov chain method for simulating geological uncertainty. *Engineering Geology*, *298*, 106539. https://doi.org/10.1016/j.enggeo.2022.106539

Zhang, P., Yin, Z.-Y., & Sheil, B. (2023). A physics-informed data-driven approach for consolidation analysis. *Géotechnique*, 1–12. https://doi.org/10.1680/jgeot.22.00046

Zhao, C., Gong, W., Li, T., Juang, C. H., Tang, H., & Wang, H. (2021). Probabilistic characterization of subsurface stratigraphic configuration with modified random field approach. *Engineering Geology*, *288*, 106138. https://doi.org/10.1016/j.enggeo.2021.106138

Zhou, C., Ouyang, J., Ming, W., Zhang, G., Du, Z., & Liu, Z. (2019). A Stratigraphic Prediction Method Based on Machine Learning. *Applied Sciences*, *9*(17). https://doi.org/10.3390/app9173553