Bulletin of Engineering Geology and the Environment (2021) 80:9157–9171
https://doi.org/10.1007/s10064-021-02478-x

ORIGINAL PAPER

Machine learning-enhanced soil classification by integrating borehole and CPTU data with noise filtering

Te Xiao[1] · Hai-Feng Zou[1] · Ke-Sheng Yin[1] · Yu Du[2,3] · Li-Min Zhang[1,4]

Received: 30 July 2020 / Accepted: 20 October 2021 / Published online: 26 October 2021

Abstract

Integrating borehole and piezocone penetration test (CPTU) data in site characterization helps to achieve a more comprehensive understanding of ground conditions. However, soil types at CPTU and nearby borehole locations may not always be consistent. The presence of noisy data or thin layers will mislead the interpretation of CPTU data in soil type classification and soil property evaluation. This study proposes a coupled machine learning method to integrate the borehole and CPTU data under a rigorous Bayesian framework and to identify and separate the noisy CPTU data without subjective judgment, which contributes to more reliable soil classification and property evaluation. The borehole-reported soil type and CPTU data are treated as two types of evidence of the authentic soil type. A lateral transition of soil type from the CPTU location to the borehole location is allowed to capture the discrepancy of soil types. The proposed approach is applied to the marine site characterization of the Hong Kong-Zhuhai-Macao Bridge that crosses the Pearl River Estuary of China. The soil seams embedded in the dominant soil strata are successfully detected, producing a more reliable soil profile and interpreting more compatible soil properties with engineering practice. Additionally, the integration of borehole and CPTU data significantly reduces the stratification uncertainty in site characterization.

Keywords Site investigation · Soil classification · Piezocone penetration test · Machine learning · Hong Kong-Zhuhai-Macao Bridge

Highlights

- A coupled machine learning method is proposed to integrate borehole and CPTU data under a rigorous Bayesian framework.
- A lateral transition of soil type from the CPTU location to the borehole location is allowed to capture the discrepancy of soil types.
- The method is applied to the marine site characterization of the Hong Kong-Zhuhai-Macao Bridge.
- Noisy CPTU data are filtered to achieve more reliable soil type classification and soil property evaluation.
- The integration of borehole and CPTU data significantly reduces the stratification uncertainty.

✉ Li-Min Zhang
cezhangl@ust.hk

Te Xiao
xiaote@ust.hk

Hai-Feng Zou
zhf@ust.hk

Ke-Sheng Yin
kyinac@connect.ust.hk

Yu Du
duy@fhdigz.com

Introduction

The piezocone penetration test (CPTU) has been widely used in geotechnical site investigation for soil type classification and property evaluation in terms of theoretical (e.g., Mo et al. 2017), empirically deterministic (e.g., Robertson 1990; Lunne et al. 1997; Mayne 2007; Schneider et al. 2008; Li et al. 2020; Yahşi and Ersoy 2021; Yin et al. 2021), and empirically probabilistic (e.g., Ching et al. 2015; Li et al. 2016c; Zhao et al. 2018; Wang et al. 2013, 2019a, b) manners. It provides

[1] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[2] Institute of Geotechnical Engineering, Southeast University, Nanjing, China

[3] CCCC-FHDI Engineering CO., LTD, Guangzhou, China

[4] HKUST Shenzhen Research Institute, Shenzhen, China

nearly continuous but indirect mechanical measurements of soil behavior in the vertical direction. Due to the indirect nature of CPTU, a verification procedure is required by comparing soil types from drilled boreholes and soil properties from laboratory tests (i.e., direct observations/measurements) with those interpreted from in situ CPTU data. Nevertheless, the inherent spatial variability of soils (e.g., Phoon and Kulhawy 1999; Phoon 2020) and the lateral separations between CPTU and borehole locations make soil types at the two locations not always be consistent. Figure 1 gives an example of CPTU-based soil classification. The soil segment of the borehole indicates a homogenous clay layer, while a thin sandy seam exists in the adjacent CPTU sounding. Such a thin layer separates the CPTU data into a dominant dataset and a noisy dataset. The noisy data will mislead the interpretation of CPTU data in soil property evaluation, resulting in significant bias on uncertainty quantification, and will have considerable impacts on subsequent geotechnical analysis and design. It is, therefore, of great importance to identify the noisy data and filter them from the dominant data.

Current practice identifying noises or thin layers in CPTU data mainly relies on visual inspection and subjective judgment from experienced engineers. It is burdensome to manually deal with a large amount of data (e.g., hundreds of soundings) in major projects. Besides, the results may vary significantly from one engineer to another, especially when the boundaries between noisy and dominant data are not distinct. Statistical outlier detection approaches help to identify noises more objectively (e.g., Hodge and Austin 2004; Yuen and Mu 2012; Zheng et al. 2021), among which some are particularly developed for CPTU, such as the wavelet transform modulus maxima method (Ching et al. 2015), spatial filtering method (Boulanger and DeJong 2018), and empirical screening method (Du et al. 2020). These methods based on sole CPTU data cannot properly interpret the physical meaning of noisy data, namely the authentic soil type. Recall that borehole logs used in the CPTU verification are direct observations of soil types. It is desirable to integrate borehole and CPTU data to facilitate a more comprehensive understanding of soil types. Several machine learning techniques have been developed for this purpose, such as the Bayesian mixture analysis (Depina et al. 2016) and hidden Markov random field model (Wang et al. 2019a, b). However, they cannot address the aforementioned borehole-CPTU inconsistency well. It is essential and challenging to filter noises automatically during the integration of borehole and CPTU data.

The objectives of this study are threefold: (1) to propose a coupled machine learning method to identify noisy CPTU data with the help of borehole information; (2) to improve the reliability of soil type classification and soil property evaluation by separating dominant data from noisy data; and (3) to conduct a case study of the marine site characterization of the Hong Kong-Zhuhai-Macao Bridge on the Pearl River Estuary of China, the longest bridge-cum-tunnel sea crossing in the world. The development of the proposed method for soil classification, estimation of model parameters, and extension to soil property evaluation with noise filtering will be introduced, respectively. Afterwards, a detailed example of the Hong Kong-Zhuhai-Macao Bridge will be investigated to illustrate how to enhance soil classification and property evaluation results in practice using the proposed approach.

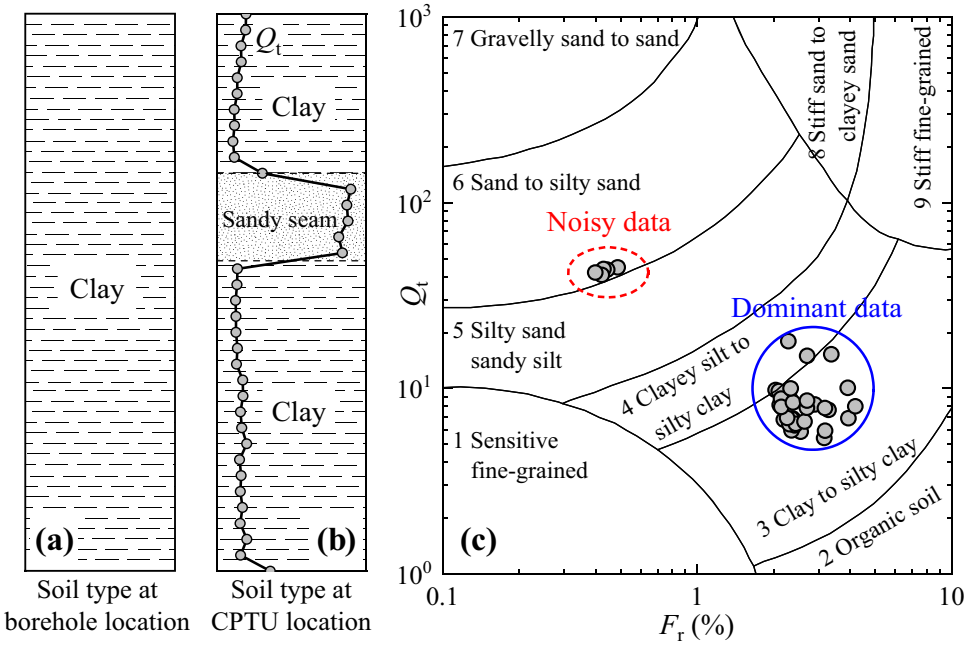

Fig. 1 Example of noisy data in soil classification: **a** and **b** soil types at borehole and CPTU locations; **c** CPTU-based soil classification

Machine learning-enhanced soil classification

Physical principles

In this study, borehole and CPTU data are connected through the authentic soil type at the CPTU location, which is what engineers actually concern about but cannot be observed directly. As shown in Fig. 2, the authentic soil type, CPTU data, and borehole-reported soil type are denoted as x_s, $\boldsymbol{y}_s$, and w_s, respectively, at a specific location s ($s = 1, 2, \ldots, S$, where S is the number of locations in terms of CPTU) and $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{w}$, respectively, over the whole site. Although the borehole location laterally deviates from the CPTU location to a certain distance, its location is also labeled as s for simpler notation. The borehole-reported soil type and CPTU data are treated as two types of evidence of the authentic soil type. Together with the vertical self-similarity of the authentic soil type, they form the three vital physical principles of the proposed method.

Firstly, the authentic soil type at one location (s) tends to remain similar to others in its neighborhood (∂s), as observed in engineering practice. Since soil classification is modeled as a one-dimensional problem along with the depth in this study, only vertical self-similarity will be considered, the degree of which depends on the vertical spatial variability of soil types.

Secondly, the CPTU data describing the mechanical behavior of soil reflect soil types at the target location indirectly, which is a common foundation of traditional CPTU-based soil classification. Among various CPTU indices, the logarithmic values of normalized cone tip resistance (Q_t) and normalized friction ratio (F_r) are the two most widely used indices in a variety of CPTU-based soil classification charts (e.g., Robertson 1990, 2009; Schneider et al. 2008); hence they are adopted in this study as well, i.e., $\boldsymbol{y} = \{\log F_r, \log Q_t\}$.

Thirdly, the nearby verification borehole data can be viewed as direct observations of soil types but not at the target location. For simplicity, previous studies usually treated the borehole-observed soil types as the authentic soil types at their adjacent CPTU locations. Such a simplification lumps the dominant and noisy CPTU data together, which may reduce the reliability of soil classification. To address this problem, a lateral transition of soil type from the CPTU location to the borehole location will be allowed in this study, which relies on the horizontal spatial variability of soil types.

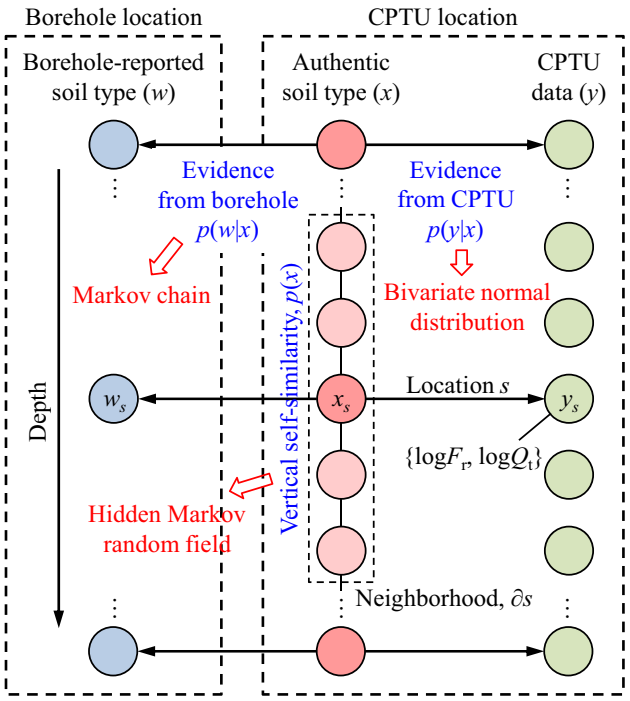

Fig. 2 Three physical principles in the coupled machine learning method

Bayesian formulation

In this study, x_s and w_s are categorical variables, namely $\{x_s = l, l = 1, 2, \ldots, L\}$ and $\{w_s = k, k = 1, 2, \ldots, K\}$, where L and K are the numbers of authentic soil types at CPTU locations and borehole-reported soil types at borehole locations, respectively, and $\boldsymbol{y}_s$ is a set of numerical variables. Based on the three physical principles, the task of soil classification is to infer the posterior probability of x_s given observations w_s and $\boldsymbol{y}_s$, i.e., $p(x_s|\boldsymbol{y}_s, w_s)$, which can be derived according to the Bayes' theorem as:

$$p(x_s|\boldsymbol{y}_s, w_s) \propto p(x_s)p(\boldsymbol{y}_s, w_s|x_s) = p(x_s)p(\boldsymbol{y}_s|x_s)p(w_s|x_s) \tag{1}$$

where $p(x_s)$ is the prior probability of x_s without any sampling or testing information; $p(\boldsymbol{y}_s, w_s|x_s)$ is the joint likelihood of observing w_s and $\boldsymbol{y}_s$ given x_s; and $p(\boldsymbol{y}_s|x_s)$ and $p(w_s|x_s)$ are the likelihoods of observing $\boldsymbol{y}_s$ and w_s given x_s, respectively. It is practically reasonable to assume that w_s and $\boldsymbol{y}_s$ are independent when conditioned on x_s. This is because the measurement of CPTU data for a given soil type only depends on the mechanical behavior of soil and can be regarded as a random observation, which is irrelevant to the soil type from borehole drilling. The prior probability, $p(x_s)$, relies on the vertical self-similarity of the authentic soil type, and the two likelihoods, $p(\boldsymbol{y}_s|x_s)$ and $p(w_s|x_s)$, correspond to the evidence from CPTU and borehole, respectively. Different machine learning models are required to describe them properly.

Prior probability of authentic soil type

The vertical self-similarity of soil types can be described using a hidden Markov random field model, in which $p(x_s)$ can be determined according to the soil types in its neighborhood (∂s) as (e.g., Li 2009; Wang et al. 2019b):

$$p(x_s) \propto \exp\left[-\sum_{j\in\partial s} V(x_s, x_j;\beta)\right] \tag{2}$$

where $V(\cdot)$ is the clique potential: $V=-\beta$ when $x_s=x_j$; otherwise, $V=0$. The model parameter, β (≥ 0), determines the degree of self-similarity between soil types. A large β value indicates a relatively homogeneous soil layer; on the contrary, $\beta=0$ means that all soil types are mutually independent and substantial heterogeneity is expected.

Likelihood of CPTU data

The likelihood of observing a set of CPTU data $\boldsymbol{y}_s=\{\log F_r, \log Q_t\}$ for a given soil type (e.g., $x_s=l$) is usually modeled as a bivariate normal distribution in geotechnical literature:

$$p(\boldsymbol{y}_s|x_s=l) = (2\pi)^{-1}|\boldsymbol{\Sigma}_l|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{y}_s-\boldsymbol{\mu}_l)^{\mathrm{T}}\boldsymbol{\Sigma}_l^{-1}(\boldsymbol{y}_s-\boldsymbol{\mu}_l)\right] \tag{3}$$

where $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are the mean vector and covariance matrix of the CPTU data for the lth soil type and $\boldsymbol{\Sigma}_l$ is consisted of two components, namely the standard deviation $\boldsymbol{\sigma}_l$ and the correlation coefficient ρ between $\log F_r$ and $\log Q_t$.

Likelihood of borehole-reported soil type

As the location of a CPTU sounding and its verification borehole do not overlap exactly, their soil types may not be completely consistent. Such an inconsistency can be modeled probabilistically as a lateral transition of soil type from the CPTU location to the borehole location using a Markov chain model (e.g., Elfeki and Dekking 2005; Li et al. 2016a; Xiao et al. 2017). According to the total probability theorem, the probabilities of soil types at borehole and CPTU locations, $p(w_s)$ and $p(x_s)$, can be linked through a conditional probability (or a transition probability), $p(w_s|x_s)$, as:

$$p(w_s=k) = \sum_l p(w_s=k|x_s=l)p(x_s=l) \tag{4}$$

Consider that $p(w_s|x_s)=p(w|x)$, i.e., the lateral transition among different soil types remains consistent over the site. Equation (4) can be rewritten in a matrix form, in the context of Markov chain model, for all locations as:

$$\boldsymbol{p}_{\boldsymbol{w}} = \boldsymbol{p}_{\boldsymbol{x}}\boldsymbol{\pi} \tag{5}$$

where $\boldsymbol{p}_{\boldsymbol{w}}$ and $\boldsymbol{p}_{\boldsymbol{x}}$ are the probabilities of possible soil types at all borehole and CPTU locations, respectively, and their dimensions are $S\times K$ and $S\times L$, respectively; and $\boldsymbol{\pi}$ is an $L\times K$ transition probability matrix from $\boldsymbol{x}$ to $\boldsymbol{w}$, in which the (l, k)th entry is the probability of observing $w=k$ given $x=l$. Slightly different from the previous two models, $p(w_s|x_s)$ or $\boldsymbol{\pi}$ itself is a model parameter that must be characterized based on borehole and CPTU data. If a site is ideally homogeneous without any spatial variability in soil type over the lateral distance, $L=K$ and x will always be the same as w, leading to an identity matrix of $\boldsymbol{\pi}$ mathematically. Incorporating transition probability makes possible the identification of any soil seams between CPTU soundings and boreholes and consequently the filtering of noisy data.

In summary, authentic soil type $\boldsymbol{x}$ is modeled as a hidden Markov random field that cannot be observed directly, and the borehole-reported soil type $\boldsymbol{w}$ and CPTU data $\boldsymbol{y}$ are connected to $\boldsymbol{x}$ by a Markov chain and a bivariate normal distribution, respectively, as shown in Fig. 2. The coupled machine learning method involves a total of four model parameters $\{\beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$. Once they are quantified, it is easy to calculate $p(x_s)$, $p(w_s|x_s)$, and $p(\boldsymbol{y}_s|x_s)$ to gain the posterior distribution of the authentic soil type at every spatial location using Eq. (1). Then, the most probable soil type and the associated stratification uncertainty can be determined according to the criterion of maximum a posterior and information entropy, respectively, based on multiple simulation realizations (e.g., Elfeki and Dekking 2005; Xiao et al. 2017; Wang et al. 2019a; Zhao and Wang 2019).

Estimation of model parameters

Given observations from multiple CPTU soundings and their adjacent verification boreholes, i.e., $\boldsymbol{y}$ and $\boldsymbol{w}$, the model parameters $\{\beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ cannot be directly estimated as they are dependent on the authentic soil types $\boldsymbol{x}$ that cannot be observed. Instead, an iterative procedure based on the expectation–maximization algorithm (e.g., Bishop 2006; Li 2009) can be applied to estimate the model parameters as follows:

(1) Provide an initial guess of model parameters;
(2) Obtain the posterior distribution of $\boldsymbol{x}$ using current model parameters;
(3) Update the model parameters using the posterior distribution of $\boldsymbol{x}$; and
(4) Repeat the above two steps until the maximization is converged.

The third step is undoubtedly the key. A closed-form solution exists for updating $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ ($l=1, 2, \ldots, L$) according to the posterior probability of $\boldsymbol{x}$ and can be referred to Bishop (2006). Updating transition probability matrix $\boldsymbol{\pi}$ needs to solve the transition equation, i.e., Eq. (5). Note that $\boldsymbol{p}_{\boldsymbol{w}}$ is

the probability of the borehole-reported soil type that has been observed. It is independent of model parameters and can be easily evaluated at the beginning. For example, if the borehole-reported soil type is k at location s, the (s, k)th entry of $\boldsymbol{p}_w$ is taken as one, and the other elements of the sth row are all zeros. Meanwhile, the posterior probability of $\boldsymbol{x}$ at all locations, $\boldsymbol{p}_x$, has been obtained in the second step. Consequently, the transition probability matrix can be obtained using the least-squares method as:

$$\boldsymbol{\pi} = \left(\boldsymbol{p}_x^{\mathrm{T}}\boldsymbol{p}_x\right)^{-1}\boldsymbol{p}_x^{\mathrm{T}}\boldsymbol{p}_w \tag{6}$$

Regarding the updating of the degree of vertical self-similarity, β, numerical methods such as maximum likelihood estimation can be applied, which requires the maximization of the joint likelihood function of observations $\boldsymbol{w}$ and $\boldsymbol{y}$, i.e., $p(\boldsymbol{w}, \boldsymbol{y}|\beta)$, over the whole site. In fact, $\boldsymbol{w}$ and $\boldsymbol{y}$ are not directly related to β, but $\boldsymbol{x}$ is. Therefore, $p(\boldsymbol{w}, \boldsymbol{y}|\beta)$ can be calculated by marginalizing $p(\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{x}|\beta)$ with respect to $\boldsymbol{x}$ according to the total probability theorem:

$$p(\boldsymbol{w}, \boldsymbol{y}|\beta) = \sum_l p(\boldsymbol{w}, \boldsymbol{y}|\boldsymbol{x} = l)p(\boldsymbol{x} = l|\beta) \tag{7}$$

Following a commonly used assumption that all observations are conditionally independent among different spatial locations, Eq. (7) can be approximated in a pseudo-likelihood form as (Li 2009):

$$p(\boldsymbol{w}, \boldsymbol{y}|\beta) \approx \prod_s \sum_l p\left(w_s|x_s = l\right)p\left(\boldsymbol{y}_s|x_s = l\right)p\left(x_s = l|\beta\right) \tag{8}$$

where $p(w_s|x_s=l)=p(w|x)=\boldsymbol{\pi}$, $p(\boldsymbol{y}_s|x_s=l)$, and $p(x_s=l|\beta)=p(x_s)$ are calculated using Eqs. (6), (3), and (2), respectively. The maximum likelihood estimation can be further extended to Bayesian updating if additional prior information on β is provided.

In fact, both borehole and CPTU data are spatially correlated, as reported in geotechnical literature (e.g., Cai et al. 2017; Xiao et al. 2017, 2018). Many models are available to describe such a spatial correlation, such as random field model (e.g., Li et al. 2016b; Xiao et al. 2016; Gong et al. 2017; Cao et al. 2019; Zhou et al. 2019) and Bayesian compressive sampling (e.g., Zhao et al. 2018, 2020; Zhao and Wang 2019). The independence assumption is still applied in Eq. (8) not only for simplicity but also because it can capture the randomly scattered soil seams embedded in a soil unit by exploring each data point individually. This assumption was found acceptable in previous soil classification studies (e.g., Wang et al. 2013, 2019a).

During the calculation of Eq. (8), the determination of soil types in the neighborhood of each location is needed, as indicated in Eq. (2), because what is obtained in the second step is the posterior probability of each soil type rather than the soil type itself. This can be achieved by random simulation of soil type from the estimated posterior probability, which is referred to as the simulated field method and whose performance has been validated by Celeux et al. (2003) and Forbes and Peyrard (2003). When the joint likelihood of Eq. (8) is calculated based on a simulated field, the Markov chain Monte Carlo method (e.g., Li 2009) can be used to obtain a sequence of samples of β from the joint likelihood and to gradually maximize the joint likelihood. The Markov chain Monte Carlo simulation moves the estimate of β from one sample to another, corresponding to the updating of β in the third step. Once β converges, the other model parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ converge as well.

Soil property evaluation with noise filtering

As highlighted previously, the lateral transition of soil type from the CPTU location to the borehole location contributes to a more reliable soil classification in the presence of noisy CPTU data. The coupled machine learning method can be regarded as unsupervised learning in terms of noise filtering, as there is no need to judge whether a data point is noise or not during the soil classification. Once the soil classification is finished, the identification of noisy data is straightforward, and they can be filtered to improve the reliability of soil property evaluation.

In practice, soil property evaluation is to establish a correlation between a soil parameter measured from laboratory tests (e.g., undrained shear strength s_u) and a CPTU index (e.g., Q_t) (e.g., Mayne 2007; Zou et al. 2017; Knuuti and Länsivaara 2019). This is usually achieved by linking the engineering behavior of a soil segment (i.e., borehole sampling interval) based on laboratory tests to the averaged CPTU data at the adjacent soundings with the same depth. Due to the discrepancy of soil types at CPTU and borehole locations, a reasonable average of CPTU data should be performed within the dominant soil type. In this study, if the probability that soil types at CPTU and borehole locations are consistent is greater than 0.5 [i.e., $p(x_s=w_s|y_s, w_s)\geq 0.5$ obtained in previous soil classification], the soil layer is considered to be dominant; otherwise, it is taken as a soil seam (i.e., noisy data) and the corresponding CPTU data shall not be related to soil properties. As shown in Fig. 3, there are three relations of soil types at borehole and CPTU locations:

(1) Completely related (Fig. 3a): no soil seam is involved. All CPTU data points in the segment can be averaged to establish the correlation;
(2) Partially related (Fig. 3b): a thin soil seam is involved, and only a portion of the CPTU data points in the segment agree with the soil type reported by borehole logs. Only those agreeable to borehole logs shall be averaged in the correlation analysis; and

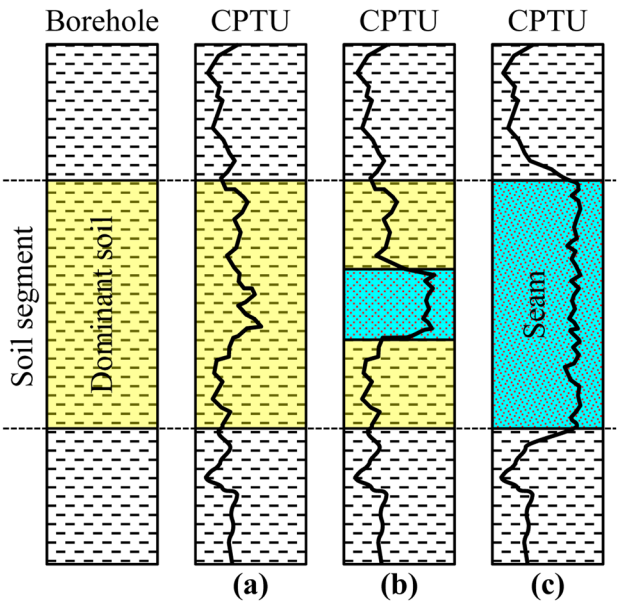

Fig. 3 Relation of soil types at borehole and CPTU locations: **a** completely related; **b** partially related; **c** none related

(3) None related (Fig. 3c): the soil segment is entirely a seam from the perspective of CPTU. All CPTU data in the segment should be abandoned.

By this means, the impact of soil seams on soil property evaluation can be eliminated. A case study of the Hong Kong-Zhuhai-Macao Bridge will be investigated in the next section to validate the proposed coupled machine learning method in both soil classification and soil property evaluation.

Case study of the Hong Kong-Zhuhai-Macao Bridge

The Hong Kong-Zhuhai-Macao Bridge (Fig. 4), straddling the Pearl River Estuary of China and comprising a 12 km Hong Kong Link Road, a 29.6 km Main Bridge (including a 6 km immersed tunnel along with two artificial transition islands) and a 13.4 km Zhuhai Link Road, is the longest bridge-cum-tunnel sea crossing in the world. The immersed tunnel is the most challenging civil work of the Hong Kong-Zhuhai-Macao Bridge, which can be referred to Hu et al. (2015) for more details. A detailed site investigation, including a large number of boreholes, CPTU soundings, and laboratory tests, was conducted around the immersed tunnel to characterize the soil strata below the seabed and evaluate the engineering properties of soils.

Geological conditions and site characterization scheme

The study site is along a strip section of artificial island-immersed tunnel-artificial island. It has a generic lithostratigraphy of marine deposits-continental sediments-marine

Fig. 4 The Hong Kong-Zhuhai-Macao Bridge and the study site

deposits-alluvial sediments, analogous to deep offshore deposits in Hong Kong (over 50 m at Chek Lap Kok, the Hong Kong International Airport), due to the marine transgressions in geologic history (e.g., Lee and Ng 1999; Xia et al. 2013). The seabed level varies from 7 to 23 m beneath the sea level. The thickness of Quaternary deposits is generally over 60 m at the site. In the design, the buried depth of the immersed tunnel varies in a range of 0 ~ 30 m beneath the seabed.

A set of 71 offshore boreholes were sunk, and 406 CPTU soundings were carried out along the immersed tunnel by the Fourth Harbour Engineering Investigation and Design Institute of the China Communications Construction Company (CCCC-FHDI Engineering Co., LTD), as shown in Fig. 5. Among these, 36 boreholes have detailed records of soil profiles with sampling intervals ranging from 0.1 to 1 m, and each has one adjacent CPTU sounding for verification (i.e., CPTU No. 1 ~ 36, black circles in Fig. 5). These 36 pairs of boreholes and CPTU soundings will be used to train the coupled machine learning model. Then, the enhanced soil classification result will be extended to the rest 370 CPTU soundings. The lateral distance between each borehole and its paring CPTU sounding was controlled within 6 m. Although such a small distance ensures most borehole data can indicate the soil types at their comparative CPTU sounding locations, inconsistency of soil types caused by soil seams is still inevitable. In addition to boreholes and CPTU soundings, 14 isotropic and 30 anisotropic triaxial compression tests were performed on undisturbed clayey soil samples in the laboratory to evaluate the undrained shear strength ratio s_u/σ'_{v0}, where σ'_{v0} is the effective overburden stress.

Soil classification and property evaluation ignoring noises

All soil samples from boreholes were classified and some of them were tested according to BS 5930 (BSI 2010). The site mainly involves six soil types (i.e., $K = 6$): 1 CLAY, 2 sandy CLAY, 3 SILT and sandy SILT, 4 clayey SAND, 5 SAND, and 6 gravelly SAND. Note that SILT is merged into sandy SILT as the corresponding samples are too few to support meaningful statistical analysis. A representative borehole log and its comparative CPTU profile, including cone tip resistance (q_t), sleeve frictional resistance (f_s), pore water pressure (u_2), Q_t, and F_r, are shown in Fig. 6. The borehole log indicates that a major soil interface, namely a clayey soil layer overlaying on a sandy soil layer, exists at a depth of about 30 m. The CPTU data may have abrupt variations at locations where the transition of soil type occurs because the mechanical behavior of soil changes with its material type.

Figure 7 compares the soil types reported by borehole and their adjacent CPTU data on a widely used CPTU-based soil classification chart — Robertson chart (Robertson 1990). Most CPTU data agree with borehole-reported soil types and distribute within the corresponding Robertson chart zones. For example, in the first subfigure of Fig. 7, most CPTU data fall in zones 3 and 4, corresponding to clay to silty clay and clayey silt to silty clay, respectively, when the borehole samples indicate a clayey soil. Nevertheless, it is also apparent that all soil types involve many noisy data points on the soil classification chart. These noisy data are difficult to separate from the dominant data based on subjective engineering judgments as their boundaries are not distinct.

With respect to the soil property evaluation, Fig. 8 illustrates the correlation between s_u/σ'_{v0} and Q_t for the study site, in which s_u/σ'_{v0} is measured from 44 triaxial compression tests and Q_t is a set of CPTU data from the corresponding 44 segments of soil specimen. The mean and the range of one standard deviation of Q_t for each segment are illustrated in Fig. 8. In practice, a cone factor $N_{kt} = Q_t/(s_u/\sigma'_{v0})$ is widely used to link the cone tip resistance and undrained shear strength. As shown in Fig. 8, most s_u/σ'_{v0}–Q_t data points follow a general trend with $N_{kt} = 10$ for the anisotropic tests and $N_{kt} = 13$ for the isotropic tests. These two values are

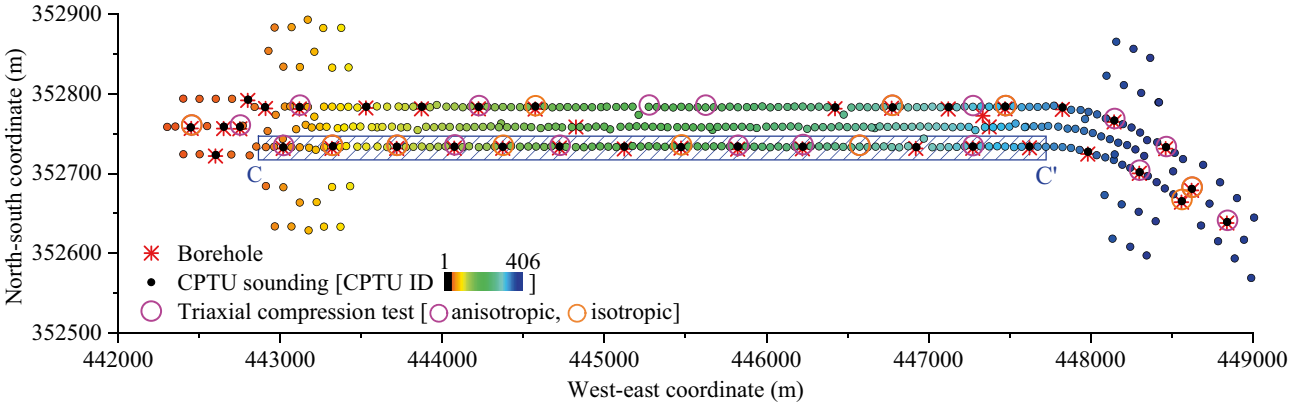

Fig. 5 Layout of CPTU soundings, boreholes, and triaxial compression tests at the study site

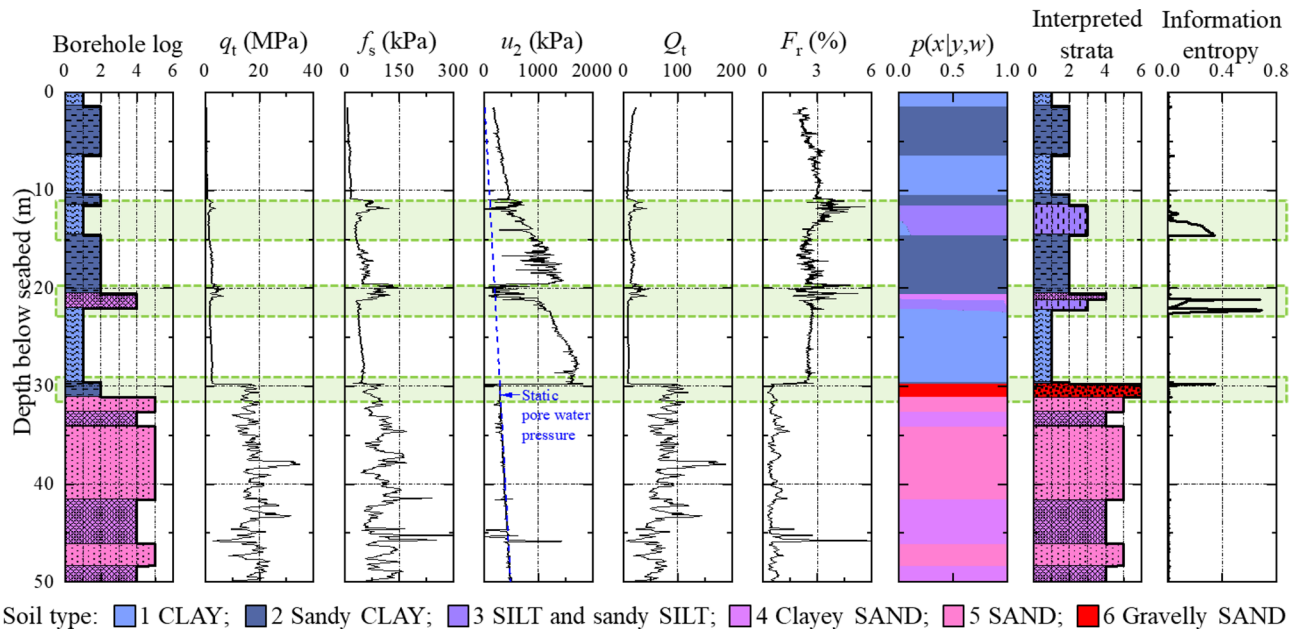

Fig. 6 Soil strata from borehole and CPTU at one representative location

consistent with existing engineering practice that N_{kt} generally ranges from 5 to 29 (e.g., Zou et al. 2017). However, two segments (i.e., A and B) are highly biased against the general trend and associated with significant variabilities. They are likely to be induced by the discrepancy of soil types at the borehole and CPTU locations and will lead to an unreasonable overestimation of N_{kt} in soil property evaluation.

Training of the coupled machine learning model

To enhance the interpretation of site investigation data, the coupled machine learning method is applied to identify and separate the noisy data from the dominant data. Recall that only 36 CPTU soundings adjacent to boreholes are utilized to train the model. A practical setting that takes borehole-reported soil types as the authentic soil types at adjacent CPTU locations is adopted to evaluate the initial model parameters. By this means, the number of authentic soil types can be determined as $L=K=6$, and the name of authentic soil types can inherit from the borehole-reported soil types, although this is not a mandatory option in the proposed method.

As six soil types are considered, the model parameters $\{\beta, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ consist of 67 elements in total, including 1, 12, 12, 6, and 36 elements for β, μ, σ, ρ, and $p(w|x)$, respectively. Figure 9 presents the variation of these elements during 3000 iterations. Parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and most elements in $\boldsymbol{\pi}$ [only except $p(w=1|x=3)$] converge fast within only a few iterations, probably because they are insensitive to the minor change of soil types and the adopted closed-form solutions of $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\}$ are relatively steady. By contrast, the convergence of β requires at least 500 iterations, as it is closely related to the fluctuations of soil type near the soil interfaces. Besides, when one soil type is often mixed with another, such as 1 CLAY and 3 SILT and sandy SILT, high uncertainties are associated with soil transitions, and the corresponding transition probability [e.g., $p(w=1|x=3)$] converges slowly as well.

Eventually, only the last 2500 samples after convergence, including both model parameters and simulated realizations of authentic soil types, are used in the subsequent analysis. The mean value of β is 3.94, and the mean estimates of $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and $\boldsymbol{\pi}$ are summarized in Tables 1 and 2, respectively. As the soil type changes from clayey soil to sandy soil, the mean values of F_r decrease, while the mean values of Q_t increase. In addition, negative correlations exist between $\log F_r$ and $\log Q_t$ for all six soil types, and the transition probability matrix $\boldsymbol{\pi}$ is approximately diagonally dominant (except 3 SILT and sandy SILT), indicating that these soils have a high probability of remaining consistent types between CPTU soundings and boreholes. All these observations agree well with engineering experience.

Enhanced soil classification and profiling

Based on the trained coupled machine learning model, the dominant CPTU data (light dots) and noisy data (dark dots) of different soil types can be identified, as shown on the Robertson chart in Fig. 10. The estimated contours of the probability distribution for dominant data clusters (solid

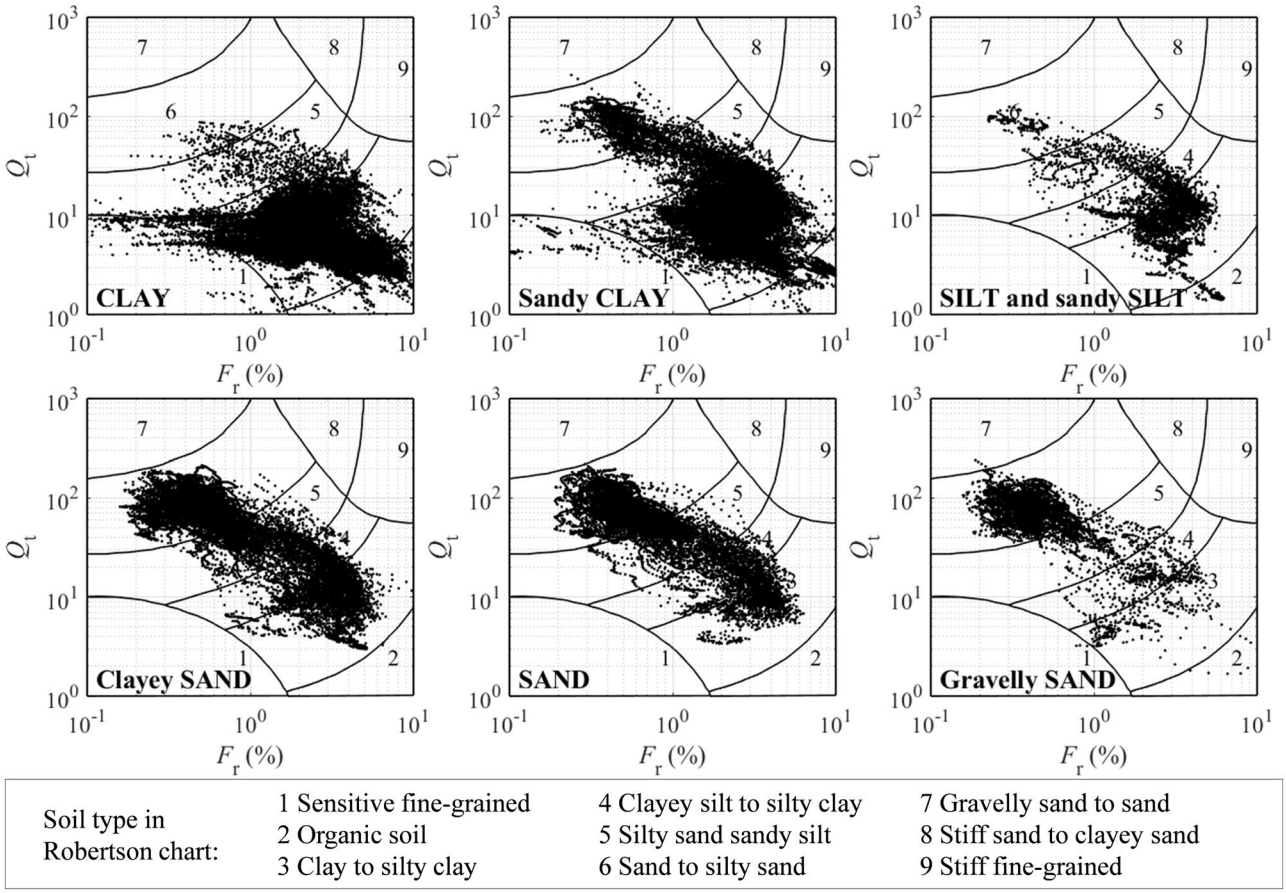

Fig. 7 CPTU data on the Robertson chart compared with borehole-reported soil types

lines) and the contours of the initial probability distribution of all CPTU data (dash lines) are also given in Fig. 10 for reference. The probability distribution contours are ellipses in shape as a bivariate normal distribution is used to describe

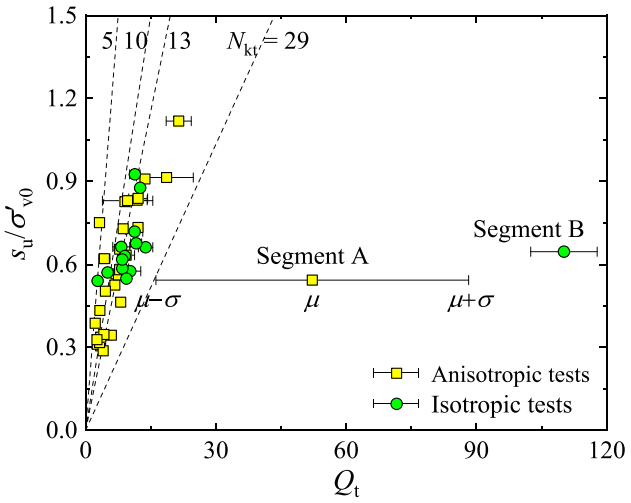

Fig. 8 Correlation between Q_t and s_u/σ'_{v0} without noise filtering

the CPTU data, and a smaller diameter indicates a higher probability density. The identified dominant data clusters have centers close to those of original data but with much smaller dispersions (i.e., smaller uncertainties in soil classification). Those noisy data points far away from the centers of dominant data clusters are successfully identified and separated. For example, many CPTU data points of SAND and gravelly SAND are in zones 3 and 4 of the Robertson chart, implying that these data shall correspond to clays and silts with low strength (Q_t) and high friction ratio (F_r); the coupled machine learning method automatically classifies these suspicious data points into CLAY, sandy CLAY, or SILT and sandy SILT. The noise filtering can significantly enhance the interpretation of CPTU data in soil classification and make it more compatible with the engineering practice. This benefits from the permission of lateral soil type transition from the CPTU location to the borehole location.

Recall the representative borehole log shown in Fig. 6. The posterior probabilities of authentic soil type, the most probable interpreted strata, and the corresponding information entropy, estimated by the coupled machine learning method, are also provided for comparison. Most enhanced

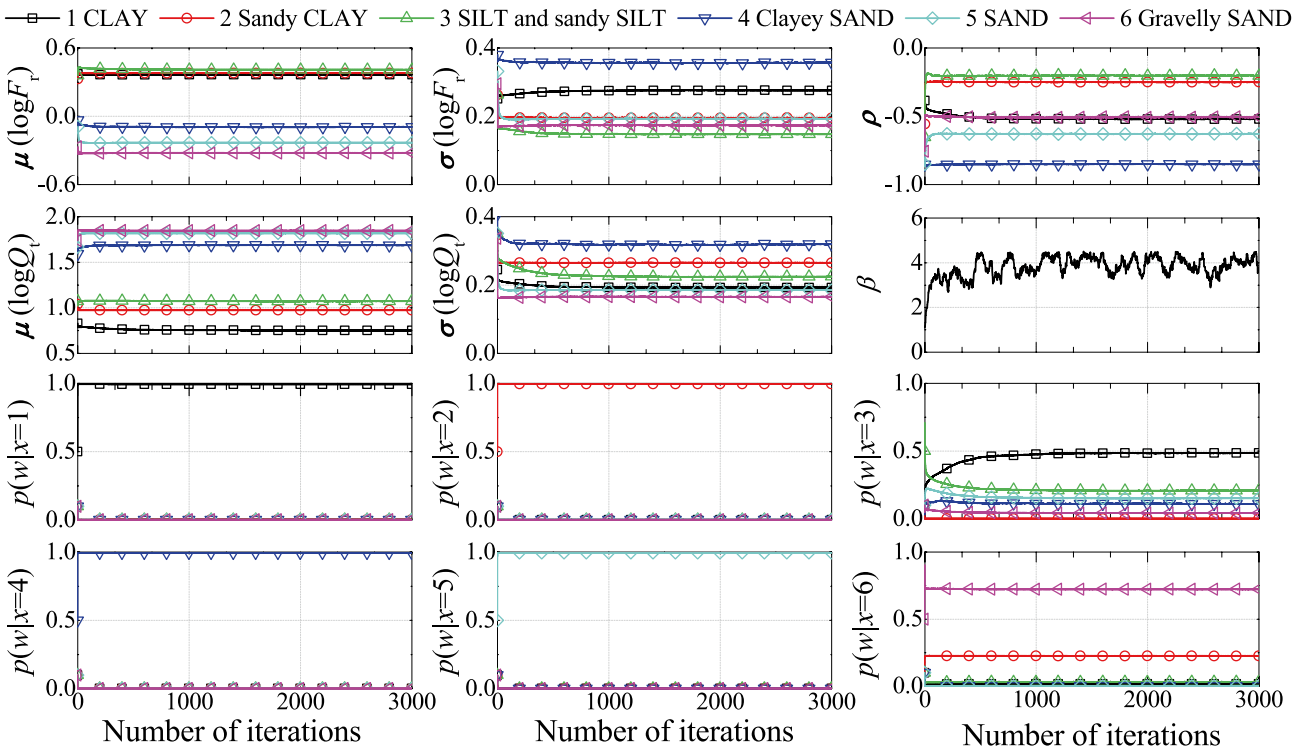

Fig. 9 Model parameter estimation in the coupled machine learning method

soil classification results agree favorably with the borehole log. Three soil seams can be identified at depths of around 13, 21, and 30 m. The information entropy also indicates that significant uncertainties exist at these three locations. This observation is interesting since the noisy data are determined by comparing the probability that soil types at the CPTU and borehole locations are consistent, rather than the degree of information entropy. The results demonstrate that the information entropy could be treated as an essential indicator of soil seam when there is no verification borehole adjacent to CPTU sounding.

Figure 11 profiles soil types along the C–C′ section in Fig. 5 from west to east at locations of boreholes, in which the left and right columns are the classification results of CPTU data using the coupled machine learning method and borehole logs, respectively. Again, they are comparable in general, but the CPTU-based soil profiles tend to contain more thin layers than the borehole logs. This is reasonable because the CPTU is more continuously sampled than the borehole and therefore has a much higher capacity to identify the embedded soil seams. The coupled machine learning method can correct the variation of soil boundaries over the distance between the borehole and the pairing CPTU, such as those dash segments in Fig. 11, and produce more reliable results of soil profiling.

Furthermore, the enhanced soil classification can be easily extended to all 406 CPTU soundings over the whole study site, as illustrated in Fig. 12. The first 36 CPTU soundings adjacent to the verification boreholes are taken as the learning set, while the rest 370 CPTU soundings are organized from west to east spatially, as shown in Fig. 5. For the rest of CPTU soundings, the coupled machine learning model retrogresses to the conventional hidden Markov random field model without borehole information. The missing data and end of sounding in CPTU data are colored in white in Fig. 12a and b. As demonstrated in the most probable soil type (Fig. 12c), a clayey soil layer with a thickness of 25 ~ 40 m overlays on a sandy soil layer, thin in the west (about 25 m) and thick in the east (about 40 m). At the eastern artificial island (i.e., the farthest east side), the

Table 1 Estimated mean, standard deviation, and correlation coefficient of CPTU data

Soil type	Mean, μ		Standard deviation, σ		Correlation coefficient, ρ
	$\log F_r$	$\log Q_t$	$\log F_r$	$\log Q_t$	
1. CLAY	0.37	0.75	0.28	0.19	−0.52
2. Sandy CLAY	0.38	0.97	0.20	0.26	−0.25
3. SILT and sandy SILT	0.41	1.07	0.15	0.22	−0.20
4. Clayey SAND	−0.09	1.69	0.36	0.32	−0.85
5. SAND	−0.23	1.82	0.19	0.18	−0.63
6. Gravelly SAND	−0.32	1.85	0.17	0.17	−0.51

Table 2 Estimated transition probability matrix from x to w

Authentic soil type	Borehole-reported soil type					
	$w=1$	$w=2$	$w=3$	$w=4$	$w=5$	$w=6$
$x=1$	0.99	0.01	0	0	0	0
$x=2$	0	1.00	0	0	0	0
$x=3$	0.49	0	0.21	0.11	0.15	0.04
$x=4$	0	0	0	1.00	0	0
$x=5$	0	0	0	0	1.00	0
$x=6$	0.02	0.23	0.03	0	0	0.72

sandy soil layer almost vanishes. Regarding the information entropy (Fig. 12d), the stratification uncertainties of the learning CPTU soundings are much less than those of other soundings, benefiting from the information integrated from the verification boreholes. The stratification uncertainties at locations where no data is available are the largest, followed by the interfaces between two soil types, particularly between two similar soil types, such as SAND and gravelly SAND, and CLAY and sandy CLAY.

Enhanced soil property evaluation

After identifying noisy CPTU data in soil classification, the site-specific s_u/σ'_{v0}–Q_t correlation analysis of soil property evaluation can be enhanced correspondingly. Recall that Fig. 8 indicates that CPTU data in segments A and B are suspicious. By comparing the soil types between borehole and CPTU obtained from the enhanced soil classification, it is found that segments A and B contain many noisy data

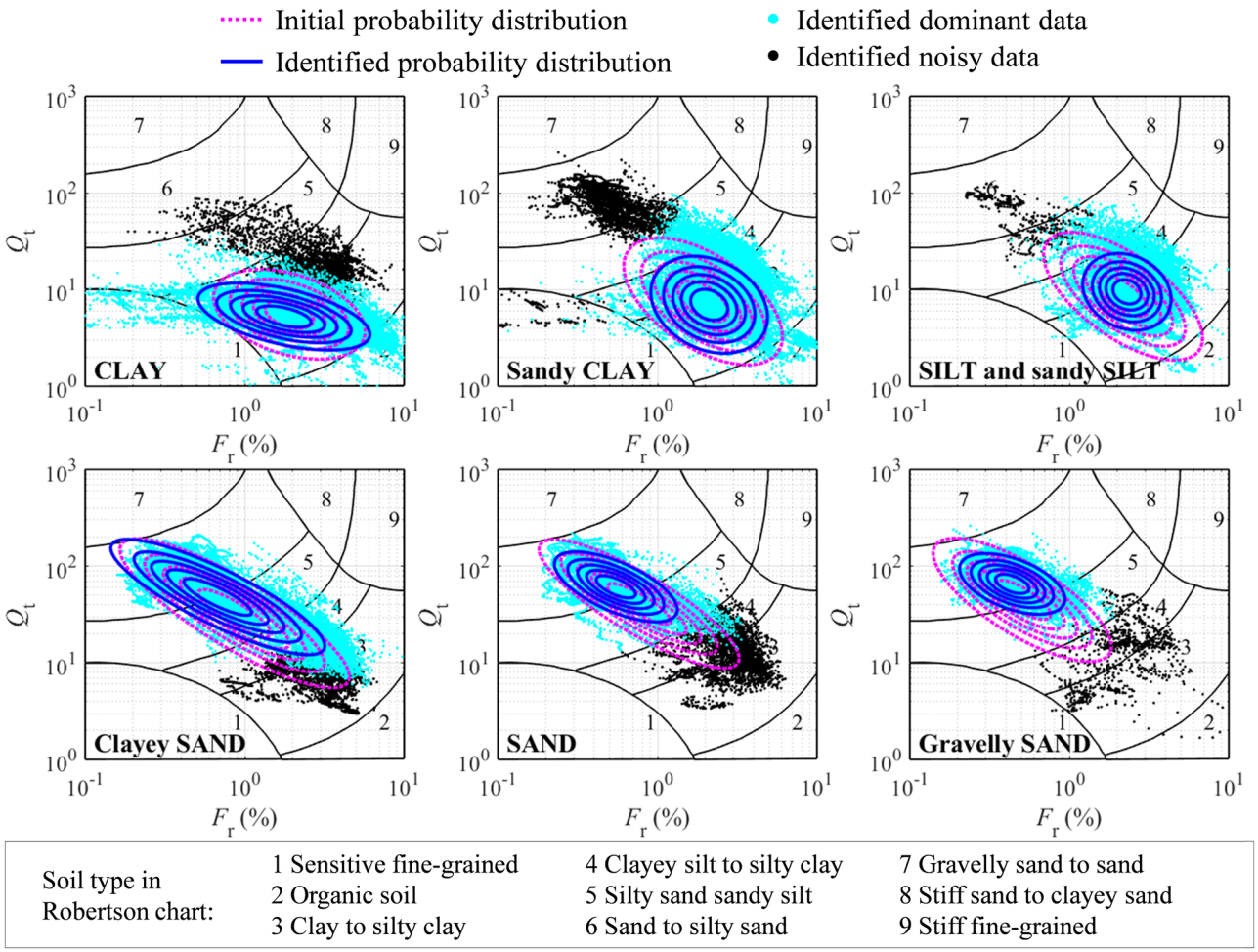

Fig. 10 Enhanced soil classification using the coupled machine learning method

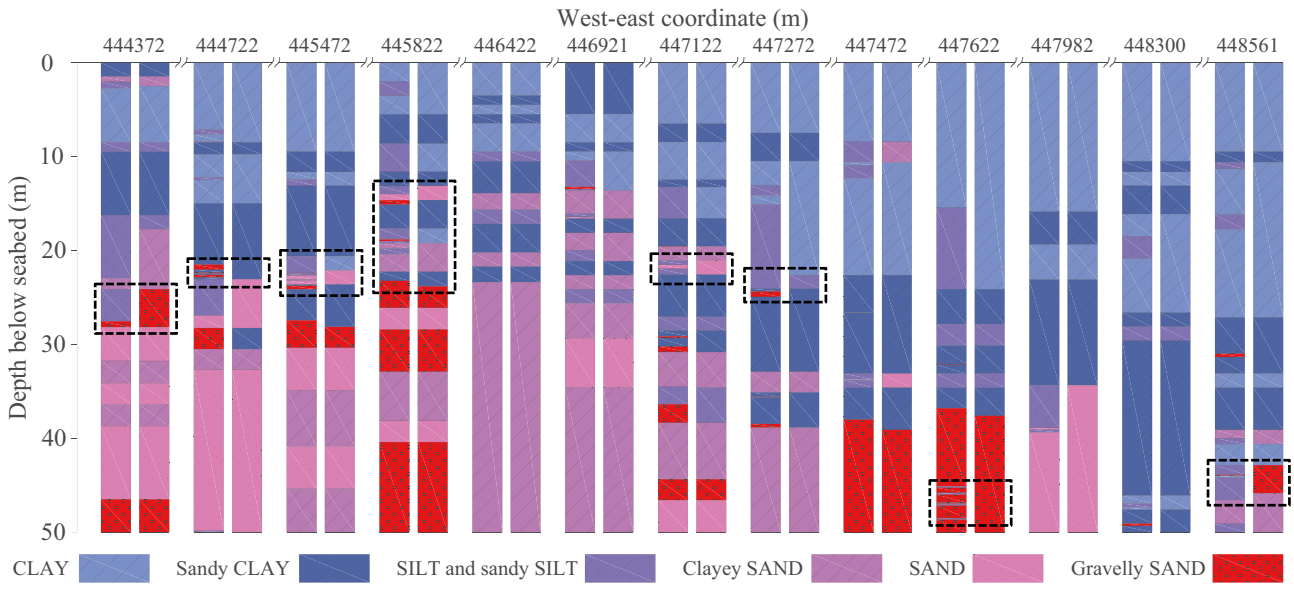

Fig. 11 Profiling soil types along the C–C′ section (Fig. 5) using the coupled machine learning method (the left and right columns are the classification results of CPTU data and borehole logs, respectively)

indeed. Specifically, segment A is classified as a partially related case (Fig. 3b), and segment B is a none related case (Fig. 3c), while other segments all belong to the completely related case (Fig. 3a). This explains why the variability of segment A is much greater than other segments, including segment B.

For segment B, the authentic soil type identified by the coupled machine learning method is SAND (Fig. 13a). Therefore, it shall not be included in the s_u/σ'_{v0}–Q_t correlation analysis because its CPTU data do not represent the behavior of cohesive soils. Regarding segment A, some data points are classified as sandy CLAY, whereas others are SAND, according to the proposed method. Through partial averaging of those sandy CLAY data, the s_u/σ'_{v0}–Q_t data points of segment A fall in exactly the common range of $N_{kt} = 5 \sim 29$, with a much smaller variability, as shown in Fig. 13b. The impact of noisy CPTU data on the s_u/σ'_{v0}–Q_t correlation is significantly alleviated. The

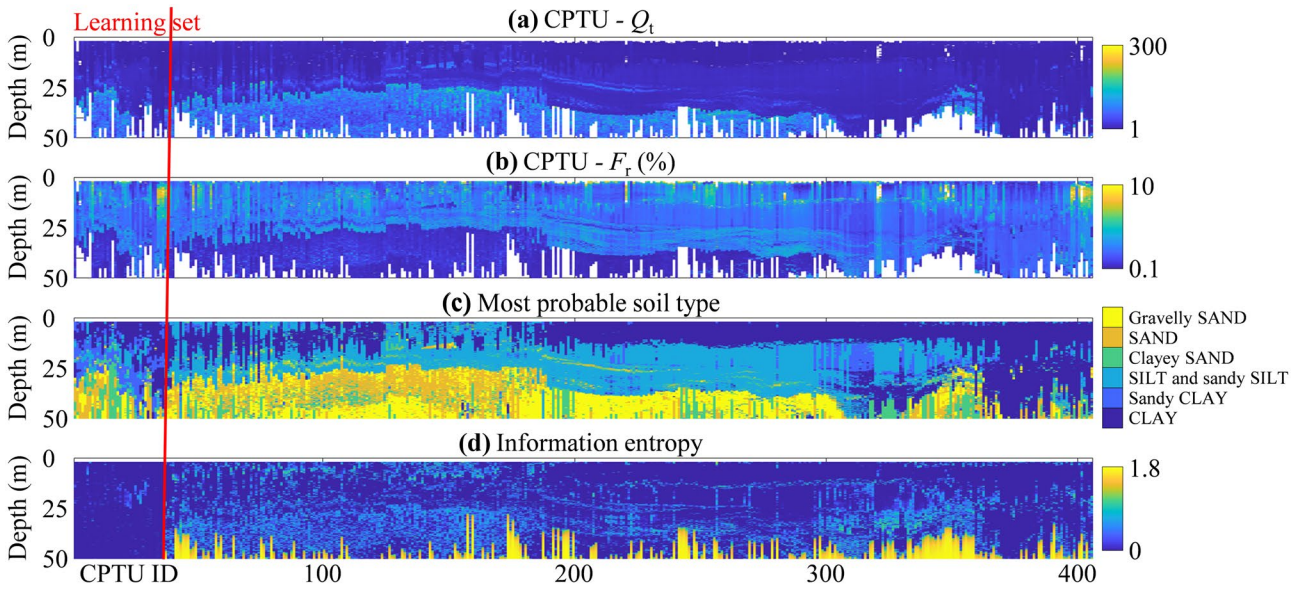

Fig. 12 Enhanced soil classification for all CPTU soundings: **a** Q_t; **b** F_r; **c** most probable soil type; **d** information entropy

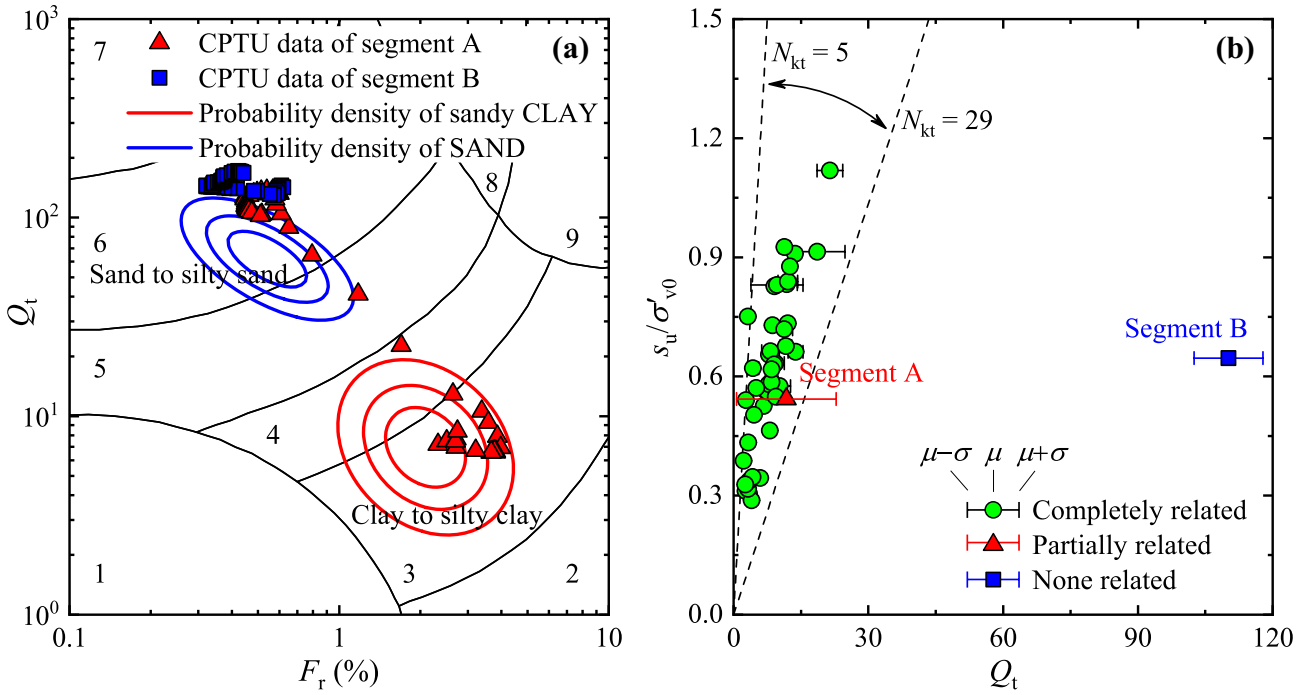

Fig. 13 Correlation between Q_t and s_u/σ'_{v0} with noise filtering of segments A and B: **a** enhanced soil classification; **b** corrected correlation

noise filtering obtains a better interpretation of CPTU data and simultaneously reduces the uncertainty of s_u/σ'_{v0}–Q_t correlation.

Summary and conclusions

This study proposes a coupled machine learning method for soil type classification and soil property evaluation to achieve a more reliable interpretation of site characterization data by integrating borehole and CPTU data with noise filtering. On the one hand, the borehole and CPTU data are systematically integrated under a rigorous Bayesian framework, in which the borehole-reported soil type and CPTU data are treated as two types of evidence of the authentic soil type. On the other hand, the noisy CPTU data are identified and separated from the dominant CPTU data, which is achieved by allowing a lateral transition of soil type from the CPTU location to the borehole location to capture the discrepancy of soil types.

A detailed marine site characterization around the 6 km immersed tunnel of the Hong Kong-Zhuhai-Macao Bridge was investigated. The performance of the coupled machine learning method in CPTU-based soil classification was validated against the widely used Robertson chart. The soil seams embedded in the dominant soil strata can be successfully detected without subjective judgment so that more reliable soil classification and profiling can be produced. Regarding the soil profile of the study site, a clayey soil layer with a thickness of around 25 ~ 40 m is found to overlay on a sandy soil layer, and the thickness generally increases from west to east. The integration of borehole and CPTU data significantly reduces the stratification uncertainty in site characterization. Meanwhile, the site-specific correlation between undrained shear strength ratio and CPTU index in soil property evaluation becomes more compatible with engineering practice after filtering the identified noisy CPTU data.

Acknowledgements This work was supported by Eunsung O&C Offshore Marine and Construction (Project No. EUNSUNG19EG01) and the Science and Technology Plan of Shenzhen, China (Project No. JCYJ20180507183854827).

Declarations

Competing interests The authors declare no competing interests.

References

Bishop CM (2006) Pattern recognition and machine learning. Springer-Verlag, Berlin, Heidelberg

Boulanger RW, DeJong JT (2018) Inverse filtering procedure to correct cone penetration data for thin-layer and transition effects. In: Proc of the 4th Int'l Symp on Cone Penetration Testing (CPT'18), p 25–44

British Standards Institution (BSI) (2010) BS 5930:1999+A2:2010 - code of practice for site investigations. British Standards Institution, London

Cai GJ, Zou HF, Liu SY, Puppala AJ (2017) Random field characterization of CPTU soil behavior type index of Jiangsu quaternary soil deposits. Bull Eng Geol Environ 76:353–369. https://doi.org/10.1007/s10064-016-0854-x

Cao ZJ, Zheng S, Li DQ, Phoon KK (2019) Bayesian identification of soil stratigraphy based on soil behaviour type index. Can Geotech J 56(4):570–586. https://doi.org/10.1139/cgj-2017-0714

Celeux G, Forbes F, Peyrard N (2003) EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognit 36:131–144. https://doi.org/10.1016/S0031-3203(02)00027-4

Ching JY, Wang JS, Juang CH, Ku CS (2015) Cone penetration test (CPT)-based stratigraphic profiling using the wavelet transform modulus maxima method. Can Geotech J 52(12):1993–2007. https://doi.org/10.1139/cgj-2015-0027

Depina I, Le TMH, Eiksund G, Strøm P (2016) Cone penetration data classification with Bayesian Mixture Analysis. Georisk 10(1):27–41. https://doi.org/10.1080/17499518.2015.1072637

Du Y, Zhu L, Zou H, Zhang L, Cai G, Liu S (2020) Evaluation of CPTU-based soil classification charts for offshore sediments in Pearl River Delta, China. In: Geo-Congress 2020: Modeling, Geomaterials, and Site Characterization (GSP 317), p 663–639. https://doi.org/10.1061/9780784482803.067

Elfeki AMM, Dekking FM (2005) Modelling subsurface heterogeneity by coupled Markov chains: directional dependency, Walther's law and entropy. Geotech Geol Eng 23(6):721–756. https://doi.org/10.1007/s10706-004-2899-z

Forbes F, Peyrard N (2003) Hidden Markov random field model selection criteria based on mean field-like approximations. IEEE Trans Pattern Anal Mach Intell 25(9):1089–1101. https://doi.org/10.1109/TPAMI.2003.1227985

Gong W, Tien YM, Juang CH, Martin JR, Luo Z (2017) Optimization of site investigation program for improved statistical characterization of geotechnical property based on random field theory. Bull Eng Geol Environ 76(3):1021–1035. https://doi.org/10.1007/s10064-016-0869-3

Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2):85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Hu ZN, Xie YL, Wang J (2015) Challenges and strategies involved in designing and constructing a 6 km immersed tunnel: a case study of the Hong Kong-Zhuhai-Macao Bridge. Tunn Undergr Sp Tech 50:171–177. https://doi.org/10.1016/j.tust.2015.07.011

Knuuti M, Länsivaara T (2019) Variation of CPTu-based transformation models for undrained shear strength of Finnish clays. Georisk 13(4):262–270. https://doi.org/10.1080/17499518.2019.1644525

Lee KM, Ng PCC (1999) A geotechnical investigation of marine deposits in a nearshore seabed for land reclamation. Can Geotech J 36:981–1000. https://doi.org/10.1139/t99-063

Li A, Jafari NH, Tsai FTC (2020) Modelling and comparing 3-D soil stratigraphy using subsurface borings and cone penetrometer tests in coastal Louisiana, USA. Georisk 14(2):158–176. https://doi.org/10.1080/17499518.2019.1637528

Li DQ, Qi XH, Cao ZJ, Tang XS, Phoon KK, Zhou CB (2016a) Evaluating slope stability uncertainty using coupled Markov chain. Comput Geotech 73:72–82. https://doi.org/10.1016/j.compgeo.2015.11.021

Li DQ, Xiao T, Cao ZJ, Zhou CB, Zhang LM (2016b) Enhancement of random finite element method in reliability analysis and risk assessment of soil slopes using Subset Simulation. Landslides 13(2):293–303. https://doi.org/10.1007/s10346-015-0569-2

Li JH, Cassidy MJ, Huang J, Zhang LM, Kelly R (2016c) Probabilistic Identification of Soil Stratification Géotechnique 66(1):16–26. https://doi.org/10.1680/jgeot.14.P.242

Li SZ (2009) Markov random field modeling in image analysis. Springer Science & Business Media, London

Lunne T, Robertson PK, Powell JJM (1997) Cone penetration testing in geotechnical practice. Blackie Academic and Professional, London

Mayne PW (2007) Cone penetration testing: a synthesis of highway practice. NCHRP Synthesis 368, Transportation Research Board, Washington, D.C.

Mo P, Marshall AM, Yu H (2017) Interpretation of cone penetration test data in layered soils using cavity expansion analysis. J Geotech Geoenviron Eng 143(1):04016084. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001577

Phoon KK, Kulhawy FH (1999) Characterization of geotechnical variability. Can Geotech J 36(4):612–624. https://doi.org/10.1139/t99-038

Phoon KK (2020) The story of statistics in geotechnical engineering. Georisk 14(1):3–25. https://doi.org/10.1080/17499518.2019.1700423

Robertson PK (1990) Soil classification using the cone penetration test. Can Geotech J 27(1):151–158. https://doi.org/10.1139/t90-014

Robertson PK (2009) Interpretation of cone penetration tests - a unified approach. Can Geotech J 46(11):1337–1355. https://doi.org/10.1139/T09-065

Schneider JA, Randolph MF, Mayne PW, Ramsey NR (2008) Analysis of factors influencing soil classification using normalizing piezocone tip resistance and pore pressure parameters. J Geotech Geoenviron Eng 134(11):1569–1586. https://doi.org/10.1061/(ASCE)1090-0241(2008)134:11(1569)

Wang H, Wang X, Wellmann F, Liang RY (2019a) A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data. Can Geotech J 56(8):1184–1205. https://doi.org/10.1139/cgj-2017-0709

Wang X, Wang H, Liang RY, Liu X (2019b) A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data. Eng Geol 248:102–116. https://doi.org/10.1016/j.enggeo.2018.11.014

Wang Y, Huang K, Cao Z (2013) Probabilistic identification of underground soil stratification using cone penetration tests. Can Geotech J 50(7):766–776. https://doi.org/10.1139/cgj-2013-0004

Xia Z, Jia P, Ma S, Liang K, Shi Y, Waniek JJ (2013) Sedimentation in the Lingdingyang Bay, Pearl River Estuary. Southern China J Coastal Res 66(SP1):12–24. https://doi.org/10.2112/SI_66_2

Xiao T, Li DQ, Cao ZJ, Au SK, Phoon KK (2016) Three-dimensional slope reliability and risk assessment using auxiliary random finite element method. Comput Geotech 79:146–158. https://doi.org/10.1016/j.compgeo.2016.05.024

Xiao T, Li DQ, Cao ZJ, Zhang LM (2018) CPT-based probabilistic characterization of three-dimensional spatial variability using MLE. J Geotech Geoenviron Eng 144(5):04018023. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001875

Xiao T, Zhang LM, Li XY, Li DQ (2017) Probabilistic stratification modeling in geotechnical site characterization. ASCE-ASME J Risk Uncertain Eng Syst – A Civ Eng 3(4):04017019. https://doi.org/10.1061/AJRUA6.0000924

Yahşi BK, Ersoy H (2021) Effect of mineralogical composition related to profile depth on index and strength properties of regolith soil. Bull Eng Geol Environ 80(2):1791–1808. https://doi.org/10.1007/s10064-020-01968-8

Yin KS, Zhang LM, Wang HJ, Zou HF, Li JH (2021) Marine soil behaviour classification using piezocone penetration test (CPTu) and borehole records. Can Geotech J 58(2):190–199. https://doi.org/10.1139/cgj-2019-0571

Yuen KV, Mu HQ (2012) A novel probabilistic method for robust parametric identification and outlier detection. Probabilistic Eng Mech 30:48–59. https://doi.org/10.1016/j.probengmech.2012.06.002

Zhao T, Hu Y, Wang Y (2018) Statistical interpretation of spatially varying 2D data from sparse measurements using Bayesian

compressive sampling. Eng Geol 246:162–175. https://doi.org/10.1016/j.enggeo.2018.09.022

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

Zhao T, Xu L, Wang Y (2020) Fast non-parametric simulation of 2D multi-layer cone penetration test (CPT) data without pre-stratification using Markov Chain Monte Carlo simulation. Eng Geol 273:105670. https://doi.org/10.1016/j.enggeo.2020.105670

Zheng S, Zhu YX, Li DQ, Cao ZJ, Deng QX, Phoon KK (2021) Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning. Geosci Front 12(1):425–439. https://doi.org/10.1016/j.gsf.2020.03.017

Zhou XP, Zhu BZ, Juang CH, Wong LNY (2019) A stability analysis of a layered-soil slope based on random field. Bull Eng Geol Environ 78(4):2611–2625. https://doi.org/10.1007/s10064-018-1266-x

Zou HF, Liu SY, Cai GJ, Puppala AJ, Bheemasetti T (2017) Multivariate correlation analysis of seismic piezocone penetration (SCPTU) parameters and design properties of Jiangsu quaternary cohesive soils. Eng Geol 228:11–38. https://doi.org/10.1016/j.enggeo.2017.07.005