Acta Geotechnica
https://doi.org/10.1007/s11440-025-02732-6

RESEARCH PAPER

Cone penetration test (CPT)-based soil classification and stratification with consideration of data cross-correlation and noises

Jun-Cheng Yao[1] · Yu Wang[2] · Zheng Guan[3] · Kostas Senetakis[1]

Received: 8 March 2025 / Accepted: 14 July 2025

Abstract
In geotechnical site investigation, Robertson's soil behavior type (SBT) chart is widely used for soil classification based on two quantities measured during a cone penetration test (CPT), the normalized cone resistance Q_t and the normalized friction ratio F_R. Q_t and F_R are negatively correlated and provide complementary information for soil classification. However, this cross-correlation between Q_t and F_R has not been explicitly modelled in previous studies of subsurface soil classification and stratification using an often-limited number of CPT soundings from a specific site. This study aims to leverage such cross-correlation for improving CPT-based stratification and zonation by a joint sparse representation of Q_t and F_R in a vertical cross-section, as well as quantifying their uncertainty under a Bayesian framework. In addition, direct application of the SBT chart to a vertical cross-section often leads to noisy results (e.g., SBTs fluctuate rapidly and unrealistically within short distances). The noises are subsequently removed mainly by subjective engineering judgment in current practices. In this study, a randomization of input measurements is proposed to filter out the noise and improve computational efficiency simultaneously. Both simulated and real data examples are used to illustrate the proposed method. The results indicate that the proposed method significantly improves accuracy of the soil classification and stratification and automatically removes the noise.

Keywords Bayesian compressive sensing · Data cross-correlation · Joint sparse representation · Noise removal · Soil classification and stratification · Uncertainty quantification

1 Introduction

Soil classification and stratification (or zonation) (e.g., obtain the spatial distribution of different soil types) are essential in geotechnical site investigation before geotechnical design and construction of tunnels, foundations, pipelines, and other infrastructure systems [6, 18, 26, 31, 41]. Because of various geological processes, such as erosion, transportation, weathering, and metamorphic processes [8, 16], underground soil profiles often contain multiple geological layers or units with different physical and mechanical properties; very often the variabilities in stratification make it difficult to develop reliable ground models especially when the ground formations involve complex mechanisms of chemical weathering or because of seasonal influences altering periodically the mechanical properties of the strata. Cone penetration test (CPT) is used widely for identifying such layers or units, as it can provide near-continuous

✉ Yu Wang
wang.yu@ust.hk

Jun-Cheng Yao
jcyao2-c@my.cityu.edu.hk

Zheng Guan
zhengguan@tudelft.nl

Kostas Senetakis
ksenetak@cityu.edu.hk

[1] Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

[2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[3] Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Published online: 16 August 2025

information on physical and mechanical properties of soils and pore water pressure at the tested location along depth [9, 37, 42]. To classify soil types based on CPT data, Robertson [35] developed a soil behavior type (SBT) classification chart, as shown in Fig. 1a, to identify in-situ soil types based on two normalized CPT parameters, i.e., the normalized cone resistance $Q_t = (q_t - \sigma_{v0})/\sigma'_{v0}$, and the normalized friction ratio $F_R = 100 f_s/(q_t - \sigma_{v0})$, where f_s, q_t, σ_{v0}, and σ'_{v0} are the sleeve friction, corrected cone tip resistance, vertical total stress, and vertical effective stress, respectively. The Robertson chart in Fig. 1a is divided into nine zones for nine distinct soil behavior types (SBT1 to SBT9). Table 1 summarizes soil descriptions for these nine soil types [36]. In addition, the SBT index (I_c), calculated as $I_c = \sqrt{(3.47 - \log Q_t)^2 + (\log F_R + 1.22)^2}$, has been widely used as a surrogate of Q_t and F_R [23], and it provides a reasonable fit to the boundaries in the center of the Robertson chart [22, 36]. According to the I_c values, six soil behavior types, i.e., SBT2 to SBT7, can be classified as summarized in Table 1.

Consider, for example, Fig. 1b that shows a set of real CPT data obtained from an experimental site in the South Parklands in South Australia [20]. The variations of Q_t and F_R with depth are shown in the first column of Fig. 1b, and the soil types are determined using the Robertson chart and shown in the second column of Fig. 1b. It is observed that the measured Q_t and F_R are highly correlated, with a correlation coefficient calculated as –0.74 in this example. The cross-correlation between Q_t and F_R is of practical significance as it can provide complementary information for soil classification and stratification.

Although the CPT soundings can provide near-continuous information in the vertical direction, their horizontal spacing is usually large, e.g., 10–100 m, due to limited

Table 1 Description of soil behavior types in Robertson's soil behavior type (SBT) classification chart (modified from [35, 36])

Area	Soil description	Range of SBT index I_c
1	Sensitive, fine-grained	N/A
2	Organic soils (peats)	$I_c > 3.60$
3	Clays (clay to silty clay)	$2.95 < I_c < 3.60$
4	Silt mixtures (clayey silt to silty clay)	$2.60 < I_c < 2.95$
5	Sand mixtures (silty sand to sandy silt)	$2.05 < I_c < 2.60$
6	Sands (clean sand to silty sand)	$1.31 < I_c < 2.05$
7	Gravelly sand to sand	$I_c < 1.31$
8	Very stiff sand to clayey sand	N/A
9	Very stiff, fine-grained	N/A

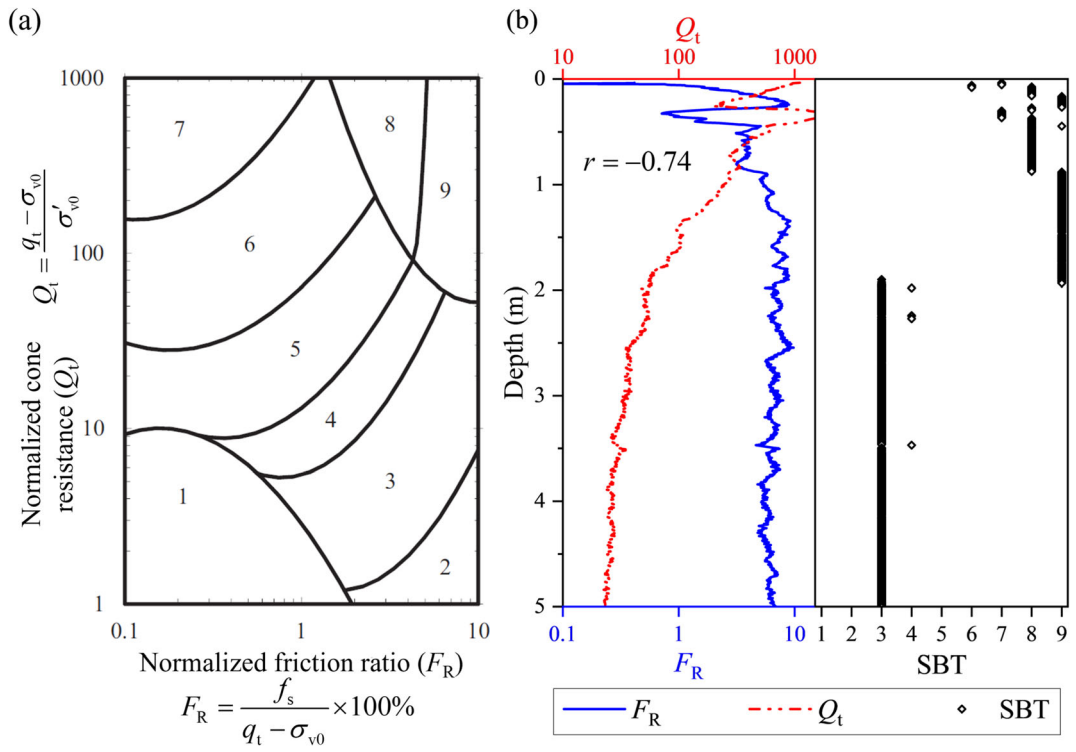


Fig. 1 An illustrative example of soil classification using Robertson's soil behavior type (SBT) classification chart: **a** Robertson's soil behavior type classification chart (modified from [35]); and **b** a set of real CPT results at the South Parklands site in Australia (data from [20]) with soil classification results from Robertson's SBT chart

budget, time, or site constraints [34]. It is, therefore, necessary to interpret information of soil properties and identify the soil type at the untested location in a specific site. That is, CPT-based soil classification and stratification in a two-dimensional (2D) vertical cross-section often have to be inferred from a limited number of one-dimensional (1D) profiles of CPT data [17, 48, 50, 53]. Since the information along the horizontal direction is limited with epistemic (statistical) uncertainties [34], the strong cross-correlation between Q_t and F_R (e.g., see Fig. 1b) might offer additional valuable information for improving data interpretation along the horizontal direction [13] and CPT-based soil classification and stratification, although such a cross-correlation has not been modelled in previous studies.

In addition, the vertical resolution of Q_t and F_R along depth can be as small as 0.005–0.05 m [37], with many high-frequency variations caused by noise along depth (e.g., see Fig. 1b). If the CPT data with noises (e.g., see Fig. 1b) is directly used to identify soil type and strata, many soil types or layers may be identified within a short depth range. For example, the soil profile with a depth of about 5 m below the ground surface in Fig. 1b is divided into about 16 soil layers, which is not practical in engineering applications. In current practice, because of no soil sample retrieved from CPT for visual inspection, subjective engineering judgment is often exercised to manually smooth the noisy results into a soil stratification with only a limited number of soil layers [34, 46]. As different engineers might have different engineering judgments, the soil strata identified by different practitioners may be different. This underscores a need of an objective and automatic method to deal with noisy CPT data and improve CPT-based soil classification and stratification.

To model CPT data cross-correlation and noises, this study proposes a joint sparse representation of Q_t and F_R for improving CPT-based soil classification and stratification. A randomization of input measurements is also proposed to remove noises and improve computational efficiency simultaneously. Following this introduction, a framework of the proposed method is introduced in Sect. 2, and the detailed procedure is presented step by step in Sect. 3. Then, a simulated example is used to illustrate the proposed method in Sect. 4. A real data example is also provided in Sect. 5, followed by concluding remarks in Sect. 6.

2 Proposed method based on joint sparse representation of CPT data

Underground soils are formed by multiple geological, physical, and chemical processes, leading to notable spatial patterns and variability [33]. Hence, geotechnical data, e.g., Q_t and F_R from a CPT, are usually auto-correlated spatially, enabling a sparse representation in an appropriate coordinate system [44, 45, 52, 54, 55]. In other words, when the datasets or signals are auto-correlated, they are compressible and can be represented by a relatively small number of significant components [5, 40]. The basic idea of sparse representation is that an auto-correlated signal, or CPT data in this study, can be represented as a linear combination of many basis functions, e.g., discrete cosine transform (DCT) functions, and only a limited number of basis functions are non-trivial or important [19, 28, 49]. Mathematically, a set of geotechnical data within a 2D cross-section, defined by a matrix $\mathbf{F}$ with a size of $N_1 \times N_2$, can be expressed as [47]:

$$\mathbf{F} = \sum_{t=1}^{N} \omega_t \mathbf{B}_t^{2D} \tag{1}$$

where $N = N_1 \times N_2$; $\mathbf{B}_t^{2D}$ is the t^{th} 2D basis function with the same size as $\mathbf{F}$, i.e., $N_1 \times N_2$; and ω_t is the weight coefficient corresponding to $\mathbf{B}_t^{2D}$. $\mathbf{B}_t^{2D}$ may be constructed from two 1D basis function matrices $\mathbf{B}^1$ and $\mathbf{B}^2$, which is independent of $\mathbf{F}$. $\mathbf{B}^1$ and $\mathbf{B}^2$ has a dimension of $N_1 \times N_1$ and $N_2 \times N_2$, respectively. In this study, different columns of $\mathbf{B}^1$ and $\mathbf{B}^2$ represent the orthonormal discrete cosine functions with different frequencies. Under a sparse representation of $\mathbf{F}$, only a relatively small number of ω_t are non-zero or important.

As illustrated in the Introduction, Q_t and F_R data from CPTs exhibit high cross-correlation, due to the physical and mechanical properties of the soil. Therefore, they may share a common spatial pattern [13, 51], especially within the same soil layer. The inherent auto-correlated and cross-correlated structure in Q_t and F_R make it possible to have a joint sparse representation in an appropriate normalized coordinate system [2, 27]. Under a joint sparse representation of Q_t and F_R, each normalized dataset is represented by a sum of two components [1, 12]: (1) a common component with a shared spatial feature for both Q_t and F_R, and (2) an individual component with unique spatial features for Q_t or F_R, specifically. Subsequently, the inherent auto-correlated and cross-correlated structure of Q_t and F_R can be modelled simultaneously by a joint sparse representation. Note that, when modelling the auto- and cross-correlated structure by joint sparse representation, no hyper-parameter, e.g., trend function or horizontal correlation length, needs to be modelled from limited measurements.

This is different from parametric methods like Gaussian Process Regression (GPR), where accurate estimation of such parameter is often a significant challenge due to horizontal data sparsity in geotechnical site investigation [56].

To properly model the cross-correlated structure, a normalization of Q_t and F_R to a common scale is required before the joint sparse representation. In addition, since the Q_t and F_R are negatively correlated (see Fig. 1b), their spatial patterns are opposite to each other. On the other hand, the common component under a joint sparse representation framework shall represent positive cross-correlation between two variables considered (e.g., Q_t and F_R in this study). To satisfy the requirement of positive cross-correlation in the joint sparse representation, a negative sign shall be added to the F_R data for converting its negative correlation with Q_t data into a positive correlation. After that, the normalized datasets of Q_t and F_R within a 2D cross-section, denoted as $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$, can be jointly represented as:

$$\begin{aligned}\mathbf{F}'_{Q_t} &= \mathbf{F}^C + \mathbf{F}^U_{Q_t} = \sum_{t=1}^{N} \omega_t^C \mathbf{B}_t^{2D} + \sum_{t=1}^{N} \omega_t^{U_1} \mathbf{B}_t^{2D} \\ \mathbf{F}'_{F_R} &= \mathbf{F}^C + \mathbf{F}^U_{F_R} = \sum_{t=1}^{N} \omega_t^C \mathbf{B}_t^{2D} + \sum_{t=1}^{N} \omega_t^{U_2} \mathbf{B}_t^{2D}\end{aligned} \tag{2}$$

where ω_t^C is the t^{th} weight coefficient for common component $\mathbf{F}^C$; $\omega_t^{U_1}$ and $\omega_t^{U_2}$ are the t^{th} weight coefficient for individual component $\mathbf{F}^U_{Q_t}$ and $\mathbf{F}^U_{F_R}$, respectively. Under a joint sparse representation framework, most of the weight coefficients are close to zero. Therefore, the important weight coefficients can be estimated using measurements from a limited number of CPT soundings. The normalized measurements of Q_t and F_R, denoted respectively as $\mathbf{Y}'_{Q_t}$ and $\mathbf{Y}'_{F_R}$ with a size of $M_1 \times M_2$, can be determined by their locations of measurements in $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$:

$$\begin{aligned}\mathbf{Y}'_{Q_t} &= \boldsymbol{\Psi}_1 \mathbf{F}'_{Q_t} \boldsymbol{\Psi}_2 = \sum_{t=1}^{N} \omega_t^C \boldsymbol{\Psi}_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 + \sum_{t=1}^{N} \omega_t^{U_1} \boldsymbol{\Psi}_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 \\ \mathbf{Y}'_{F_R} &= \boldsymbol{\Psi}_1 \mathbf{F}'_{F_R} \boldsymbol{\Psi}_2 = \sum_{t=1}^{N} \omega_t^C \boldsymbol{\Psi}_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 + \sum_{t=1}^{N} \omega_t^{U_2} \boldsymbol{\Psi}_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2\end{aligned} \tag{3}$$

where $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are problem-specific measurement matrices, with a dimension of $M_1 \times N_1$ and $N_2 \times M_2$, respectively. Because each pair of Q_t and F_R data is measured by CPT at the same location, measurement matrices $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are the same for both Q_t and F_R. Since $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ reflect the locations of elements of $\mathbf{Y}'_{Q_t}$ and $\mathbf{Y}'_{F_R}$ in $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$ along the column and row directions, respectively, each measurement matrix can be constructed from an identity matrix based on the measurement locations [54].

2.1 Randomization of input measurements

Noisy results of CPT-based soil classification and stratification typically exhibit high-frequency variations along the vertical direction (e.g., Fig. 1b). A basic idea of removing the noise is to map the CPT data from the original space to a lower-dimension space. A random matrix (e.g., Gaussian or Bernoulli random matrix), can directly sense the projected CPT data in its original domain, thereby maintaining sensing efficiency without requiring additional prior knowledge [7, 25]. The high-frequency noises might be filtered out automatically during this process. Therefore, a random matrix is used in this study to compress the input measurements along the vertical direction (i.e., dimension of N_1 in this study). This is the so-called randomization of input measurements in data analytics literature [7]. It is important to emphasize that this randomization does not involve down-sampling or dropping some measurements directly, but uses a random linear combination of measurements with all measured information preserved. Simultaneously, the basis function matrix in Eq. (3) is compressed using the same random matrix. Note that such a randomization might also improve computational efficiency by reducing volume of the input data.

To realize the randomization of input measurements, the choice of a good random matrix is essential for joint sparse representation of Q_t and F_R. Gaussian random projections are often chosen for representation due to their favorable mathematical properties and the richness of information extracted, minimizing the risk of losing crucial information of the original data [4]. The normalized measurements $\mathbf{Y}'_{Q_t}$ and $\mathbf{Y}'_{F_R}$ are compressed from a vertical length of M_1 to M_c (e.g., $M_c = 101$) using a Gaussian random matrix in this study. The compressed measurements, denoted as $\mathbf{Y}'^{,c}_{Q_t}$ and $\mathbf{Y}'^{,c}_{F_R}$ with a size of $M_c \times M_2$, can be expressed as follows:

$$\begin{aligned}\mathbf{Y}'^{,c}_{Q_t} &= \Phi \mathbf{Y}'_{Q_t} = \sum_{t=1}^{N} \omega_t^C \Phi \Psi_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 + \sum_{t=1}^{N} \omega_t^{U_1} \Phi \Psi_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 \\ \mathbf{Y}'^{,c}_{F_R} &= \Phi \mathbf{Y}'_{F_R} = \sum_{t=1}^{N} \omega_t^C \Phi \Psi_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2 + \sum_{t=1}^{N} \omega_t^{U_2} \Phi \Psi_1 \mathbf{B}_t^{2D} \boldsymbol{\Psi}_2\end{aligned} \tag{4}$$

where Φ is a Gaussian random matrix with a size of $M_c \times M_1$. Each element in Gaussian random matrix follows a random Gaussian distribution with a zero mean and variance equal to $1/M_c$. For derivation convenience, $\mathbf{Y}'^{,c}_{Q_t}$ and $\mathbf{Y}'^{,c}_{F_R}$ are vectorized by stacking sequentially the columns of

$\mathbf{Y}'^{,c}_{Q_t}$ and $\mathbf{Y}'^{,c}_{F_R}$. This leads to two column vectors $\boldsymbol{y}'^{,c}_{Q_t}$ and $\boldsymbol{y}'^{,c}_{F_R}$, both with a length of $M = (M_c \times M_2)$:

$$\begin{aligned} \boldsymbol{y}'^{,c}_{Q_t} &= \mathbf{A}^R\boldsymbol{\omega}^C + \mathbf{A}^R\boldsymbol{\omega}^{U_1} \\ \boldsymbol{y}'^{,c}_{F_R} &= \mathbf{A}^R\boldsymbol{\omega}^C + \mathbf{A}^R\boldsymbol{\omega}^{U_2} \end{aligned} \tag{5}$$

where the sensing matrix $\mathbf{A}^R$ is a random sensing matrix with a size of $M \times N$. Each column of $\mathbf{A}^R$, denoted as $\boldsymbol{\alpha}^R_t$, is obtained from a vectorization of $\Phi\Psi_1\mathbf{B}^{2D}_t\Psi_2$, i.e. $\boldsymbol{\alpha}^R_t = vec(\Phi\Psi_1\mathbf{B}^{2D}_t\Psi_2)$ (t = 1, 2, …, N). $\boldsymbol{\omega}^C = [\omega^C_1, \omega^C_2, ..., \omega^C_N]^T$ is the weight coefficient vector for the common component. $\boldsymbol{\omega}^{U_1} = [\omega^{U_1}_1, \omega^{U_1}_2, ..., \omega^{U_1}_N]^T$ and $\boldsymbol{\omega}^{U_2} = [\omega^{U_2}_1, \omega^{U_2}_2, ..., \omega^{U_2}_N]^T$ are the weight coefficient vectors for the individual components of Q_t and F_R, respectively. Then, Eq. (5) can be further re-written as:

$$\boldsymbol{y}^e = \begin{bmatrix} \boldsymbol{y}'^{,c}_{Q_t} \\ \boldsymbol{y}'^{,c}_{F_R} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^R & \mathbf{A}^R & \mathbf{0} \\ \mathbf{A}^R & \mathbf{0} & \mathbf{A}^R \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}^C \\ \boldsymbol{\omega}^{U_1} \\ \boldsymbol{\omega}^{U_2} \end{bmatrix} \tag{6}$$

Equation (6) can be further simplified as:

$$\boldsymbol{y}^e = \mathbf{A}^e\boldsymbol{\omega}^e \tag{7}$$

where $\boldsymbol{y}^e = \begin{bmatrix} \boldsymbol{y}'^{,c}_{Q_t} \\ \boldsymbol{y}'^{,c}_{F_R} \end{bmatrix}$, $\mathbf{A}^e = \begin{bmatrix} \mathbf{A}^R & \mathbf{A}^R & \mathbf{0} \\ \mathbf{A}^R & \mathbf{0} & \mathbf{A}^R \end{bmatrix}$, and $\boldsymbol{\omega}^e = \begin{bmatrix} \boldsymbol{\omega}^C \\ \boldsymbol{\omega}^{U_1} \\ \boldsymbol{\omega}^{U_2} \end{bmatrix}$.

Since correlated Q_t and F_R usually have a sparse representation, most of the elements in $\boldsymbol{\omega}^e$ are close to zero. Therefore, when important weights in $\boldsymbol{\omega}^e$ are probabilistically estimated from compressed measurement ensemble, i.e., $\boldsymbol{y}^e$, the Q_t and F_R in the target cross-section can be reconstructed jointly by Eq. (2).

2.2 Joint Bayesian reconstruction of CPT data

Because the CPT soundings are limited along horizontal directions, substantial statistical uncertainty might be induced when estimating $\boldsymbol{\omega}^e$ and reconstructing $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$, particularly along horizontal directions. To quantify the uncertainty in an approximation of $\boldsymbol{\omega}^e$, denoted as $\hat{\boldsymbol{\omega}}^e$, a Bayesian framework is used for the joint estimation of $\hat{\boldsymbol{\omega}}^e$. Under a Bayesian framework, the compressed measurement ensemble, $\boldsymbol{y}^e$, can be used to update the distribution of $\hat{\boldsymbol{\omega}}^e$, with the expression as follows [3]:

$$p(\hat{\boldsymbol{\omega}}^e|\boldsymbol{y}^e) = \frac{p(\boldsymbol{y}^e|\hat{\boldsymbol{\omega}}^e) \times p(\hat{\boldsymbol{\omega}}^e)}{p(\boldsymbol{y}^e)} \tag{8}$$

where $p(\hat{\boldsymbol{\omega}}^e)$ represents the prior probability distribution function (PDF) of $\hat{\boldsymbol{\omega}}^e$, reflecting the prior knowledge about $\hat{\boldsymbol{\omega}}^e$; $p(\boldsymbol{y}^e|\hat{\boldsymbol{\omega}}^e)$ represents the likelihood function, reflecting the probability of observing the $\boldsymbol{y}^e$ given $\hat{\boldsymbol{\omega}}^e$; $p(\boldsymbol{y}^e)$ represents the normalizing constant ensuring an integration of the posterior $p(\hat{\boldsymbol{\omega}}^e|\boldsymbol{y}^e)$ to be one. It is seen from Eq. (8) that the likelihood function $p(\boldsymbol{y}^e|\hat{\boldsymbol{\omega}}^e)$ and the prior PDF $p(\hat{\boldsymbol{\omega}}^e)$ are two essential ingredients of the Bayesian framework. A zero-mean Gaussian measurement error with unknown variance, σ^2, is adopted for each measurement, leading to the Gaussian likelihood function [43]:

$$p(\boldsymbol{y}^e|\hat{\boldsymbol{\omega}}^e, \tau) = \left(\frac{\tau}{2\pi}\right)^{M_e/2} \exp\left(\frac{-\tau(\boldsymbol{y}^e - \mathbf{A}^e\hat{\boldsymbol{\omega}}^e)^T(\boldsymbol{y}^e - \mathbf{A}^e\hat{\boldsymbol{\omega}}^e)}{2}\right) \tag{9}$$

where τ represents the reciprocal of the unknown variance of σ^2 for modelling convenience, i.e., $\tau = \sigma^{-2}$; M_e is the total data point number in the measurement ensemble $\boldsymbol{y}^e$.

To promote the sparsity within the weight coefficient vector and facilitate identification of the important or non-trivial weight coefficients, a three-level hierarchical model is used to model the prior PDF of $\hat{\boldsymbol{\omega}}^e$. First, each element of $\hat{\boldsymbol{\omega}}^e$, i.e., $\hat{\omega}^e_t$, is taken to follow a Gaussian distribution with a mean of zero and unknown variance of ${\alpha_t}^{-1}$, t = 1, 2, …, $3N$, with an expression as [24]:

$$p(\hat{\boldsymbol{\omega}}^e|\boldsymbol{\alpha}) = \prod_{t=1}^{3N} \frac{\alpha_t^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\alpha_t(\hat{\omega}^e_t)^2}{2}\right) \tag{10}$$

where $\boldsymbol{\alpha}$ represents α_t expressed in a vector form, i.e., $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{3N}]$; $\mathbf{D}^\alpha$ is a diagonal matrix with diagonal element $D^\alpha(t,t) = \alpha_t$ (t = 1, 2, …, $3N$). Note that the prior distribution given in Eq. (10) is conjugate to the likelihood function given in Eq. (9). This suggests that the conditional posterior PDF to be derived later also follows a Gaussian distribution. Then, each element of $\boldsymbol{\alpha}$, i.e., α_t, is taken to follow an inverse Gamma distribution with a parameter $\gamma/2$ ($\gamma > 0$) and a constant parameter 1, leading to $p(\alpha_t \mid \gamma) = \gamma/2\ {\alpha_t}^{-2} \exp[-\gamma/2\ {\alpha_t}^{-1}]$. Then, the prior PDF of $\boldsymbol{\alpha}$ is expressed as:

$$p(\boldsymbol{\alpha}|\boldsymbol{\gamma}) = \prod_{t=1}^{3N} p(\alpha_t|\gamma) = \prod_{t=1}^{3N} \frac{\gamma}{2}\alpha_t^{-2} \exp[-\frac{\gamma}{2}\alpha_t^{-1}] \tag{11}$$

To avoid manual selection of γ values, the unknown parameter γ is also taken to follow a Gamma distribution, which is expressed as:

$$p(\gamma) = b_0^{a_0}\gamma^{a_0-1}\exp(-b_0\gamma)/\Gamma(a_0) \tag{12}$$

where a_0 and b_0 are non-negative constants. In addition, τ is taken to follow a Gamma distribution, and the prior PDF of τ is expressed as [24, 47]:

$$p(\tau) = \frac{d_0^{c_0}\tau^{c_0-1}}{\Gamma(c_0)}\exp(-d_0\tau) \tag{13}$$

where c_0 and d_0 are two non-negative parameters. According to Bayes' rule, the posterior distribution is derived as:

$$p(\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \tau|\boldsymbol{y}^e) = \frac{p(\boldsymbol{y}^e|\hat{\boldsymbol{\omega}}^e, \tau)p(\hat{\boldsymbol{\omega}}^e|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\gamma)p(\gamma)p(\tau)}{p(\boldsymbol{y}^e)} \quad (14)$$

Because of the term $p(\boldsymbol{y}^e)$ in Eq. (14), there is no analytical expression for the joint posterior PDF of $\hat{\boldsymbol{\omega}}^e$, $\boldsymbol{\alpha}$, τ and γ. It is therefore difficult to solve the above formulation analytically. To address this issue, a Markov Chain Monte Carlo (MCMC) simulation method called Gibbs sampling is used to simulate $\hat{\boldsymbol{\omega}}^e$ samples for estimating the posterior distribution of $\hat{\boldsymbol{\omega}}^e$ [10, 11]. Gibbs sampling requires analytical expressions of conditional PDFs, such as $p(\hat{\boldsymbol{\omega}}^e|\gamma, \tau, \boldsymbol{\alpha}, \boldsymbol{y}^e)$, $p(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}}^e, \gamma, \tau, \boldsymbol{y}^e)$, $p(\tau|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \boldsymbol{y}^e)$, $p(\gamma|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \tau, \boldsymbol{y}^e)$, which can be obtained from the above equations. Zhao et al. [54] showed that the posterior distribution of $\hat{\boldsymbol{\omega}}^e$, $p(\hat{\boldsymbol{\omega}}^e|\gamma, \tau, \boldsymbol{\alpha}, \boldsymbol{y}^e)$, follows a multivariate normal distribution with a mean $\boldsymbol{\mu}_{\hat{\boldsymbol{\omega}}^e}$ and a covariance $\mathbf{COV}_{\hat{\boldsymbol{\omega}}^e}$:

$$\begin{aligned} \boldsymbol{\mu}_{\hat{\boldsymbol{\omega}}^e} &= \mathbf{COV}_{\hat{\boldsymbol{\omega}}^e}(\mathbf{A}^e)^{\mathrm{T}}\boldsymbol{y}^e\tau \\ \mathbf{COV}_{\hat{\boldsymbol{\omega}}^e} &= [(\mathbf{A}^e)^{\mathrm{T}}\mathbf{A}^e\tau + \mathbf{D}^{\alpha}]^{-1} \end{aligned} \quad (15)$$

$p(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}}^e, \gamma, \tau, \boldsymbol{y}^e)$ follows a generalized inverse Gaussian distribution:

$$p(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}}^e, \gamma, \tau, \boldsymbol{y}^e) = \prod_{t=1}^{3N} \exp\left(-\frac{\alpha_t(\hat{\omega}_t^e)^2 + \alpha_t^{-1}\gamma}{2}\right) \alpha_t^{p-1} \frac{(\alpha_t/\gamma)^{p/2}}{2K_p\sqrt{\alpha_t\gamma}} \quad (16)$$

where K_p is a modified Bessel function of the second kind with parameter p of –1/2. $p(\tau|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \boldsymbol{y}^e)$ and $p(\gamma|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \tau, \boldsymbol{y}^e)$ follow two Gamma distributions:

$$\begin{aligned} p(\tau|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \boldsymbol{y}^e) &= Gamma(c_n, d_n) \\ p(\gamma|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \tau, \boldsymbol{y}^e) &= Gamma(\gamma_a, \gamma_b) \end{aligned} \quad (17)$$

where $c_n = M_e/2 + 1$; $d_n = d_0 + 1/2[(\boldsymbol{y}^e)^{\mathrm{T}}\boldsymbol{y}^e - 2(\hat{\boldsymbol{\omega}}^e)^{\mathrm{T}}(\mathbf{A}^e)^{\mathrm{T}}\boldsymbol{y}^e + (\hat{\boldsymbol{\omega}}^e)^{\mathrm{T}}(\mathbf{A}^e)^{\mathrm{T}}\mathbf{A}^e\hat{\boldsymbol{\omega}}^e]$; $\gamma_a = 3N + a_0$ and $\gamma_b = b_0 + \sum_{t=1}^{3N}\alpha_t^{-1}$. When a_0, b_0 and d_0 are taken as a small value (e.g., $a_0 = b_0 = d_0 = 10^{-4}$), a non-informative prior for $\hat{\boldsymbol{\omega}}^e$ is achieved, and the posterior distribution of $\hat{\boldsymbol{\omega}}^e$ will be driven by measurements. Using Eqs. (15)-(17), many (e.g., N_r = 200) cross-correlated RFSs of $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$ pair can be generated using Gibbs sampling. The procedure of Gibbs sampling is as shown in the following steps [54]:

(1) Initialize $\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\gamma}$ with some arbitrary values (e.g., $\alpha_t = 10^{-4}$, τ = 100/(variance of $\boldsymbol{y}^e$) and γ = 1) [43].
(2) Calculate $\boldsymbol{\mu}_{\hat{\boldsymbol{\omega}}^e}$ and $\mathbf{COV}_{\hat{\boldsymbol{\omega}}^e}$ using Eq. (15).
(3) Generate a sample of $\hat{\boldsymbol{\omega}}^e$ using a random field generator, such as Cholesky decomposition with calculated $\boldsymbol{\mu}_{\hat{\boldsymbol{\omega}}^e}$ and $\mathbf{COV}_{\hat{\boldsymbol{\omega}}^e}$.
(4) Update $p(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}}^e, \tau, \gamma, \boldsymbol{y}^e)$ using the generated sample of $\hat{\boldsymbol{\omega}}^e$ and γ.
(5) Sample $\boldsymbol{\alpha}$ from the updated $p(\boldsymbol{\alpha}|\hat{\boldsymbol{\omega}}^e, \gamma, \tau, \boldsymbol{y}^e)$.
(6) Update $p(\tau|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \boldsymbol{y}^e)$ and $p(\gamma|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \tau, \boldsymbol{y}^e)$ using the current sample of $\hat{\boldsymbol{\omega}}^e$ and $\boldsymbol{\alpha}$.
(7) Sample τ and γ from the updated $p(\tau|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \gamma, \boldsymbol{y}^e)$ and $p(\gamma|\hat{\boldsymbol{\omega}}^e, \boldsymbol{\alpha}, \tau, \boldsymbol{y}^e)$.
(8) Substitute $\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\gamma}$ in Step (1) with the current sample of $\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\gamma}$ and repeat Step (2) to Step (8) until prescribed number (e.g., N_r = 200) of $\hat{\boldsymbol{\omega}}^e$ samples are obtained. In MCMC simulation, a burn-in period is typically required, with an initial number of samples (e.g., the first 500) discarded. Furthermore, independent samples are usually preferred in MCMC simulation, by sampling $\hat{\boldsymbol{\omega}}^e$ at every n_b (e.g., n_b = 20) simulations.

In this study, N_r = 200 pairs of $\mathbf{F}'_{Q_t}$ and $\mathbf{F}'_{F_R}$ 2D random field samples (RFSs) can be simultaneously generated using Gibbs sampling. Since the generated RFS pairs of Q_t and F_R are in a normalized scale, a postprocessing step is required to convert the reconstructed Q_t and F_R back to the original data space.

2.3 Quantification of uncertainty and accuracy

In the proposed method, N_r pairs of the Q_t and F_R RFSs are sequentially generated for a target cross-section (i.e., with a size of $N_1 \times N_2$) with a target spatial resolution. Using each pair of Q_t and F_R data and the Robertson's SBT chart (e.g., Fig. 1a), the corresponding SBT is determined at each respective location or point within the 2D cross-section, leading to a probable SBT map sample. Repeating N_r times the above steps leads to N_r probable SBT map samples from N_r RFS pairs of the Q_t and F_R obtained using the joint sparse representation from limited CPT soundings.

In the target cross-section, each point has N_r probable SBT. The statistics of N_r probable SBT at a given point p with a coordinate (x_1, x_2) can be easily calculated. The probability of different specific SBT, i.e., $P(\mathrm{SBT}_p = s)$ (s = 1, 2, …, 9), can be computed based on the occurrence frequency of each SBT as following [17]:

$$P(\mathrm{SBT}_p = s) = \frac{N_p^s}{N_r} \times 100\% \quad (18)$$

where SBT_p is the SBT at point p; N_p^s is the number of map samples with SBT equal to s at point p. The most probable prediction (MPP) at one point p is the SBT with the highest frequency of occurrence. After repeating the above steps for all points, the MPP SBT map of the target cross-section

is obtained for soil stratification. In addition, the uncertainty of SBT at a data point p can be quantified by dispersion, i.e., $Dp(p)$, reflecting the percentage of non-matching SBT among N_r random samples in comparison to MPP [29, 38]:

$$Dp(p) = \frac{\sum_{k=1}^{N_r} I_k}{N_r},\ I_k = \begin{cases} 1\ \mathrm{SBT}_k(p) \neq \mathrm{SBT}_{\mathrm{MPP}}(p) \\ 0\ \mathrm{SBT}_k(p) = \mathrm{SBT}_{\mathrm{MPP}}(p) \end{cases} \tag{19}$$

where $\mathrm{SBT}_k(p)$ and $\mathrm{SBT}_{\mathrm{MPP}}(p)$ represent the SBT at point p of the k^{th} 2D random SBT samples and MPP, respectively. The dispersion value varies from 0 to 1, with a large value corresponding to a low prediction confidence level and vice versa. The spatial distribution of dispersion in a vertical cross-section of interest can be determined accordingly, and areas with large dispersion values often indicate boundary between strata with different soil types [38].

In addition, accuracy is used to evaluate the performance of the proposed method, when the ground true SBT map is available for the target vertical cross-section. The accuracy of the proposed method can be validated by comparing MPP with the ground true 2D SBT map, using the following expression [30, 39]:

$$\mathrm{Acc} = \frac{\sum_{i=1}^{N_1 \times N_2} I_i}{N_1 \times N_2},\ I_i = \begin{cases} 1\ \mathrm{SBT}_{\mathrm{T}}(p_i) = \mathrm{SBT}_{\mathrm{MPP}}(p_i) \\ 0\ \mathrm{SBT}_{\mathrm{T}}(p_i) \neq \mathrm{SBT}_{\mathrm{MPP}}(p_i) \end{cases} \tag{20}$$

where $N_1 \times N_2$ represents the total number of data points in the target cross-section; $\mathrm{SBT}_{\mathrm{T}}(p_i)$ and $\mathrm{SBT}_{\mathrm{MPP}}(p_i)$ denote the SBT at point p_i of the true 2D SBT map and MPP, respectively. A larger value of Acc in Eq. (20) corresponds to a higher accuracy of the proposed method. Acc can also be used to evaluate the prediction accuracy of individual CPT sounding by comparing MPP and true CPT sounding data observed at the same locations. Note that, in real practices, the ground true is unknown at the untested locations. The accuracy of the whole cross-section therefore cannot be calculated in real practices. It is used only for validation purposes in the illustrative examples.

3 Implementation procedure of the proposed method

Figure 2 shows the procedure of the proposed method for soil classification and stratification. The implementation of the proposed method is described below. Firstly, a number of 1D CPT data (i.e., Q_{t} and F_{R}) within a 2D vertical cross-section are collected and used as input data. Then, the input data is pre-processed to normalize Q_{t} and F_{R} and convert them into positively correlated data pairs. To transform the original measurement $\boldsymbol{y}_{Q_{\mathrm{t}}}$ and $\boldsymbol{y}_{F_{\mathrm{R}}}$ into a common scale, z-score normalization is used in this study. Consider, for example, the original measurements of Q_{t} and F_{R}, $\boldsymbol{y}_{Q_t} = [y_{Q_t,1}, y_{Q_t,2}, \ldots, y_{Q_t,M}]^{\mathrm{T}}$ and $\boldsymbol{y}_{F_{\mathrm{R}}} = [y_{F_{\mathrm{R}},1}, y_{F_{\mathrm{R}},2}, \ldots, y_{F_{\mathrm{R}},M}]^{\mathrm{T}}$. The j^{th}, $j = 1, 2, \ldots, M$, normalized Q_{t} and F_{R}, $y'_{Q_{\mathrm{t}},j}$ and $y'_{F_{\mathrm{R}},j}$ can be computed as $y'_{Q_{\mathrm{t}},j} = \frac{y_{Q_{\mathrm{t}},j} - \mu_{Q_{\mathrm{t}}}}{\sigma_{Q_{\mathrm{t}}}}$ and $y'_{F_{\mathrm{R}},j} = -\frac{y_{F_{\mathrm{R}},j} - \mu_{F_{\mathrm{R}}}}{\sigma_{F_{\mathrm{R}}}}$, where $\mu_{Q_{\mathrm{t}}}$ and $\mu_{F_{\mathrm{R}}}$ are the mean of Q_{t} and F_{R} measurements; $\sigma_{Q_{\mathrm{t}}}$ and $\sigma_{F_{\mathrm{R}}}$ are the standard deviation of Q_{t} and F_{R} measurements. A negative sign is also added to F_{R} to convert negative correlation to positive correlation.

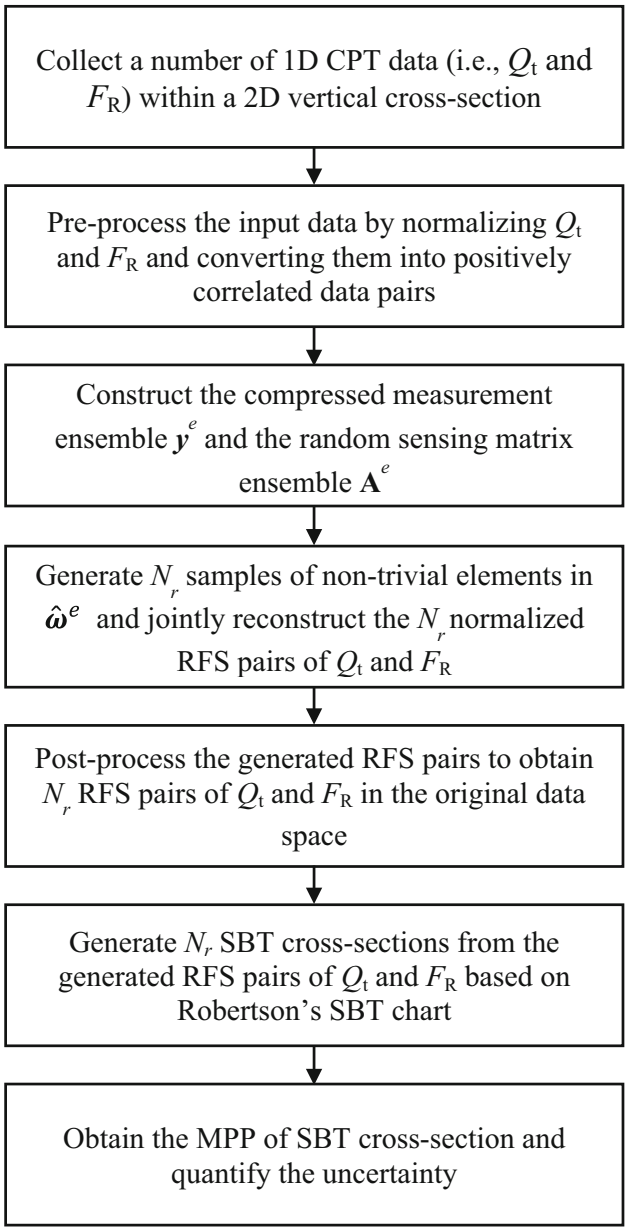


Fig. 2 Flowchart for the proposed method

After pre-processing the input data, compressed measurement ensemble $\boldsymbol{y}^e$ and the random sensing matrix ensemble $\mathbf{A}^e$ in Eq. (7) can be constructed. The Q_t and F_R measurements are compressed using a Gaussian random matrix by Eq. (4), respectively, leading to a compressed measurement ensemble $\boldsymbol{y}^e = \begin{bmatrix} \boldsymbol{y}^{l,c}_{Q_t} \\ \boldsymbol{y}^{l,c}_{F_R} \end{bmatrix}$. The t^{th}, $t = 1, 2, \ldots, N$, basis function $\mathbf{B}_t^{2D}$ can be compressed, using the same Gaussian random matrix by Eq. (4), to construct the random sensing matrix ensemble $\mathbf{A}^e = \begin{bmatrix} \mathbf{A}^R & \mathbf{A}^R & \mathbf{0} \\ \mathbf{A}^R & \mathbf{0} & \mathbf{A}^R \end{bmatrix}$, based on locations of measurements. Then, N_r samples of non-trivial elements in $\hat{\boldsymbol{\omega}}^e$ are generated from the compressed measurement ensemble $\boldsymbol{y}^e$ using Gibbs sampling, following the procedure described in Sect. 2.2. Consequently, N_r RFS pairs of normalized Q_t and F_R in the target cross-section can be jointly reconstructed using Eq. (2).

Once the normalized RFS pairs are reconstructed, they are converted to obtain N_r RFS pairs of Q_t and F_R in the original data space by post-processing the Q_t and F_R. The generated RFSs of Q_t and F_R are mapped back to original data space by multiplying σ_{Q_t} and σ_{F_R}, and subsequently adding the multiplication results to μ_{Q_t} and μ_{F_R}, respectively. After that, N_r SBT cross-sections are generated from the RFS pairs of Q_t and F_R based on Robertson's SBT chart. Finally, the MPP of SBT cross-section can be obtained using the Eq. (18) and the uncertainty can be quantified using the Eq. (19).

4 Simulated examples of CPT data for soil classification and stratification

In this section, the proposed method is illustrated and validated using a set of simulated CPT data. To facilitate the illustration, a 2D vertical cross-section with four different soil types is simulated, as shown in Fig. 3. The four soil types simulated are clay, silt mixtures, sand mixtures, and sand, with the corresponding SBT value of 3 to 6, respectively. Because the subsurface condition is unknown in practice, the simulated cross-section in Fig. 3 is for validation only in this study.

In this example, two CPT parameters, i.e., Q_t and F_R, are firstly simulated within each soil unit with a resolution of 0.01 m along depth and 0.2 m along horizontal direction using a spectral random field generator [32]. The random field simulation produces 2D cross-section pairs of Q_t and F_R, which are stored in two matrices with a size of 1280 × 256. Each matrix represents a 2D vertical cross-section with a thickness of 12.79 m and a width of 51 m. In other words, the cross-section contains 256 CPT soundings (i.e., $N_2 = 256$), and each CPT sounding has a depth of 12.79 m with 1280 data points (i.e., $N_1 = 1280$). Table 2 summarizes the parameters used for random field simulation of each soil unit. The random field parameters include mean μ, standard deviation σ, vertical and horizontal correlation lengths, λ_v and λ_h, and cross-correlation coefficient of ln Q_t and ln F_R. ln Q_t and ln F_R are logarithms of Q_t and F_R, respectively. ln Q_t and ln F_R are taken to follow normal distributions, respectively, so Q_t and F_R follow lognormal distributions. In this example, an exponential autocorrelation function is adopted and expressed as:

$$\rho = \exp\left[-2\sqrt{\frac{(\Delta x_v)^2}{\lambda_v} + \frac{(\Delta x_h)^2}{\lambda_h}}\right] \tag{21}$$

where ρ is the auto-correlation coefficient between two points in the cross-section; Δx_v and Δx_h indicate the absolute distances between any two points along vertical and horizontal directions, respectively. Then, two cross-correlated RFSs of Q_t and F_R are generated in the 2D cross-section with spatially varying layer boundary, as shown in Fig. 4.

Six CPT soundings (M1-M6) at horizontal coordinates of 0 m, 10.2 m, 20.4 m, 30.6 m, 40.8 m, and 51 m, respectively (see the six dotted lines in Fig. 4), are adopted as input to the proposed method. The Q_t and F_R profiles of these six CPT soundings are shown in Fig. 5. The measurement number of input data points accounts for about 6 / 256 × 100% ≈ 2.3% of the whole cross-section. Note that the complete Q_t and F_R data in the whole cross-section, i.e., Figs. 4a and b, are only used for validation and comparison of the results from the proposed method.

4.1 *Generation of RFS pairs of Q_t and F_R data*

Using the six sets of CPT data shown in Fig. 5, $N_r = 200$ 2D RFS pairs of complete Q_t and F_R cross-sections are generated by the proposed method. Four examples of generated RFS pairs are shown by colormap in Fig. 6. Note that a total of 400 RFSs are generated in pairs, i.e., 200 RFS pairs of Q_t and F_R, from only six CPT soundings. Mean and standard deviation of the 200 RFS pairs are

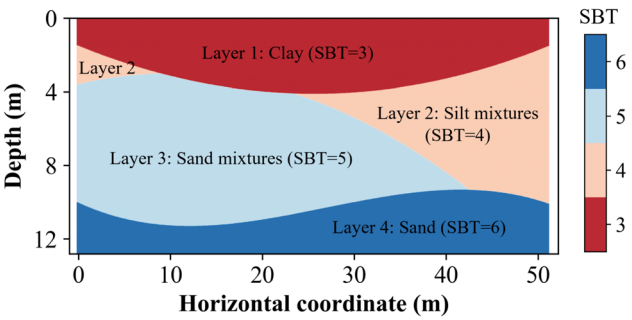


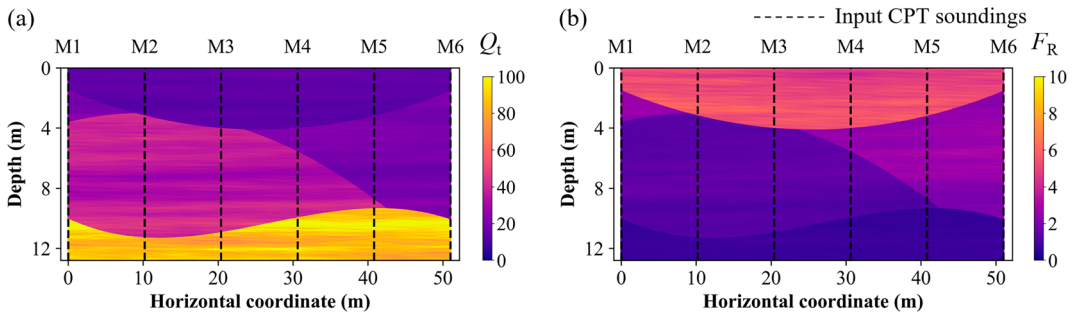
Fig. 3 A simulated 2D cross-section with different soil zones

Table 2 Parameters used in random field simulation of Q_t and F_R data in each layer

Soil layer	Mean		Standard deviation		Horizontal correlation length (m)		Vertical correlation length (m)		Cross-correlation coefficient
	ln F_R	ln Q_t	ln F_R	ln Q_t	ln F_R	ln Q_t	ln F_R	ln Q_t	
Layer 1	1.7	2.6	0.15	0.15	25	25	4	6	–0.7
Layer 2	0.9	3	0.15	0.15	40	40	4	6	–0.7
Layer 3	0.3	3.6	0.15	0.15	35	35	4	6	–0.7
Layer 4	–0.3	4.4	0.15	0.15	30	30	4	6	–0.7

shown in Fig. 7. The averages of 200 RFS pairs of Q_t and F_R are shown in Figs. 7a and b, respectively, and they are very similar to the original simulated data in Figs. 4a and b, although some local differences are observed. This is due to the uncertainty of joint sparse representation arising from limited number of CPT soundings. For quantification of uncertainty, standard deviation of 200 Q_t and F_R RFSs are calculated and shown in Figs. 7c and d, respectively. It is observed that the uncertainty of Q_t and F_R is relatively small at the locations with CPT soundings, but increases at locations without measurements.

To clearly show the performance of the joint sparse representation, the 1D profiles at three locations without measurements U1-U3 (see the three thin solid lines in



Simulated **a** Q_t and **b** F_R 2D cross-section with spatially varying soil layer boundaries

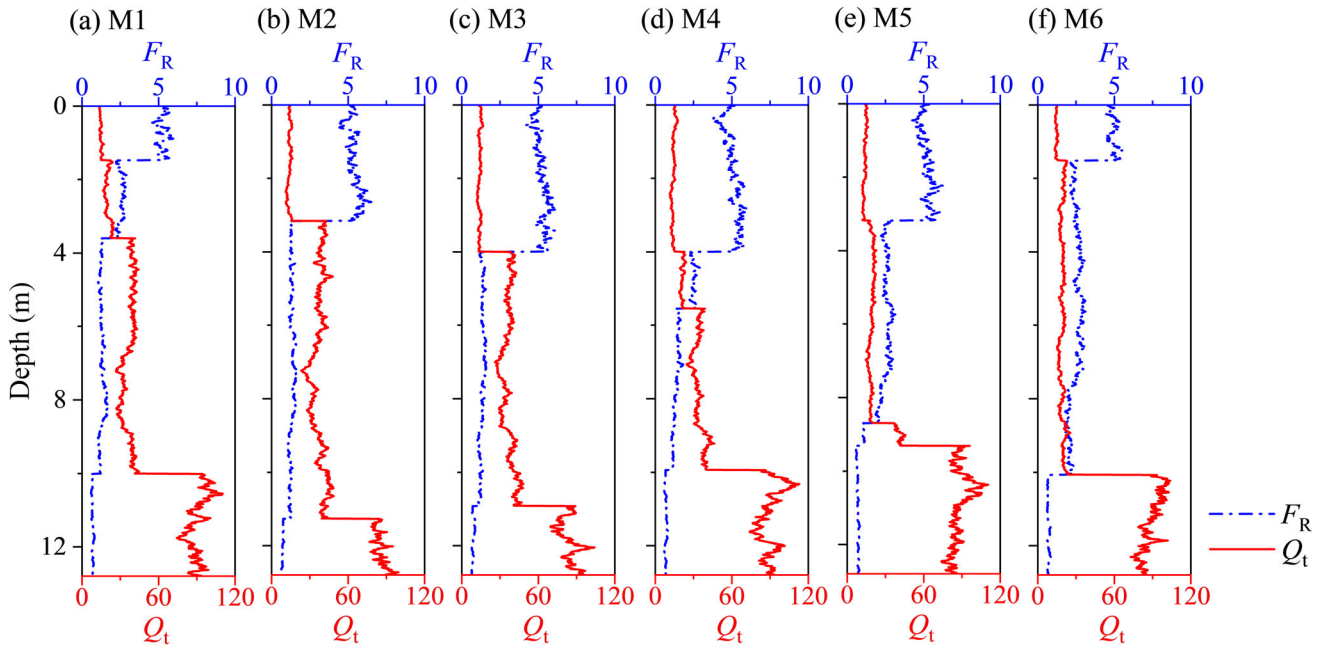


Fig. 5 Profiles of simulated CPT soundings at M1–M6: **a** M1; **b** M2; **c** M3; **d** M4; **e** M5; and **f** M6

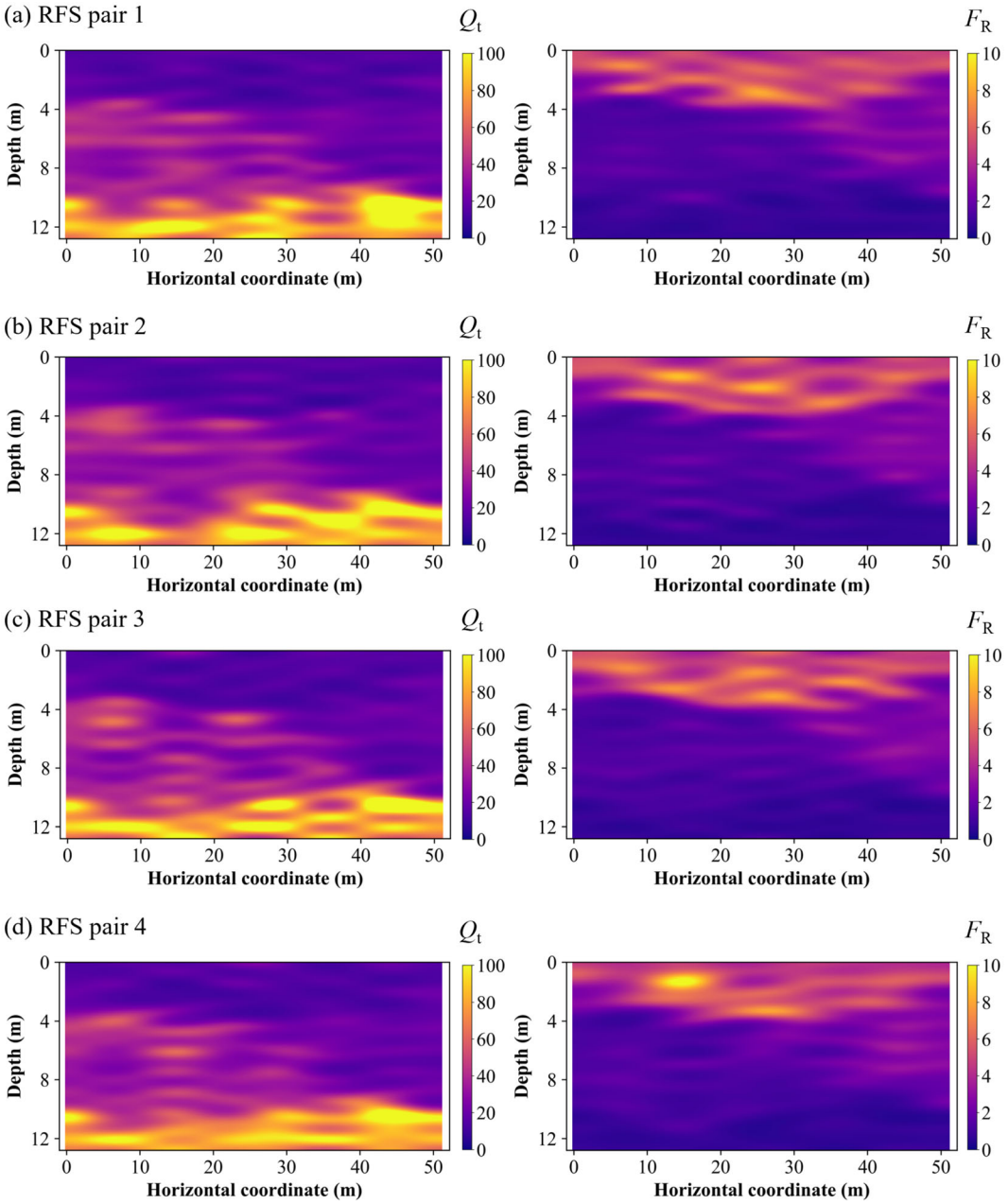


Fig. 6 Four examples of generated RFS pairs of Q_t and F_R: **a** RFS pair 1; **b** RFS pair 2; **c** RFS pair 3; and **d** RFS pair 4

Fig. 7) are provided respectively in Fig. 8. These three locations are all located at the midpoint between two input CPT soundings, each representing the worst situation for prediction. Figure 8 shows the original Q_t and F_R profiles by black solid lines and the mean of 200 RFSs from the proposed method by orange solid lines. It is evident that two lines have consistent trends, although some differences occur at the locations with local variations and abrupt jumps. A 95% confidence interval is depicted by two blue dashed lines. Most local variations fall within the 95% confidence interval. In addition, four examples of the RFSs are shown in each subplot by four grey solid lines. The results indicate that the proposed method for the joint sparse representation of Q_t and F_R performs well with properly quantified uncertainty.

Then, the generated $N_r = 200$ RFS pairs of Q_t and F_R are used to produce $N_r = 200$ random samples of SBT map according to Robertson's SBT chart. The details of the soil classification and stratification based on the generated RFS pairs of Q_t and F_R are described in the next subsection.

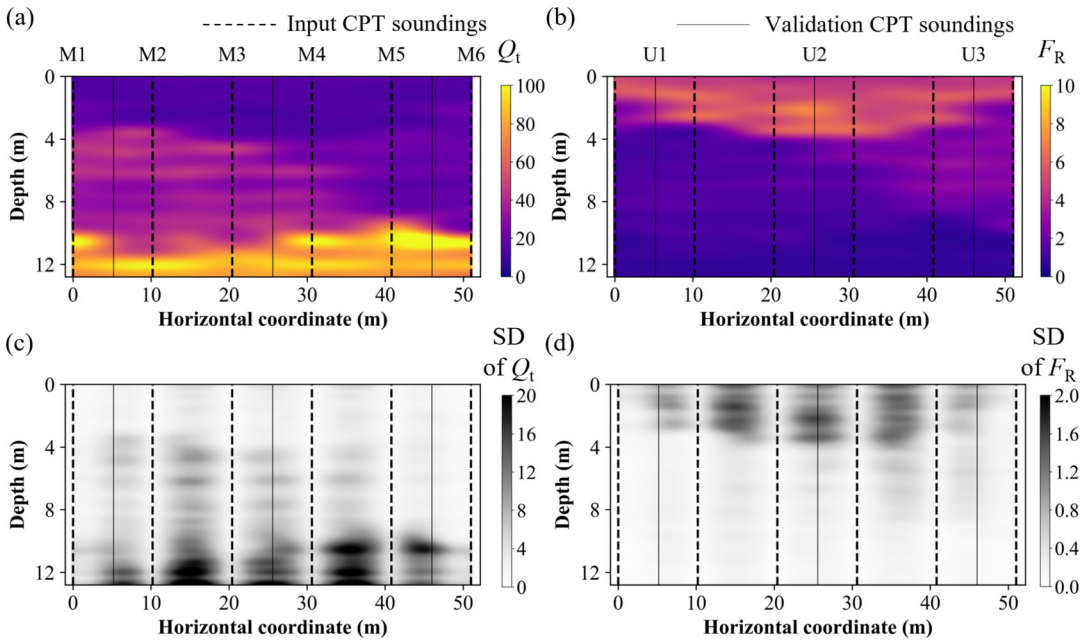


Fig. 7 Statistics of 200 RFS pairs of Q_t and F_R: **a** mean of Q_t; **b** mean of F_R; **c** standard deviation (SD) of Q_t; and **d** standard deviation (SD) of F_R

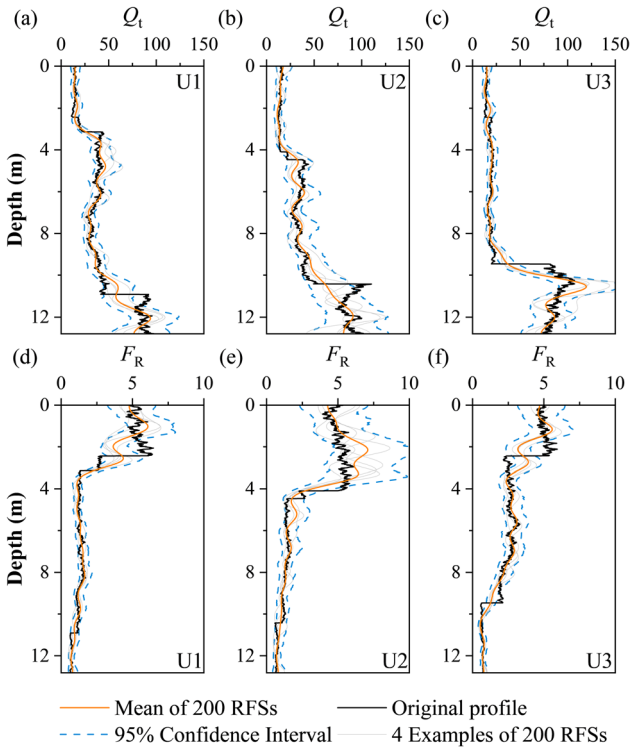


Fig. 8 Comparisons between original CPT profiles and representation profiles of Q_t and F_R at three untested locations U1-U3

4.2 Generation of SBT cross-sections and uncertainty quantification

Using one pair of 2D Q_t and F_R RFSs, SBT at each point within the cross-section is determined based on Robertson's SBT chart, leading to a random sample of SBT map. By repeating this step 200 times for 200 RFS pairs of Q_t and F_R, 200 SBT cross-sections are generated. Four examples of SBT cross-sections are depicted in Figs. 9a–d. These four 2D SBT cross-sections correspond respectively to the four RFS pairs of Q_t and F_R shown in Figs. 6a–d.

Each point in the cross-section has 200 probable SBT samples generated by the proposed method. The MPP of SBT cross-section is then obtained by Eq. (18) for showing the most probable SBT at all points, as illustrated in Fig. 10a. In the MPP cross-section, four SBTs (i.e., SBT3 to SBT6) are identified in Fig. 10a. The MPP is generally comparable to the ground true cross-section (i.e., Fig. 2) with an accuracy of 93.5% by Eq. (20), although the MPP is not perfectly identical to the ground true condition because of the uncertainty arising from a limited number of CPTs used as input. The uncertainty is quantified by calculating the dispersion by Eq. (19), as shown in Fig. 10b. Most areas within each soil layer are shown as dark purple or dark blue, indicating the small dispersion or small uncertainty of SBT within each soil layer. In contrast, the light yellow indicates the large dispersion or high uncertainty of SBT, mainly appearing around the boundaries between different soil types.

The proposed method is performed using a desktop computer with a CPU of Intel Core i7-10700 at 2.9 GHz and 32 GB RAM in this study. In this simulated example, it takes about 109.2 s, less than two minutes, for generating 200 RFS pairs of Q_t and F_R in joint sparse representation. To benchmark the computational efficiency of the proposed method, a similar study but with a deterministic

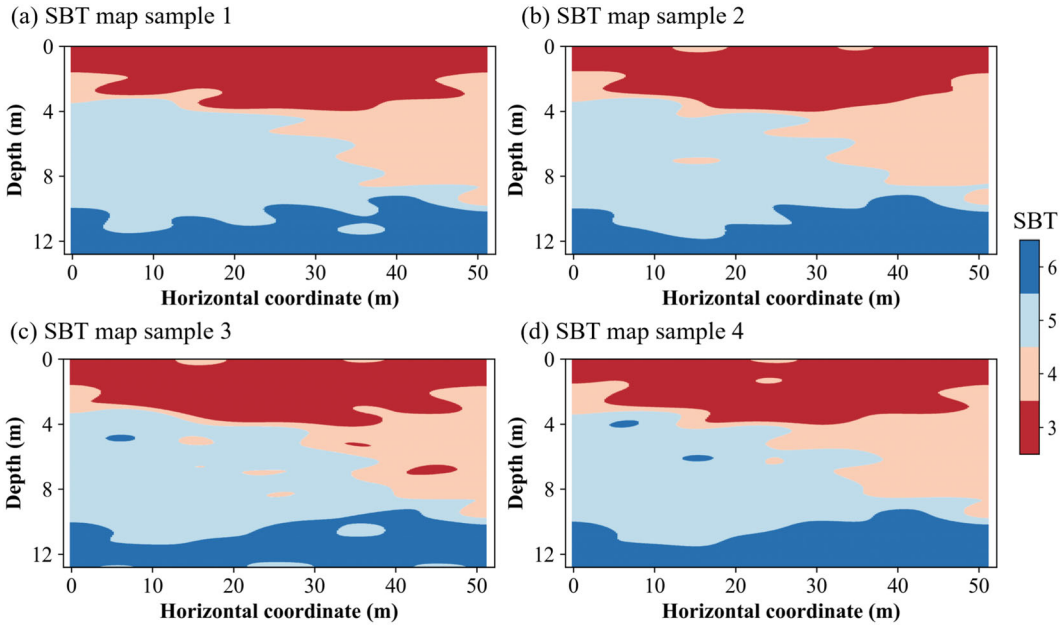


Fig. 9 SBT cross-sections from the four Q_t and F_R RFS pairs shown in Fig. 6: **a** SBT map sample 1; **b** SBT map sample 2; **c** SBT map sample 3; and **d** SBT map sample 4

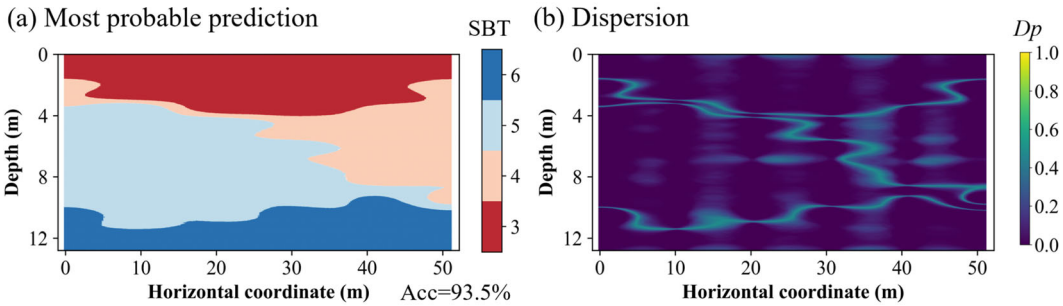


Fig. 10 Soil classification and stratification results from joint sparse representation of Q_t and F_R: **a** most probable prediction; and **b** dispersion

sensing matrix is also performed. It takes about 1491.6 s, over twenty minutes, under the same conditions. Using a randomization of the input CPT measurements significantly improves computational efficiency. It is more than ten times faster than a similar study without randomization of input measurements.

4.3 Effect of data noise

To simulate the noisy CPT data and validate the ability of denoising using the proposed method, a white Gaussian noise with a signal-to-noise ratio (SNR) of 15 dB is simulated and added to the input Q_t and F_R. An SNR of 15 dB indicates that the signal power is $10^{1.5}$ (approximately 31.6) times greater than the noise power. A Gaussian random variable with zero mean and variance scaled to achieve this SNR is added to the original data, resulting in noisy profiles of Q_t and F_R. The profiles of input Q_t and F_R with a white Gaussian noise are illustrated in Fig. 11. For fair comparison, the noisy inputs are positioned at the same locations as those in Fig. 5. Then, 200 RFS pairs of Q_t and F_R are generated by joint sparse representation from the simulated noisy CPT data. Subsequently, 200 probable SBT cross-sections are obtained from the 200 generated RFS pairs according to the SBT chart. The MPP of SBT cross-section is obtained from the statistics of 200 probable SBT cross-sections using Eq. (18), as shown in Fig. 12a. The uncertainty is quantified by calculating the dispersion using Eq. (19), as shown in Fig. 12b.

Four SBTs (i.e., SBT3 to SBT6) are identified in Fig. 12a, which are the same as Fig. 10a. Although the input data contains white Gaussian noise, the spatial variation trend can be clearly captured using the proposed method. The overall accuracy of the prediction from noisy input is calculated as 93.4% by Eq. (20). That is virtually identical to 93.5% when using input data without noises.

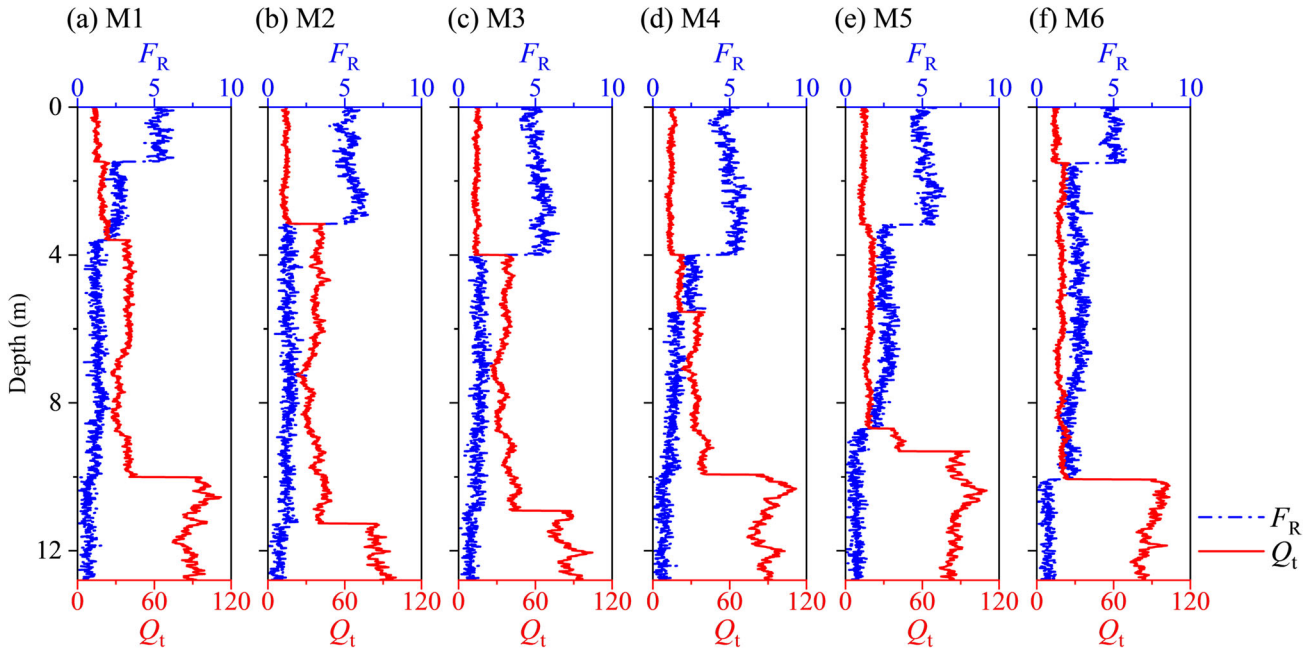


Fig. 11 Profiles of input Q_t and F_R with a white Gaussian noise at M1–M6: **a** M1; **b** M2; **c** M3; **d** M4; **e** M5; and **f** M6

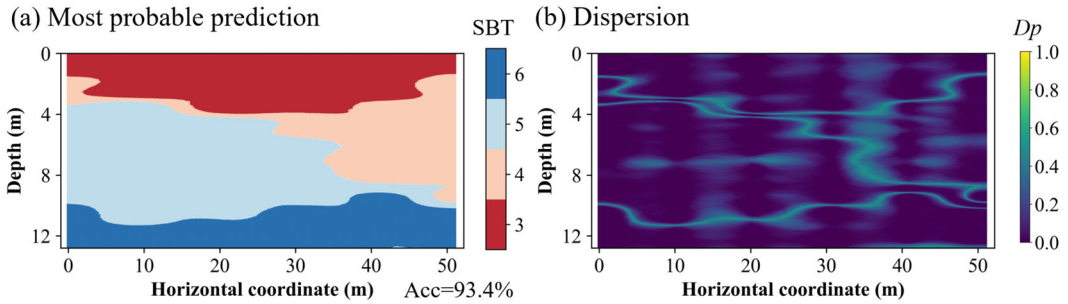


Fig. 12 Soil classification and stratification results from joint sparse representation of Q_t and F_R with a white Gaussian noise: **a** most probable prediction; and **b** dispersion

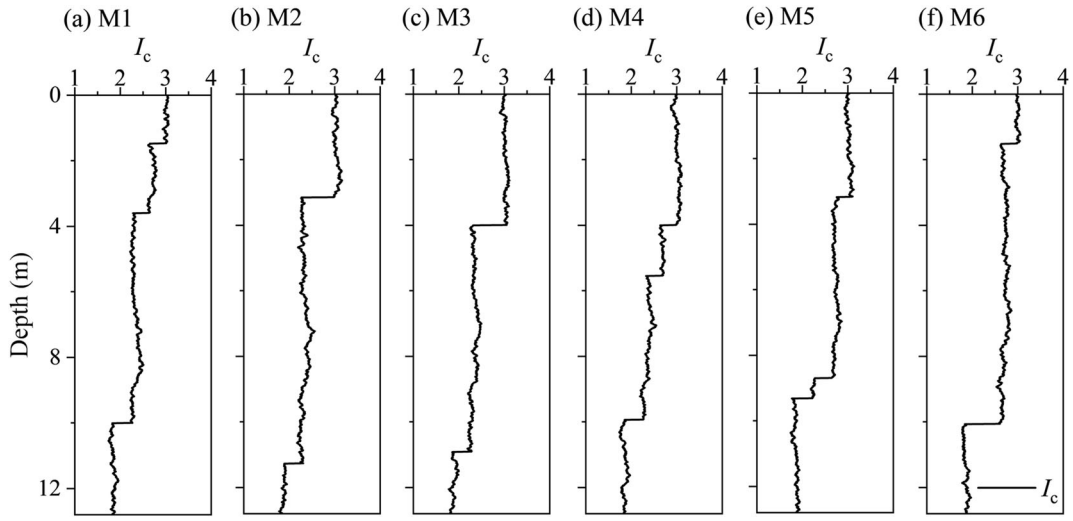


Fig. 13 Input I_c calculated from Q_t and F_R in Fig. 4 at M1–M6: **a** M1; **b** M2; **c** M3; **d** M4; **e** M5; and **f** M6

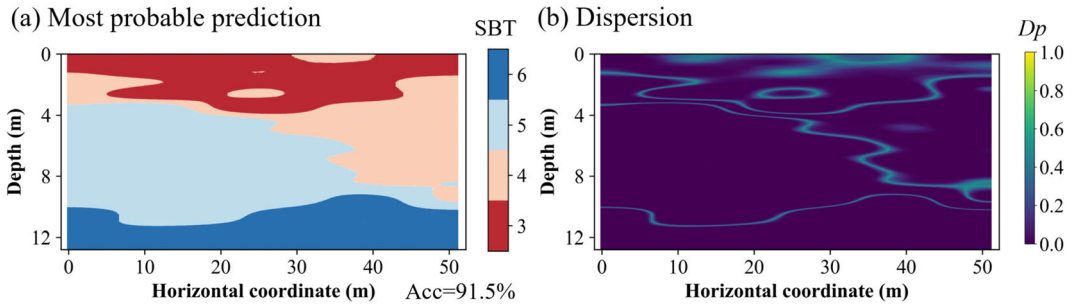


Fig. 14 Soil classification and stratification results from individual representation of I_c: **a** most probable prediction; and **b** dispersion

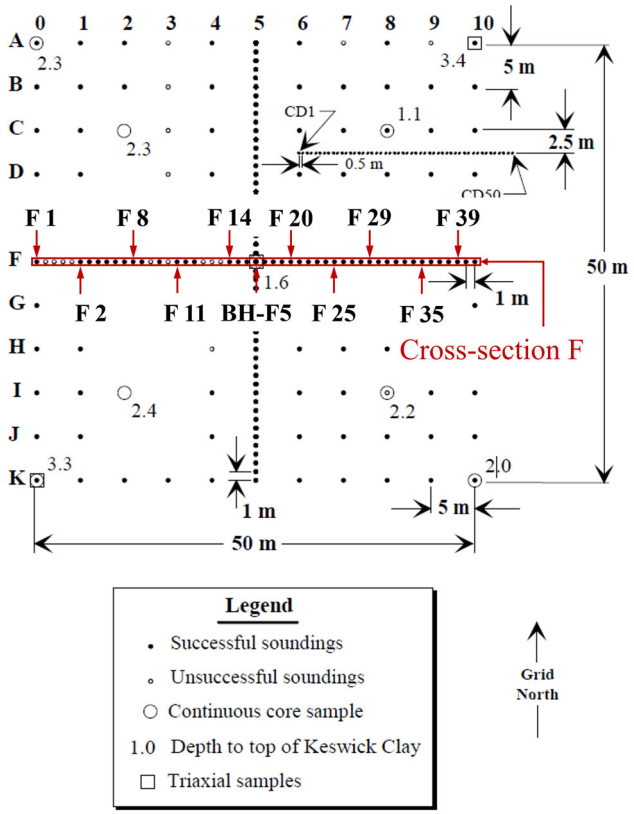


Fig. 15 Layout of 41 CPT soundings (F1-F41) in a cross-section at the South Parklands site in Australia (modified from [21])

Similarly, large dispersion or high uncertainty of SBT appears around the boundaries between different soil layers, and they are also close to the true boundaries in Fig. 3. In addition, white Gaussian noises with different SNR levels are tested to validate its denoising capability. The results indicate that the proposed method can denoise automatically when using noisy input data for the soil classification and stratification.

4.4 Effect of joint sparse representation

To demonstrate the advantage of modelling data cross-correlation, a comparative study is conducted using a single representation of SBT index (I_c) for CPT-based soil classification and stratification. As shown in Fig. 13, the input I_c data from the six simulated CPT soundings are calculated using the input Q_t and F_R shown in Fig. 5. For a fair comparison, a randomization of input measurements is also performed on the I_c obtained. Then 200 RFSs of I_c are generated from the input I_c data by Eq. (1). According to the I_c-based soil classification summarized in Table 1, 200 probable SBT cross-sections are obtained from 200 generated RFSs of I_c. The MPP of SBT cross-section is then obtained from statistical analysis of the 200 probable SBT cross-sections by Eq. (18) and shown in Fig. 14a. The uncertainty is quantified by a dispersion plot calculated by Eq. (19) and shown in Fig. 14b.

In Fig. 14a, four SBTs (i.e., SBT3 to SBT6) are determined in this example. Although I_c-based prediction can identify all soil behavior types in this example, it may miss the information of SBT1, SBT8, and SBT9 in other situations, e.g., the real data example in the next section. The overall accuracy from I_c-based prediction is calculated as 91.5%, which is slightly lower than the prediction from joint sparse representation under the same conditions. Figure 14a shows that I_c-based prediction performs poorly in identifying soils of SBT3 when compared to prediction using joint sparse representation. That is because I_c-based soil classification disregards the cross-correlation between Q_t and F_R. This cross-correlation between Q_t and F_R provides additional information for improving accuracy of CPT-based soil classification and stratification. Moreover, the quantified uncertainty between 0 and 2 m deep in Fig. 14b is significantly greater than that in Fig. 10b. In general, the performance of the proposed method is better than the I_c-based prediction in this simulated example. Leveraging data cross-correlation by joint sparse representation of Q_t and F_R improves CPT-based soil classification and stratification.

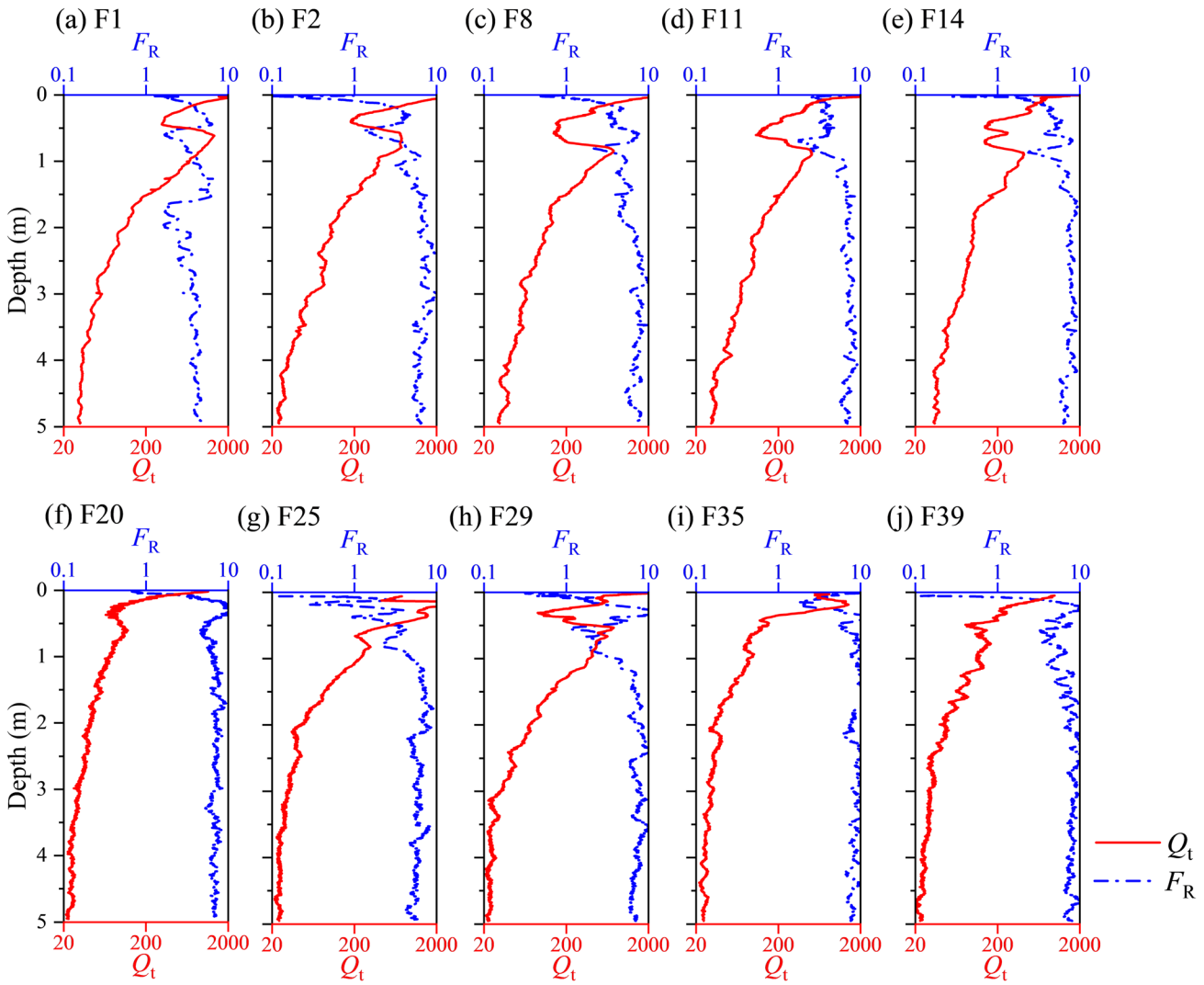


Fig. 16 Q_t and F_R profiles of the 10 real CPT soundings used as input data

5 Real data example

The proposed method is illustrated and validated using a set of real CPT data in this section. The CPT data is taken from an experimental site in the South Parklands in South Australia [20, 21]. Figure 15 shows the layout of 41 CPT soundings in the cross-section F of interest at the study site. The cross-section F starts at location of CPT sounding F1 and ends at location of CPT sounding F41, with the length of 50 m and minimum interval between each nearby CPT soundings of 1 m. In this example, 10 CPT soundings (i.e., F1, F2, F8, F11, F14, F20, F25, F29, F35, and F39) in this cross-section are used as input to the proposed method for the prediction of soil classification and stratification. The Q_t and F_R profiles of these 10 input CPT soundings are shown in Fig. 16. In the vertical direction, the interval is 0.005 m, and the max depth is 4.945 m. The vertical and horizontal dimensions of the cross-section are 4.945 m and 50 m with the target spatial resolutions of 0.005 m and 0.5 m, respectively. Hence, the target cross-section has a matrix dimension of 990 × 101. The measurement data point number of each input parameter (e.g., Q_t) is 8,469, and the measurement ratio in this study is about 8,469 / (990 × 101) × 100% ≈ 8.5%.

Using the 10 CPT data as input, N_r = 200 2D SBT cross-sections are generated using the proposed method. The MPP and dispersion of SBT cross-section are then obtained and shown in Figs. 17a and b, respectively. Seven SBTs (i.e., SBT3 to SBT9) are identified in Fig. 17a. In addition, the large dispersion mainly appears at the boundaries between different SBTs, as shown in Fig. 17b, similar to the results of the simulated example shown in the previous subsection. Note that the I_c-based soil classification cannot be used in this example, because it cannot identify very stiff soils, i.e., SBT8 and SBT9, which occur about half of the cross section.

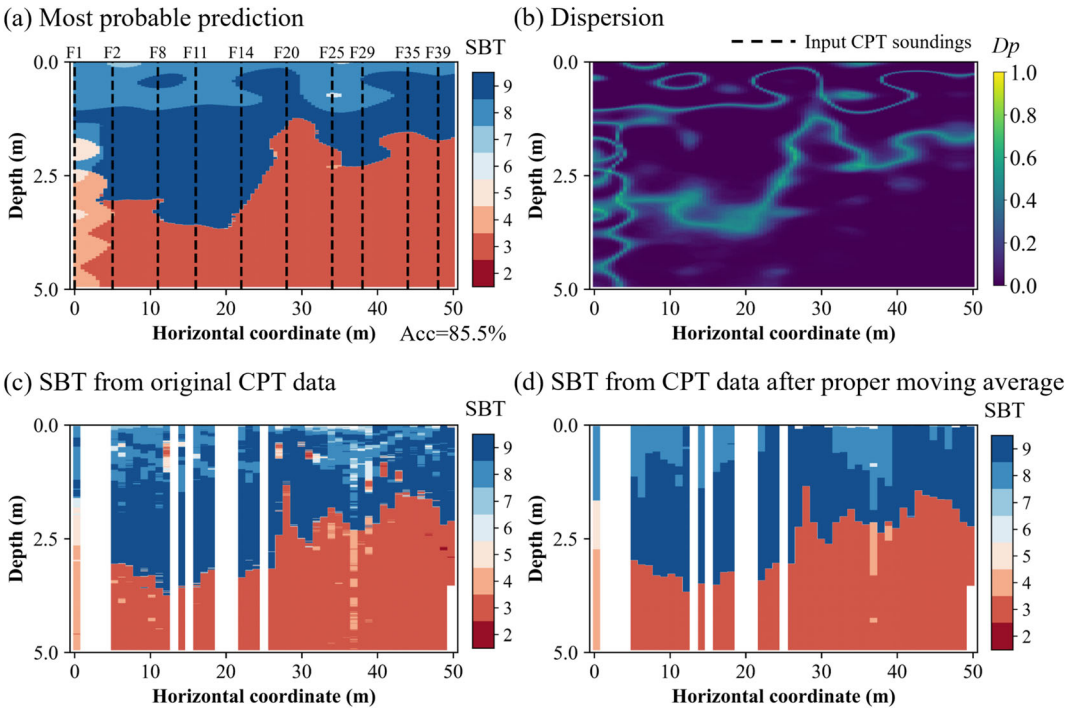


Fig. 17 Soil classification and stratification results in real data example: **a** most probable prediction; **b** dispersion; **c** SBT distribution from the original CPT data; and **d** SBT distribution from the CPT data after proper moving average

In real practices, the ground truth is unknown at the untested locations. Hence, to evaluate performance of the proposed method, the SBT results from the proposed method are compared with those directly calculated from the remaining 31 CPT soundings. The SBT distribution from original CPT data directly based on Robertson chart is shown in Fig. 17c. It is obvious that the results are noisy and unrealistic, with rapid fluctuation of SBTs within a short distance. To filter out the noises in the engineering practices, moving average of CPT data is a simple and effective method, particularly when there is no prior information or subjective judgement. Therefore, in this real example, moving averages of original Q_t and F_R data are conducted along the vertical direction. Because the typical vertical correlation lengths of CPT data range from 0.1 m to 2.2 m, with a mean of 0.9 m [33], the length of the moving window is set as 1 m. Figure 17d shows the SBT distribution after moving average of CPT data, and the results in Fig. 17d is used for comparison with the results obtained from the method proposed in this study. The comparison shows that the SBT results from the proposed method are consistent with results after removing data noises from the 31 remaining CPT data by moving average. The overall accuracy of the whole cross-section is computed as 85.5% by Eq. (20). Note that, while the original CPT data can not provide complete stratification information (see Fig. 17c), the proposed method, using only 10 CPT soundings, is able to predict a complete cross-section with high accuracy and quantified uncertainty.

For detailed comparison of individual CPT soundings, 4 specific CPT locations (i.e., F5, F15, F27, and F31) are chosen. Figures 18a–d respectively plot the SBT derived directly from the original CPT data, the SBT obtained from CPT data after moving average, and predictions using the proposed method of these four CPT soundings. Although the SBTs directly from original CPT data exhibit rapid fluctuation, the proposed method could provide predictions consistent with the moving-averaged results at these four validation locations with high accuracy. That indicates the proposed method can not only capture the spatial variation trend of soil stratification, but also remove the noises automatically. Furthermore, the statistical analysis of accuracy of 31 validation CPT soundings is performed and shown in Fig. 18e. The histogram plot in Fig. 18e shows that prediction accuracy of 31 CPT soundings ranges from 56.0% to 97.2%, with a mean of 84.1% and a median of 86.0%.

In addition, for cross-validation, a comparison of most probable prediction with soil logging from the borehole BH-F5 is performed (see Fig. 15 for the borehole location). The borehole BH-F5 contains four layers, as shown in Fig. 19a. Because the soil types in the second and third layer are almost the same, those four layers can be simplified into three layers, i.e., a clayey sand layer at depths from 0 to 0.3 m, followed by a fine to medium grained sand

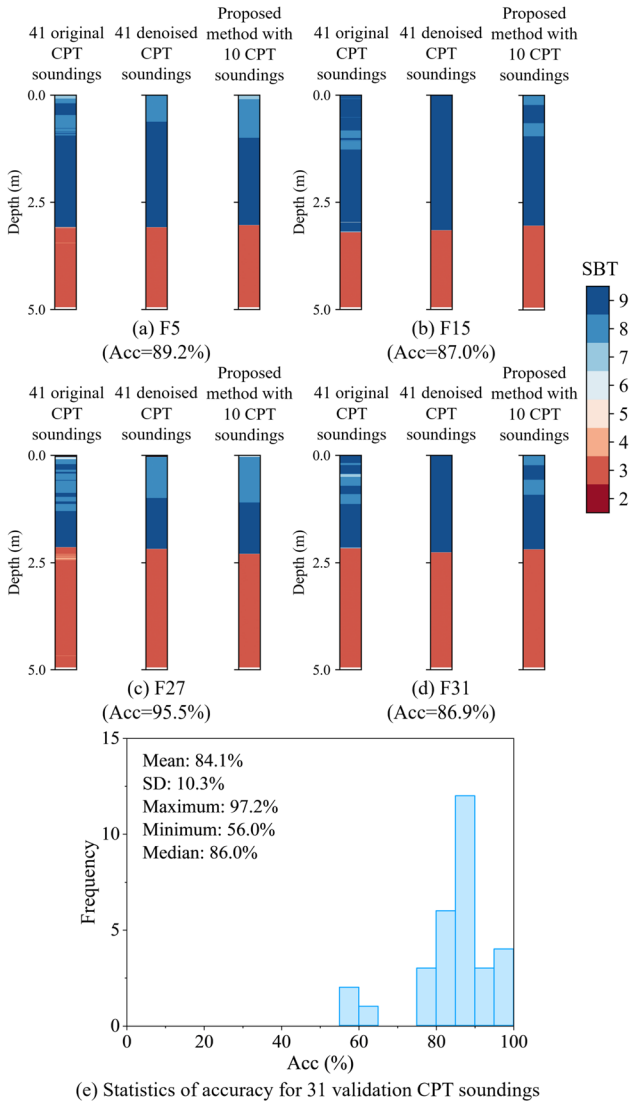


Fig. 18 Detailed comparison of accuracy at **a** F5, **b** F15, **c** F27, and **d** F31; and **e** statistics of accuracy for 31 validation CPT soundings

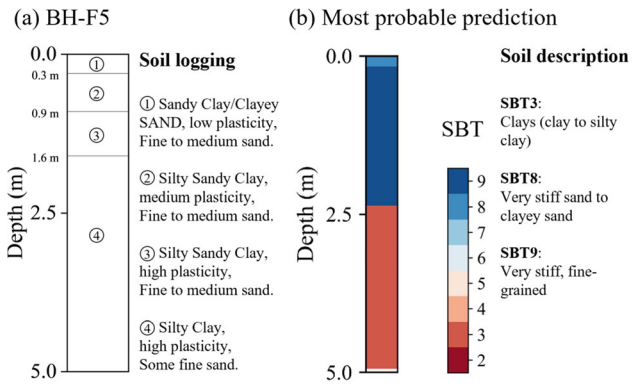


Fig. 19 Comparison between **a** soil logging from the borehole BH-F5 (modified from [20]) and **b** most probable prediction using proposed method

layer at depths from 0.3 to 1.6 m and a silty clay layer with some fine sand at depths from 1.6 to 5.0 m. The location of borehole BH-F5 is at the center of cross-section F, as shown in Fig. 15. The most probable prediction obtained from the proposed method at the same location is shown in Fig. 19b. This SBT profile shows that soils at depths from 0 to 0.16 m are mainly very stiff sand to clayey sand, and soils at depths from 0.16 to 2.40 m are very stiff and fine-grained, and soils at depths from 2.40 to 4.95 m are mainly clay (clay to silty clay). The results from the proposed method are generally consistent with the borehole logging, although there may be some differences between SBT boundaries and the boundaries obtained from borehole. That is because the stratification methods for CPT and borehole are different, i.e., the CPT stratification is based on soil behavior types, while borehole stratification is obtained through visual inspection of soil sample compositions [35]. The proposed method performs well for both simulated and real data.

6 Summary and conclusions

A novel method was developed in this study for CPT-based soil classification and stratification using joint sparse representation of CPT data with consideration of data cross-correlation and noises. The proposed method jointly reconstructs the strong-correlated Q_t and F_R from limited CPT soundings in a target cross-section. Under the joint sparse representation framework, the auto- and cross-correlated structure of Q_t and F_R can be modelled simultaneously. To filter out the high-frequency noises automatically, a randomization of input measurements is conducted. In addition, the uncertainty is quantified under a Bayesian framework. The implementation procedure of the proposed method was described in detail. The proposed method was illustrated and validated using both simulated and real data examples.

The results show that the proposed method accurately and efficiently predicts the spatial distribution of SBT from limited CPT soundings by joint sparse representation. Many probable samples of SBT cross-sections are generated to quantify the uncertainty, and it is found that large uncertainty mainly occurs at boundaries of different soil layers. The proposed method efficiently leverages on additional valuable information from the strong data cross-correlation between Q_t and F_R. Modelling such cross-correlated structure by the proposed method improves the CPT-based soil classification and stratification from sparse CPT soundings. In addition, when handling the noisy CPT data, a scenario often encountered in practice, the proposed method is able to automatically filter out the noises and capture the important spatial pattern for soil stratification

and zonation in a data-driven manner. Furthermore, the proposed method also significantly improves computational efficiency, when compared to existing methods without a randomization of input measurements. The proposed method might also be extended to adopt complicated basis functions, such as those constructed using data obtained from nearby or similar sites [14, 15], in future studies.

Acknowledgements The work described in this paper was supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region (Project Nos: 11203322 and 11207724). The financial support is gratefully acknowledged.

Author contributions J.C.Y. contributed to methodology, software, validation, formal analysis, data curation, and writing–original draft, review and editing. Y.W. contributed to the study conception, methodology, validation, writing–original draft, review and editing, supervision, and funding acquisition. Z.G. contributed to the code and edited the manuscript. K.S. provided valuable comments and edited the manuscript.

Data availability The data generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Baron D, Duarte MF, Sarvotham S, Wakin MB, Baraniuk RG (2005) An information-theoretic approach to distributed compressed sensing. In: Proc. 43rd Conf. Commun. Control Comput
2. Baron D, Wakin MB, Duarte MF, Sarvotham S, Baraniuk RG (2006) Distributed compressed sensing. Technical report TREE0612, Rice University, Houston, TX
3. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
4. Brunton SL, Kutz JN (2022) Data-driven science and engineering: machine learning, dynamical systems, and control. Cambridge University Press, Cambridge
5. Candès EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25:21–30. https://doi.org/10.1109/MSP.2007.914731
6. Clayton CR, Matthews MC, Simons NE (1995) Site investigation: a handbook for engineers. Blackwell Science
7. Do TT, Gan L, Nguyen NH, Tran TD (2012) Fast and efficient compressive sensing using structurally random matrices. IEEE Trans Signal Process 60(1):139–154. https://doi.org/10.1109/TSP.2011.2170977
8. Duan L, Hao J, Xie S, Zhou Z, Ye X (2002) Determining weathering rates of soils in China. Geoderma 110(3–4):205–225. https://doi.org/10.1016/S0016-7061(02)00231-8
9. Ecemis N, Arik MS, Taneri H (2023) Effect of drainage conditions on CPT resistance of silty sand: physical model and field tests. Acta Geotech 18:6709–6724. https://doi.org/10.1007/s11440-023-01915-3
10. Gelfand AE, Hills SE, Racine-Poon A, Smith AF (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. J Am Stat Assoc 85(412):972–985. https://doi.org/10.1080/01621459.1990.10474968
11. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741. https://doi.org/10.1109/TPAMI.1984.4767596
12. Guan Z, Wang Y (2023) Data-driven simulation of two-dimensional cross-correlated random fields from limited measurements using joint sparse representation. Reliab Eng Syst Saf 238:109408. https://doi.org/10.1016/j.ress.2023.109408
13. Guan Z, Wang Y, Phoon KK (2024) Fusion of sparse non-co-located measurements from multiple sources for geotechnical site investigation. Canadian Geotech J 61(8):1574–1592. https://doi.org/10.1139/cgj-2023-0289
14. Guan Z, Wang Y, Phoon KK (2025) Data-driven geotechnical site recognition using machine learning and sparse representation. Eng Geol 346:107893. https://doi.org/10.1016/j.enggeo.2024.107893
15. Guan Z, Wang Y, Phoon KK (2024) Dictionary learning of spatial variability at a specific site using data from other sites. J Geotech Geoenviron Eng 150(9):04024072. https://doi.org/10.1061/jggefk.gteng-12408
16. Han Y, Zhao W, Ding J, Ferreira CSS (2023) Soil erodibility for water and wind erosion and its relationship to vegetation and soil properties in China's drylands. Sci Total Environ 903:166639. https://doi.org/10.1016/j.scitotenv.2023.166639
17. Hu Y, Wang Y (2020) Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation. Comput Geotech 124:103634. https://doi.org/10.1016/j.compgeo.2020.103634
18. Huang J, Griffiths DV (2010) One-dimensional consolidation theories for layered soil and coupled and uncoupled solutions by the finite-element method. Géotechnique 60(9):709–713. https://doi.org/10.1680/geot.08.P.038
19. Huang M, Shuku T, Shibata T, Nishimura S (2024) A novel data-driven 3D site characterisation method: tucker decomposition-bayesian compressive sensing and benchmarking study. Georisk assess manage risk eng sys geohazards 19(2):247–266. https://doi.org/10.1080/17499518.2024.2395555
20. Jaksa MB (1995) The influence of spatial variability on the geotechnical design properties of a stiff, overconsolidated clay. Doctoral dissertation. The University of Adelaide
21. Jaksa MB, Kaggwa WS, Brooker PI (2002) An improved statistically based technique for evaluating the CPT friction ratio. Geotech Test J 25(1):61–69. https://doi.org/10.1520/GTJ11080J
22. Jefferies MG, Davies MP (1993) Use of CPTU to estimate equivalent SPT N60. Geotech Test J 16(4):458–468. https://doi.org/10.1520/GTJ10286J
23. Jefferies MG, Davies MP (1991) Soil classification by the cone penetration test: discussion. Can Geotech J 28(1):173–176. https://doi.org/10.1139/t91-023
24. Ji S, Xue Y, Carin L (2008) Bayesian compressive sensing. IEEE Trans Signal Process 56(6):2346–2356. https://doi.org/10.1109/TSP.2007.914345
25. Jin SJ, Sun WZ, Huang L (2023) Joint optimization methods for Gaussian random measurement matrix based on column coherence in compressed sensing. Signal Process 207:108941. https://doi.org/10.1016/j.sigpro.2023.108941
26. Lee KK, Cassidy MJ, Randolph MF (2013) Bearing capacity on sand overlying clay soils: experimental and finite-element investigation of potential punch-through failure. Géotechnique 63(15):1271–1284. https://doi.org/10.1680/geot.12.P.175

27. Li J, Zhang H, Zhang L, Ma L (2015) Hyperspectral anomaly detection by the use of background joint sparse representation. IEEE J Selected Topics Appl Earth Obs Remote Sensing 8(6):2523–2533. https://doi.org/10.1109/JSTARS.2015.2437073
28. Li P, Wang Y (2023) Interpretation of spatio-temporal variation of precipitation from spatially sparse measurements using Bayesian compressive sensing (BCS). Georisk Assess Manag Risk Eng Systems Geohazards 17(3):554–571. https://doi.org/10.1080/17499518.2023.2188464
29. Lyu B, Wang Y, Shi C (2024) Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge. Comput Geotech 170:106336. https://doi.org/10.1016/j.compgeo.2024.106336
30. Lyu B, Wang Y, Miao C, Yao J, Shum LKW, Wong AL, Ho RCM (2025) Fusion of limited site-specific borehole logs and geophysical data from a different site for three-dimensional subsurface geological modelling using Multi-Scale Generative Adversarial Network. J Geotech Geoenvironmental Eng. https://doi.org/10.1061/JGGEFK.GTENG-13369
31. Mayne PW, Christopher BR, Berg R, DeJong J (2002) Subsurface investigations—geotechnical site characterization Publication Number FHWA-NHI-01–031 Washington, DC: National Highway Institute, Federal Highway Administration
32. Müller S, Schüler L, Zech A, Heße F (2022) GSTools v1.3: a toolbox for geostatistical modelling in Python. Geoscientific Model Develop 15:3161–3182. https://doi.org/10.5194/gmd-2021-301
33. Phoon KK, Kulhawy FH (1999) Characterization of geotechnical variability. Can Geotech J 36(4):612–624. https://doi.org/10.1139/t99-038
34. Phoon KK, Ching J, Shuku T (2022) Challenges in data-driven site characterization. Georisk Assess Manag Risk Eng Systems Geohazards 16(1):114–126. https://doi.org/10.1080/17499518.2021.1896005
35. Robertson PK (1990) Soil classification using the cone penetration test. Can Geotech J 27(1):151–158. https://doi.org/10.1139/t90-014
36. Robertson PK, Wride C (1998) Evaluating cyclic liquefaction potential using the cone penetration test. Can Geotech J 35(3):442–459. https://doi.org/10.1139/t99-101
37. 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
38. Shi C, Wang Y (2022) Data-driven construction of three-dimensional subsurface geological models from limited site-specific boreholes and prior geological knowledge for underground digital twin. Tunn Undergr Space Technol 126:104493. https://doi.org/10.1016/j.tust.2022.104493
39. Shi C, Wang Y (2023) Data-driven sequential development of geological cross-sections along tunnel trajectory. Acta Geotech 18:1739–1754. https://doi.org/10.1007/s11440-022-01707-1
40. Strang G (2016) Introduction to linear algebra. Wellesley-Cambridge Press, MA, USA
41. Tian HM, Cao ZJ, Li DQ, Du W, Zhang FP (2022) Efficient and flexible Bayesian updating of embankment settlement on soft soils based on different monitoring datasets. Acta Geotech 17:1273–1294. https://doi.org/10.1007/s11440-021-01378-4
42. Tian HM, Wang Y, Shi C (2025) Machine learning-aided selection of CPT-based transformation models using field monitoring data from a specific project. Acta Geotech 20:439–459. https://doi.org/10.1007/s11440-024-02475-w
43. Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1:211–244. https://doi.org/10.1162/15324430152748236
44. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2008) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227. https://doi.org/10.1109/TPAMI.2008.79
45. Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S (2010) Sparse representation for computer vision and pattern recognition. Proc IEEE 98(6):1031–1044. https://doi.org/10.1109/JPROC.2010.2044470
46. 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
47. 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
48. Wang Y, Hu Y, Zhao T (2020) CPT-based subsurface soil classification and zonation in a 2D vertical cross-section using Bayesian compressive sampling. Can Geotech J 57(7):947–958. https://doi.org/10.1139/cgj-2019-0131
49. Wang Y, Hu Y, Phoon KK (2022) Non-parametric modelling and simulation of spatiotemporally varying geo-data. Georisk Assess Manag Risk Eng Systems Geohazards 16(1):77–97. https://doi.org/10.1080/17499518.2021.1971258
50. Xie J, Zeng C, Huang J, Zhang Y, Lu J (2024) A back analysis scheme for refined soil stratification based on integrating borehole and CPT data. Geoscience Frontier 15:101688. https://doi.org/10.1016/j.gsf.2023.101688
51. Xu J, Wang Y, Zhang L (2021) Interpolation of extremely sparse geo-data by data fusion and collaborative Bayesian compressive sampling. Comput Geotech 134:104098. https://doi.org/10.1016/j.compgeo.2021.104098
52. Zhang Z, Xu Y, Yang J, Li X, Zhang D (2015) A survey of sparse representation: algorithms and applications. IEEE Access 3:490–530. https://doi.org/10.1109/ACCESS.2015.2430359
53. Zhao LS, Zhuo S, Shen B (2023) An efficient model to estimate the soil profile and stratigraphic uncertainty quantification. Eng Geol 315:107025. https://doi.org/10.1016/j.enggeo.2023.107025
54. 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
55. Zhao T, Wang Y, Lu S, Xu L (2023) Fast stratification of geological cross-section from CPT results with missing data using multitask and modified Bayesian compressive sensing. Can Geotech J 60(12):1812–1834. https://doi.org/10.1139/cgj-2022-0131
56. Zinas O, Papaioannou I, Schneider R, Cuéllar P (2025) Multivariate Gaussian process regression for 3d site characterization from CPT and categorical borehole data. Eng Geol 352:108052. https://doi.org/10.1016/j.enggeo.2025.108052