Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5182-7_01-007-cd

Determination of Optimal CPT Locations for Characterizing Nonstationary Spatial Variability of Geotechnical Properties Using Efficient Bayesian Compressive Sensing

Tengyuan Zhao[1], Yu Wang[2], and Ling Xu[3]

[1]School of Human Settlements and Civil Engineering, Xi'an Jiaotong University, No.28, West Xianning Road, Xi'an, Shaanxi, 710049,P.R. China,
E-mail: tyzhao@xjtu.edu.cn
[2]Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, SAR,
E-mail: yuwang@cityu.edu.hk
[3]School of Human Settlements and Civil Engineering, Xi'an Jiaotong University, No.28, West Xianning Road, Xi'an, Shaanxi, 710049,P.R. China,
E-mail: xuling82@xjtu.edu.cn

Abstract: Spatial variability of geotechnical properties within multiple soil layers plays an essential role in geotechnical design or analysis, especially in probability-based analysis for geo-structures (e.g., slopes, tunnels, piles). This is often determined via laboratory methods using samples from drilled boreholes, alternatively in-situ testing methods, such as cone penetration test (CPT). Note that CPT has been widely used in recent decades, because it is fast, inexpensive, repeatable, and can obtain almost continuous soil response data when its cone is pushed into the ground. Because of time and/or technical constraints, the number of CPT in a specific site is often small. Besides, note that subsurface conditions are often inhomogeneous, and CPT at different locations may reveal different spatial variability of geotechnical properties in terms of accuracy. In this case, it is of great interest, but of great difficulty, to determine the optimal locations for CPT soundings such that as accurate as possible information on multi-layer geotechnical properties can be obtained. This is often encountered during the multi-stage geotechnical site characterization, and additional CPT locations are often needed in later site characterization. This paper presents an efficient Bayesian compressive sensing method for addressing this issue, which consists of two components: 1) information entropy for determination of optimal CPT locations, and 2) kronecker product to improve its computational efficiency given almost continuous CPT data. The method is demonstrated using numerical datasets. The results indicate that the locations determined by the presented method are effective and can properly characterize the spatial variability of multiple soil layers.

Keywords: Bayesian methods; non-parametric methods; data-driven method; site investigation optimization; non-stationary spatial variability

1 Introduction

Characterization of spatial variability of soil properties, i.e. variations of soil properties with depth and horizontal directions (e.g., Figure 1a), is one of important tasks to geotechnical site investigation. This is attributed to the fact that spatial variability of soil properties is a crucial factor affecting the performance of geo-structures, such as slopes (e.g., Griffiths et al. 2009; Liu et al. 2019) and foundations (e.g., Fenton and Griffiths 2002; Naghibi and Fenton 2017). This is often determined through laboratory tests or in-situ method, e.g., cone penetration test (CPT). Note that CPT is one of the most commonly used in-situ procedures for measuring the spatial variability of geotechnical parameters during site investigation, especially the spatial variability along the depth direction. In comparison with other in-situ methods, CPT can obtain almost continuous soil response data, i.e., soil resistance q_c and sleeve friction f_s, as its cone is pressed into the ground at a steady rate.

Although CPT offers numerous data points along the depth direction, the number of horizontal CPT soundings is typically small. Because subsurface site conditions are often inhomogeneous, different CPT locations may result in varying volumes of information gathered during site characterization (e.g., Jiang et al. 2017; Pinheiro et al. 2017; Yang et al. 2019). This raises the challenge of how to choose appropriate locations for a certain number of additional CPT soundings in order to acquire as much information as possible on the spatial variability of underlying soils. This is often encountered during the multi-stage geotechnical site characterization, and additional CPT locations are often needed in later site characterization. Although several methods have been proposed in literature, they become less applicable when the site of interest is of multilayer and nonstationary. For example, Jiang et al. (2017) recently proposed a method to identify the optimal borehole locations for slope stability assessment by combining information theory and Bayesian updating methods; Yang et al. (2019) performed similar work by utilizing hypothesis testing and/or conditional random field theory. Note that the above-mentioned approaches perform well in finding the ideal drilling locations that contribute the most to slope reliability analysis, they cannot be utilized directly for site characterization. In addition, it is important to

note that a reliable study or design of a geo-structure at a particular site is only possible after a geotechnical site assessment at that site (e.g., Zhao and Wang 2019). In this research, determining appropriate locations for CPT soundings may not be related to the engineering reaction of a geotechnical construction.

This paper presents a novel method for determining the optimal locations for additional CPT soundings, especially when a limited number of CPT soundings are available during preliminary site investigation. The suggested method systematically integrates information entropy with data-driven and non-parametric Bayesian compressive sensing (BCS). Note that, despite the fact that BCS has been developed to determine the optimal sampling locations for one-dimensional (1D) problems (Zhao and Wang 2019), the 1D method cannot be used directly for 2D problems due to the BCS formulation and computational efficiency in Zhao and Wang (2019) when dealing with a large number of data points for CPT data. After this introduction, the information entropy is briefly discussed in this work, followed by the development of proposed method for additional CPT soundings in a later stage of site investigation.

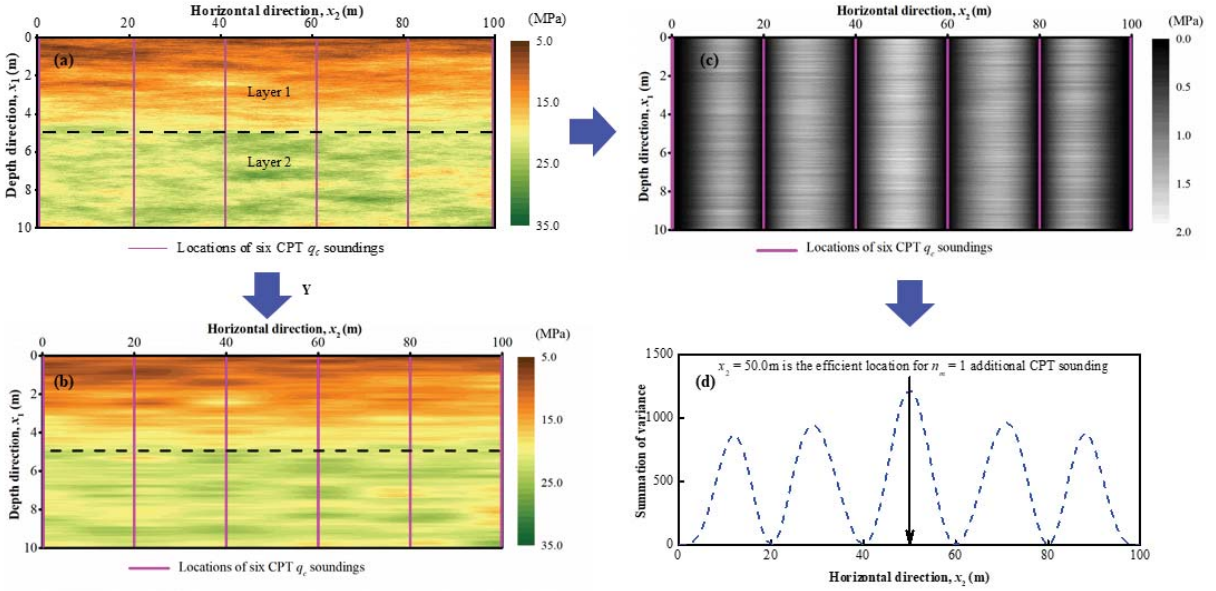

Figure 1. (a) An example of 2D spatially varying CPT tip resistance q_c data (b) Estimated 2D q_c profile from the six sets of q_c data; (c) Quantified uncertainty in terms of standard deviation; and (d)

2 Introduction to information entropy

Information entropy was introduced by Claude Shannon in 1948 as part of his communication theory (e.g., Shannon 1948). It is used to express the average quantity of information associated with a random process, which may be understood as a collection of numerous random variables. Intuitively, "entropy" implies uncertainty associated with the process: if the process is highly uncertain (i.e., random variables exhibit significantly big variability), it has a high information entropy; and if it is not uncertain, it has a low information entropy. If all random variables in a process are well-learned through measurement and there is no uncertainty associated with the process, the information entropy of the process approaches zero. The information entropy of a collection of random variables is defined as follows:

$$H(\boldsymbol{X}) = -\int p(\boldsymbol{X}) \ln[p(\boldsymbol{X})] d\boldsymbol{X} \tag{1}$$

where $H(\boldsymbol{X})$ and $p(\boldsymbol{X})$ respectively denote the information entropy and probability density function (PDF) of $\boldsymbol{X}$, respectively.

As Eq. (1) provides a quantitative measure of information, it is utilized to facilitate quantitative information learning for various applications. MacKay (1992) offered a way for adaptively selecting new data points that contribute most to the desired regression analysis; Ji et al. (2008) established a method for adaptively determining the ideal additional measurement in image recovery utilizing BCS and information entropy. Information entropy was also incorporated into global optimization algorithms to facilitate the search for objective function maxima and minima (e.g., Hennig and Schuler 2012). In spite of the complex mathematics involved in these applications, the fundamental concept was to seek out the new measurement that could potentially yield the most information gain for the challenge at hand.

Let $H(\boldsymbol{X}_{new})$ denote the new information entropy following the observation of new measurements. The data acquired from new measurements is then expressed as:

$$\Delta = H(\boldsymbol{X}) - H(\boldsymbol{X}_{new}) \tag{2}$$

Note that as new measurements are observed, the process's uncertainty is generally reduced, so $H(\boldsymbol{X}_{new})<H(\boldsymbol{X})$ and $\Delta > 0$. For geographic coordinate problems (e.g., the characterization of spatial variability in geotechnical engineering), different measurement locations may result in varying information gain. In order to gather as much information as possible from the new measurements, it is necessary to sample them at locations that maximize Eq. (2). Consequently, the information theory based on Eqs. (1) and (2) can be utilized to determine the optimal measurement locations when describing the spatial variability of soil parameters using CPT soundings, as detailed in the next section.

3 Optimal locations for additional CPT soundings

In this section, the limited number (e.g., n_b) of pre-existing CPT q_c data are utilized to find the ideal locations for additional CPT soundings, in order to further characterizing the spatial variability of the multilayer geotechnical properties using information entropy. As indicated by Eq. (1), the PDF of two dimensional (2D) spatially varying property by CPT, denoted as **F**, i.e., $p(\mathbf{F})$ is required for computing the information entropy, which may be determined using the Bayesian compressive sensing (BCS) method with n_b sets of observed CPT (e.g., q_c) data **Y** as input.

3.1 Efficient Bayesian compressive sensing for interpolation of 2D CPT data

Bayesian compressive sensing (BCS) is a novel sampling technique for reconstructing spatially or temporally changing signals from far fewer data than those required by the Nyquist-Shannon theorem (e.g., Candès and Wakin 2008; Ji et al. 2008). It has been utilized to describe the spatial variability of geotechnical parameters in 1D, 2D, and 3D scenarios (e.g., Wang and Zhao 2017; Zhao et al. 2018; Zhao and Wang 2020). In these applications, the signal decomposition idea that any geographically or temporally variable quantity, such as the 2D q_c data in Figure 1a, can be seen as a weighted sum of 2D basis functions is utilized. Let matrix **F** with dimensions $N_1 \times N_2$ represent the 2D q_c data to be described, and **F** is then expressed as

$$\mathbf{F} = \mathbf{B}_1^{1D}\boldsymbol{\Omega}(\mathbf{B}_2^{1D})^{\mathrm{T}} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\Omega_{i,j}\boldsymbol{b}_i^1(\boldsymbol{b}_j^2)^{\mathrm{T}} = \sum_{t=1}^{N}\mathbf{B}_t^{2D}\omega_t^{2D} \tag{3}$$

where $\mathbf{B}_1^{1D}$ and $\mathbf{B}_2^{1D}$ are two 1D orthonormal basis matrices (e.g., discrete cosine matrix) with dimensions $N_1 \times N_1$ and $N_2 \times N_2$, respectively; and $N = N_1 \times N_2$. As demonstrated in mathematical literature (e.g., Salomon 2007), the majority of ω_t elements are near to zero, with the exception of a limited number of non-trivial elements with notably large size. Once the non-trivial ω_t is accurately predicted, 2D q_c data with a high resolution (e.g., Figure 1a) can be produced from a small number of CPT records (i.e., **Y**) from a preliminary site study. For derivation convenience, transpose of **F**, i.e., $\mathbf{F}^{\mathrm{T}}$ is rewritten as a vector using Kronecker product which has been defined in Section 2 and decomposed as below:

$$vec(\mathbf{F}^{\mathrm{T}}) = (\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})vec(\boldsymbol{\Omega}^{\mathrm{T}}) = (\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})\boldsymbol{\omega}^{2\mathrm{D}} \tag{4}$$

where $\boldsymbol{\omega}^{2D}=vec(\boldsymbol{\Omega}^{\mathrm{T}})$ is a column vector denoting ω_t (t = 1, 2,..., N). As PDF of **F** is required when employing information entropy (see Eq. (1)), a Bayesian estimation approach is used (e.g., Zhao et al., 2020), from which PDF of **F** can be obtained using Eq (4). Using the approach presented by Zhao et al. (2020), the posterior PDF of $\boldsymbol{\omega}^{2D}$ is derived to follow a Gaussian distribution, with mean and covariance matrix expressed as:

$$\begin{aligned}
\boldsymbol{\mu}_{\hat{\boldsymbol{\omega}}^{2D}} &= (\boldsymbol{\mu}^1_{\hat{\boldsymbol{\omega}}^{2D}}, \quad \cdots, \quad \boldsymbol{\mu}^{N_1}_{\hat{\boldsymbol{\omega}}^{2D}})^{\mathrm{T}} \\
\mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}} &= \begin{bmatrix} \mathbf{COV}^1_{\hat{\boldsymbol{\omega}}^{2D}} & & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{COV}^{N_1}_{\hat{\boldsymbol{\omega}}^{2D}} \end{bmatrix}
\end{aligned} \tag{5}$$

where both $\boldsymbol{\mu}^i_{\hat{\boldsymbol{\omega}}^{2D}}$ and $\mathbf{COV}^i_{\hat{\boldsymbol{\omega}}^{2D}}$ are functions of measurement data (i.e., available CPT data from preliminary site investigation). Due to the introduction of kronecker product, Eq. (5) can be computed in an efficient manner. Readers of interest can be referred to Zhao and Wang (2020) and Zhao et al. (2020) for more details.

As **F** is a weighted sum of multivariate Gaussian random variables, it is easy to derive that the estimated 2D CPT dataset **F** follows a multivariate Gaussian distribution too (e.g., Ang and Tang 2007). The mean and covariance matrix for **F**, or equivalently $vec(\mathbf{F}^{\mathrm{T}})$, is represented as follows (e.g., Zhao et al., 2020):

$$\begin{aligned}\boldsymbol{\mu}_{vec(\hat{\mathbf{F}}^{\mathrm{T}})} &= (\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})\boldsymbol{\mu}_{vec(\boldsymbol{\Omega}^{\mathrm{T}})} = (\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})\boldsymbol{\mu}_{\hat{\omega}^{2D}} \\ \mathbf{COV}_{vec(\hat{\mathbf{F}}^{\mathrm{T}})} &= (\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})\mathbf{COV}_{\hat{\omega}^{2D}}(\mathbf{B}_1^{1D} \otimes \mathbf{B}_2^{1D})^{\mathrm{T}}\end{aligned} \tag{6}$$

$\boldsymbol{\mu}_{vec(\hat{\mathbf{F}}^{\mathrm{T}})}$ represents the estimate of **F**, while diagonal components of $\mathbf{COV}_{vec(\hat{\mathbf{F}}^{\mathrm{T}})}$ indicate the estimated variance of **F**. The entropy associated with **F** can then be estimated using Eq. (1), and the optimal locations for subsequent CPT soundings can be found appropriately, as demonstrated in the following subsection. It is important to note that the suggested BCS approach is data-driven and does not require a stationary assumption during calculation. The BCS approach can therefore be used to estimate spatially variable q_c values in multiple soil layers.

3.2 Optimal location for one additional CPT sounding

According to Eq. (1) and multivariate Gaussian PDF of $\hat{\mathbf{F}}$ (or $vec(\hat{\mathbf{F}}^{\mathrm{T}})$), information entropy $H(vec(\hat{\mathbf{F}}^{\mathrm{T}}))$ is expressed as:

$$H(vec(\hat{\mathbf{F}}^{\mathrm{T}})) = \ln[\det(\mathbf{COV}_{vec(\hat{\mathbf{F}}^{\mathrm{T}})})]/2 + N[1+\ln(2\pi)]/2 \tag{7}$$

where the term "N[1+ln(2π)]" is a constant term, which is independent of CPT sounding locations. Subsequently, combining Eqs. (6) and (7), rearranging the terms, lead to

$$H(vec(\hat{\mathbf{F}}^{\mathrm{T}})) = -1/2\sum_{i=1}^{N_1} \ln\det[\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\tau + \mathbf{D}_i^{\alpha}] + N[1+\ln(2\pi)]/2 \tag{8}$$

$\mathbf{A}_2 = \boldsymbol{\Psi}\mathbf{B}_2^{1D}$. Note that transpose of $\boldsymbol{\Psi}$, i.e., $\boldsymbol{\Psi}^{\mathrm{T}}$ reflects locations of n_b sets of the measured q_c data along the x_2 direction. $\mathbf{D}_i^{\alpha}$ (i = 1, 2, …, N_1) records the i-th N_2 intermediate unknown variables, which are determined in the BCS approach in previous subsection. If a new CPT were to be conducted and related q_c data were to be gathered along the depth, the 2D q_c would be updated as $\hat{\mathbf{F}}_{new}^{\mathrm{T}}$, and new information entropy is expressed as

$$H(vec(\hat{\mathbf{F}}_{new}^{\mathrm{T}})) = -1/2\sum_{i=1}^{N_1} \ln\det[(\mathbf{A}_2^{new})^{\mathrm{T}}(\mathbf{A}_2^{new})\tau^* + \mathbf{D}_i^{*\alpha}] + N[1+\ln(2\pi)]/2 \tag{9}$$

"τ^*" and "$\mathbf{D}_i^{*\alpha}$" are the new τ and $\mathbf{D}_i^{\alpha}$ given a new set of q_c data are obtained. Because the additional CPT has not been carried out and corresponding q_c data are unknown, "τ^*" and "$\mathbf{D}_i^{*\alpha}$" are taken as $\tau^* = \tau$ and $\mathbf{D}_i^{*\alpha} = \mathbf{D}_i^{\alpha}$ respectively hereafter.

It is worth pointing out that each row of $\mathbf{B}_2^{1D}$ corresponds to a CPT sounding location along the x_2 direction. As a result, when one additional CPT sounding is determined, $\mathbf{A}_2^{new}$ shall be updated by appending a new row $\boldsymbol{r}_2^1$ of $\mathbf{B}_2^{1D}$ to $\mathbf{A}_2$, i.e., $\mathbf{A}_2^{new} = [\mathbf{A}_2, \boldsymbol{r}_2^1]^{\mathrm{T}}$. In this case, $(\mathbf{A}_2^{new})^{\mathrm{T}}\mathbf{A}_2^{new}\tau + \mathbf{D}_i^{\alpha} = (\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\tau + \mathbf{D}_i^{\alpha}) + (\boldsymbol{r}_2^1)^{\mathrm{T}}\boldsymbol{r}_2^1\tau$. Following matrix determinant lemma (e.g., Brookes 2005), $\det[(\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\tau + \mathbf{D}_i^{\alpha}) + (\boldsymbol{r}_2^1)^{\mathrm{T}}\boldsymbol{r}_2^1\tau]$ is equal to $\det\{(\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\tau + \mathbf{D}_i^{\alpha})[1+\tau\boldsymbol{r}_2^1(\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\tau + \mathbf{D}_i^{\alpha})^{-1}(\boldsymbol{r}_2^1)^{\mathrm{T}}]\}$. With these expressions, information entropy before and after a new CPT sounding may be derived as below

$$\Delta = H[vec(\hat{\mathbf{F}}^{\mathrm{T}})] - H[vec(\hat{\mathbf{F}}_{new}^{\mathrm{T}})] = 1/2\sum_{i=1}^{N_1} \ln[1+\tau\boldsymbol{r}_2^1\mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}}^{i}(\boldsymbol{r}_2^1)^{\mathrm{T}}] \tag{10}$$

According to the discussion mentioned above, the optimal location for the additional CPT sounding is the one that maximizes Δ. Eq. (10) further shows that maximization of Δ is equivalent to the search for the row of $\mathbf{B}_2^{1D}$ (i.e., r_2) that maximizes $\sum_{i=1}^{N_1} \ln[1+\tau\boldsymbol{r}_2^1\mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}}^{i}(\boldsymbol{r}_2^1)^{\mathrm{T}}]$ or its exponential form $\left(N_1 + \tau\sum_{i=1}^{N_1}[\boldsymbol{r}_2^1\mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}}^{i}(\boldsymbol{r}_2^1)^{\mathrm{T}}]\right)$. ½ is omitted because it is a constant. Besides, because τ is a constant for all rows of $\mathbf{B}_2^{1D}$ and exponential function is monotonically increasing, the abovementioned problem further reduces to searching for the row of $\mathbf{B}_2^{1D}$ that maximizes $\sum_{i=1}^{N_1}[\boldsymbol{r}_2^1\mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}}^{i}(\boldsymbol{r}_2^1)^{\mathrm{T}}]$, which surprisingly coincides with the summation of the estimated variance of a CPT sounding along the x_2 direction (see Zhao et al. 2021 for a detailed proof). Such argument means that the optimal location for the additional CPT sounding along the x_2 direction is the one with the largest summation of variance, which intuitively make sense because the largest

variance indicates largest uncertainty at that location (e.g., Zhao and Wang 2019). Once the CPT sounding is carried out at the determined optimal location, the maximum information gain about the spatial variability is obtained, and the maximum uncertainty reduction on the estimated 2D q_c data is achieved.

3.3 Optimal location for multiple additional CPT sounding

In this subsection, the optimal locations for multiple (e.g., n_m) additional CPT soundings are determined using the efficient BCS, information entropy and the n_b sets of pre-existing q_c data $\mathbf{Y}$. Suppose that n_m additional CPT soundings are required in the multi-stage geotechnical site characterization. In this case, matrix $\mathbf{A}_2$ shall be updated to $\mathbf{A}_2^{new}$ by appending n_m rows of $\mathbf{B}_2$ to $\mathbf{A}_2$, and $\mathbf{A}_2^{new} = [\mathbf{A}_2, \mathbf{R}_2^{n_m}]^{\mathrm{T}}$, where $\mathbf{R}_2^{n_m}$ denotes the n_m rows of $\mathbf{B}_2$. Following a similar procedure, the information entropy of $\hat{\mathbf{F}}_{new}^{\mathrm{T}}$, is obtained and difference of information entropy between and after n_m sets of additional CPT soundings are derived as below

$$\Delta = H(vec(\hat{\mathbf{F}}^{\mathrm{T}})) - H(vec(\hat{\mathbf{F}}_{new}^{\mathrm{T}})) = 1/2 \sum_{i=1}^{N_1} \ln \det[\mathbf{I}_{n_m} + \tau \mathbf{R}_2^{n_m} \mathbf{COV}_{\hat{\boldsymbol{\omega}}^{2D}}^{i} (\mathbf{R}_2^{n_m})^{\mathrm{T}}] \tag{11}$$

where $\mathbf{I}_{n_m}$ is an identity matrix with a dimension of $n_m \times n_m$. Therefore, the seeking the optimal locations of the n_m additional CPT soundings is equivalent to seeking a combination of n_m rows of $\mathbf{B}_2$ (i.e., $\mathbf{R}_2^{\mathrm{nm}}$) that maximize Eq. (11), which can be readily formulated as an optimization problem and addressed via a built-in optimization function in some commercial software, e.g., "*ga*" function in MATLAB. A careful examination of Eq. (11) shows that it reduced exactly to Eq. (10) when n_m is reduced to $n_m = 1$.

4 Illustrative examples

In this section, a set of two-layer 2D CPT tip resistance q_c data is simulated for demonstration purpose, as shown in Figure 1a. The 2D q_c data are distributed over a vertical cross-section which is 100m long and 10m deep, with a resolution of 0.5m and 0.02m, respectively along the x_2 and x_1 direction (see Figure 1a). The two-layer 2D q_c data is simulated from a Gaussian simulator with a linearly varying mean, e.g., μ_1=0.05x_1+10 (MPa) in the first layer and a constant mean of μ_2 = 22.5 MPa in the second layer. The variances for q_c in these two layers are σ_1^2 =4.0 and σ_2^2 = MPa, respectively. Besides, an anisotropic exponential correlation structure is adopted when simulating the 2D q_c data:

$$\rho = \exp(-2\sqrt{\frac{\Delta x_1}{\lambda_{1,k}} + \frac{\Delta x_2}{\lambda_{2,k}}}) \ (k = 1, 2) \tag{12}$$

where $\lambda_{1,1}$ = 1.0m and $\lambda_{2,1}$ = 50.0 in the first layer; while $\lambda_{1,2}$ = 1.0m and $\lambda_{2,2}$ = 30.0m in the second layer. Suppose that n_b = 6 sets of CPT soundings were carried out in the preliminary site investigation and corresponding q_c data are recorded. With the discussions mentioned above, these data can be effectively utilized to determine the optimal locations for additional CPT soundings to efficiently characterize the spatial variability of geotechnical properties.

In this subsection, the optimal location for the additional CPT sounding is determined by the presented method together with the n_b = 6 sets of pre-existing CPT data. As discussed in detail previously, the optimal location is the x_2 location with the largest summation of variances estimated from the efficient BCS method. For the current example with n_b = 6 sets of q_c data, 2D CPT data can be obtained using the presented BCS method, as shown in Figure 1b&1c. With the estimated standard deviation in Figure 1c, the optimal location for one additional CPT soundings is shown in Figure 1d by a dashed line, which is located at x_2 = 50.0m.

Table 1 Performance of the optimal locations for n_m (= 1, 2, 3, 4 and 6) additional CPT soundings from the proposed method and that of randomly selected locations for 1000 times in interpreting 2D q_c data in terms of mean absolute error (MAE)

Number of additional CPT soundings	MAE with optimal locations from the proposed method	Statistics of the MAEs with 1000 sets of n_m random locations		
		Minima	5th percentile	25th percentile
n_m = 1	0.98	0.97	0.98	0.99
n_m = 2	0.91	0.86	0.94	1.02
n_m = 3	0.84	0.79	0.89	0.96
n_m = 4	0.77	0.77	0.86	0.93
n_m = 6	0.72	0.73	0.80	0.86

To explore the effectiveness of the location determined, CPT values at this location are retrieved from the underlying true 2D q_c data and used with the n_b = 6 sets of pre-existing q_c data as input to the BCS method to

estimate probabilistically the underlying 2D q_c data. With the results obtained, the mean absolute error (MAE) between the underlying true 2D qc data and the estimated one from the presented method is obtained, which is computed as MAE = 0.98, which is smaller when comparing the MAE = 1.05 in the n_b = 6 scenario (the results in the previous paragraph). For a systematical investigation, effectiveness of the additional q_c data at x_2 = 50.0m is also compared to that of q_c data at other 201-7 = 194 x_2 locations. q_c data at each of the 194 x_2 locations with the n_b = 6 pre-existing q_c data (i.e., the dashed lines in Figure 1d) are utilized together as input to the efficient BCS approach to infer the 2D q_c data in this two-layer soil, followed by calculation of MAE. 194 x_2 locations therefore lead to 194 MAEs in total, using which the minima, 5th and 25th percentiles are determined and described in Table 1. Table 1 also shows MAE = 0.98 corresponding to the n_b = 7 scenario with q_c data at the optimal location, i.e., x_2 = 50.0m. Table 1 shows that MAE = 0.98 is equivalent to the minima (i.e., 0.97) and the 5th percentile (i.e., 0.98) of the 194 MAEs, showing again that the optimal location (i.e., x_2 = 50.0m) identified by the proposed method is effective in defining the spatial variability of soil properties.

Similar procedure is applied to determine the optimal locations for multiple (e.g., n_m = 2, 3, 4, 6) additional CPT soundings, and their performance is also examined in terms of MAE and the performance of 1000 times of randomly selection of CPT locations, as summarized in Table 1. Table 1 shows that the MAE corresponding to the optimal locations determined from the presented method is very close to the minima of the 1000 random experiments, demonstrating that the presented method is very effective in determining optimal locations for additional CPT soundings to efficiently characterize the spatially varying subsurface soils.

5 Conclusions

This paper presented a novel approach to determine the optimal locations for additional CPT soundings by effectively utilizing the information from a limited number of CPT data in the preliminary site investigation. The presented approach is developed based on the information entropy theory and non-parametric Bayesian compressive sensing (BCS) approach. Results show that the optimal location for one CPT sounding is the one with the largest summation of variances from the BCS. Numerical examples were taken to carefully evaluate the presented approach, which indicated (1) that the presented approach performs well in determining optimal locations for additional CPT soundings, and (2) that the BCS approach is applicable to characterize non-stationary soil property within multilayers due to its data-driven characteristics.

Acknowledgments

This work described in this paper was supported by grants from the National Natural Science Foundation of China (Project No. 42107204), and the Fundamental Research Funds for the Central Universities (xjh012020046). The financial supports are gratefully acknowledged.

References

Ang, A. & Tang, W.H. (2007). Probability concepts in engineering: emphasis on applications to civil & environmental engineering. 2nd ed. *Wiley, New York*.

Candès, E.J. & Wakin, M.B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25, 21-30.

Fenton, G.A. & Griffiths, D.V. (2002). Probabilistic foundation settlement on spatially random soil. *Journal of Geotechnical and Geoenvironmental Engineering*, 128, 381-390.

Griffiths, D., Huang, J. & Fenton, G.A. (2009). Influence of spatial variability on slope reliability using 2-D random fields. *Journal of Geotechnical and Geoenvironmental Engineering*, 135, 1367-1378.

Hennig, P. & Schuler, C.J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13, 1809-1837.

Ji, S., Xue, Y. & Carin, L. (2008). Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56, 2346-2356.

Jiang, S.-H., Papaioannou, I. & Straub, D. (2017). Optimizing Borehole Locations for Slope Reliability Assessment. *In: Huang, J., Fenton, G.A., Zhang, L. & Griffiths, D.V. (eds.) Geo-Risk 2017: Geotechnical Risk from Theory to Practice. ASCE, Denver*, 420-430.

Liu, L.-L., Deng, Z.-P., Zhang, S.-h. & Cheng, Y.-M. (2018). Simplified framework for system reliability analysis of slopes in spatially variable soils. *Engineering Geology*, 239, 330-343.

MacKay, D.J. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 590-604.

Naghibi, F. & Fenton, G.A. (2017). Target geotechnical reliability for redundant foundation systems. *Canadian Geotechnical Journal*, 54, 945-952.

Pinheiro, M., Emery, X., Rocha, A.M.A., Miranda, T. & Lamas, L. (2017). Boreholes plans optimization methodology combining geostatistical simulation and simulated annealing. *Tunnelling and Underground Space Technology*, 70, 65-75.

Salomon, D. (2007). Data compression: the complete reference. *Fourth Edition ed. Springer Science & Business Media, New York, USA*.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27, 379-423.

Wang, Y. & Zhao, T. (2017). Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique*, 67, 523-536.

Yang, R., Huang, J., Griffiths, D.V., Meng, J. & Fenton, G.A. (2019). Optimal geotechnical site investigations for slope design. *Computers and Geotechnics*, 114, 103111.

Zhao, T., Hu, Y. & Wang, Y. (2018). Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. *Engineering Geology*, 246, 162-175.

Zhao, T. & Wang, Y. (2019). Determination of efficient sampling locations in geotechnical site characterization using information entropy and Bayesian compressive sampling. *Canadian Geotechnical Journal*, 56, 1622-1637.

Zhao, T. & Wang, Y. (2020). Non-parametric simulation of non-stationary non-gaussian 3D random field samples directly from sparse measurements using signal decomposition and Markov Chain Monte Carlo (MCMC) simulation. *Reliability Engineering & System Safety*, 203, 107087.

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. *Engineering Geology*, 273, 105670.

Zhao, T., Wang, Y., & Xu, L. (2021). Efficient CPT locations for characterizing spatial variability of soil properties within a multilayer vertical cross-section using information entropy and Bayesian compressive sensing. *Computers and Geotechnics*, 137, 104260.