Canadian Geotechnical Journal

Determination of Efficient Sampling Locations in Geotechnical Site Characterization Using Information Entropy and Bayesian Compressive Sampling

Journal:	<i>Canadian Geotechnical Journal</i>
Manuscript ID	cgj-2018-0286.R3
Manuscript Type:	Article
Date Submitted by the Author:	27-Nov-2018
Complete List of Authors:	Zhao, Tengyuan; City University of Hong Kong, Architecture and civil engineering Wang, Yu; City University of Hong Kong, Dept of Civil and Architectural Engineering
Keyword:	Site investigation, Spatial variability, Compressive sensing, Bayesian method, Statistical method
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

SCHOLARONE™
Manuscripts

Determination of Efficient Sampling Locations in Geotechnical Site Characterization Using Information Entropy and Bayesian Compressive Sampling

Tengyuan Zhao¹ and Yu Wang²

¹Research Associate, Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. (Tel): 852-3442-2317 (Fax): 852-3442-0427 (Email): tengyzhao2@cityu.edu.hk

²Associate Professor, Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. (Tel): 852-3442-7605 (Fax): 852-3442-0427 (Email): yuwang@cityu.edu.hk (Corresponding author)

Abstract

Site characterization is indispensable in geotechnical engineering practice, and measurements on soil properties are performed through in-situ tests, laboratory tests, or other methods. However, due to time or budget limit, technical or access constraints etc., the measurements are usually taken at a limited number of locations. This leads to a question of how to select the efficient locations for measurements/sampling such that as much as possible information on the spatial variability of soil properties can be obtained from a given number of measurements. In addition, geotechnical site characterization is a multi-stage process, and additional measurements might be required at a later stage of site characterization. In this case, how to efficiently select the additional sampling locations such that the pre-existing measurements obtained from the preliminary stages of site characterization can be best used and as much as possible information on soil properties can be further obtained? This paper aims to address these two problems using information entropy and Bayesian compressive sampling (BCS). Real cone penetration test data along both vertical and horizontal directions are used to illustrate and validate the proposed methods. The results show that the proposed methods are very effective and robust in selecting efficient sampling locations for geotechnical site characterization.

Key words: Site investigation; Spatial data; Spatial variability; Compressive sensing; Bayesian method;

Introduction

Site characterization is indispensable in geotechnical engineering practice, and it aims at delineating spatial distribution of underground soils and rocks in a project site and estimating soil properties for geotechnical analysis and design through in-situ tests, laboratory tests, or other methods (e.g., Mayne et al. 2002; Wang et al. 2016). Soils are natural geo-materials, whose properties are affected by many spatially varying factors during the geological process, such as the properties of their parent materials, weathering and erosion processes, transportation agents, and sedimentation conditions (e.g., Lacasse and Nadim 1996; Phoon and Kulhawy 1999; Webster 2000; Baecher and Christian 2003). Therefore, soil properties exhibit spatial variability, proper quantification of which generally requires extensive measurements at different locations (e.g., Vanmarcke 1977; Fenton 1999; Wang et al. 2017). In geotechnical engineering practice, however, the number of site-specific measurements is usually limited (e.g., Wang and Cao 2013a; Wang and Akeju 2016; Phoon 2017; Zhao et al. 2018), and soil properties are often sparsely measured at a given limited number of locations, due to time or budget limit, technical or access constraints, etc. This leads to a question of how to efficiently select the locations of measurements or sampling such that as much as possible information on the spatial variability of soil properties can be obtained from a given number of measurements.

Determination of the efficient sampling locations is a common topic that has wide applications in various fields, such as agricultural engineering (e.g., van Groenigen et al. 1999; Lark 2002), environmental engineering (e.g., Cox 1999; Back 2000; Stein and Ettema 2003; Johnson et al. 2005), geosciences (e.g., McBratney and Webster 1981; McBratney et al. 1981; Cressie 1991), geotechnical engineering (e.g., Halim and Tang 1993; Jiang et al. 2017, 2018; Lo and Leung 2018), among others. For example, van Groenigen et al. (1999) obtained the efficient locations of measurements on spatial data by minimizing estimation variance of

a kriging estimator with a pre-specified variogram model. Cox (1999) proposed to adaptively locate sampling points using pre-existing measurements through a geostatistical method. Lo and Leung (2018) used extended Sobol's index to determine locations of measurements which maximize uncertainty reduction in the factor of safety calculated from slope stability assessment. Note that these methods generally utilize a correlation structure (e.g., a semi-variogram model or auto-correlation function form) with parameters (e.g., range in a semi-variogram model or correlation length in an auto-correlation function) estimated from extensive measurements. In geotechnical site characterization, no site-specific measurement data are available before site investigation. Even if there are some site-specific measurements obtained from preliminary stage of a site investigation, the number of measurements is usually sparse and limited (e.g., Phoon 2017). In such cases, it is difficult to reliably estimate the required correlation structure and parameters (e.g., range or correlation length) (e.g., Wang et al. 2017; Wang et al. 2018; Zhao and Wang 2018a) and determine the efficient sampling locations for measurements using the methods mentioned above.

In engineering practice, the sampling locations are often determined empirically using engineer experience or judgment. For example, if only one measurement (e.g., Atterberg's limit test or triaxial test) is to be made on a soil sample taken from a borehole within a soil layer, the soil sample is usually taken at the mid-depth of the soil layer. If multiple measurements are performed using multiple soils samples taken at different depths for delineating spatial variability of soil properties, the soil samples are often taken using an equal sampling interval throughout the whole soil layer of interest. However, there is little scientific evidence available to support the empirical methods described above. In addition, site characterization is a multi-stage process (e.g., Mayne et al. 2002; Wang et al. 2016), and additional measurements might be required at a later stage of geotechnical site characterization. This leads to another question of how to efficiently select the sampling

locations in a later stage such that the pre-existing measurements, often sparse and limited though, obtained from the preliminary stages of site characterization can be best used and as much as possible information on the spatial variability of soil properties can be further obtained from the additional measurements.

This study aims to address the questions mentioned above and develop a quantitative and objective method for determination of efficient sampling locations in geotechnical site characterization, using information entropy or/and Bayesian compressive sampling (BCS) (e.g., Wang and Zhao 2017). Both the concept of information entropy and BCS have been developed from previous studies, mainly in electronic engineering and applied mathematics. However, it is the first time that these two methods are systematically integrated and tailored to address an important and long-lasting problem of sampling location determination in geotechnical site characterization. This study focuses on one-dimensional cases, such as depths for measurements on soil samples taken from a borehole within a soil layer or horizontal coordinates for measurements on soil samples taken from multiple boreholes at a given depth. It is feasible to extend the study to two-dimensional (2D) cases to simultaneously consider the measurement locations at both vertical and horizontal directions, such as optimization of the locations of different boreholes (e.g., Jiang et al. 2018). Such an extension, however, requires development of 2D BCS, which involves construction of 2D basis functions, identification of the non-trivial 2D basis functions and estimation of their corresponding coefficients (Zhao et al. 2018).

After this introduction, the concept of information entropy is firstly reviewed, followed by formulation of the information entropy-based method for determination of the efficient sampling locations without pre-existing measurement data. When there are some sparse measurements obtained from preliminary stages of site characterization, the information entropy-based method is integrated with BCS to determine the efficient sampling

locations for additional measurements. The proposed method is illustrated and validated using real cone penetration test (CPT) tip resistance q_c data along vertical and horizontal directions.

Information entropy

Information entropy is a concept in information theory, and it refers to the average amount of information conveyed in a random process, which is a collection of many random variables (e.g., MacKay 2003). The term “entropy” denotes disorder or uncertainty in a random process. Generally speaking, if a random process exhibits large variability (i.e., the process is very uncertain), it has a large entropy; and vice versa. If all random variables of a random process are well learned and the process contains no uncertainty, entropy of the process is zero. Mathematically, information entropy H of a random vector $\mathbf{X}$ with a length of N is expressed as (e.g., MacKay 2003):

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

where $H(\mathbf{X})$ represents information entropy of $\mathbf{X}$; and $p(\mathbf{X})$ represents probability density function (PDF) of $\mathbf{X}$. The concept of information entropy has been used in experimental design (e.g., MacKay 1992), Bayesian optimization (e.g., Hennig and Schuler 2012), and objective determination of the optimal locations for additional measurements in signal processing (e.g., Ji et al. 2008). The basic idea behind these applications is to take measurements at locations where the information gain on $\mathbf{X}$ (i.e., reduction in $H(\mathbf{X})$ after the measurements) is or approaches to maxima (e.g., Ji et al. 2008). Note that $H(\mathbf{X})$ quantifies the uncertainty associated with $\mathbf{X}$, and it is relatively large when information on $\mathbf{X}$ is limited (e.g., only have sparse, or even no, measurements). $H(\mathbf{X}_{new})$ represents the new information entropy after additional information (e.g., observing new measured values) and it quantifies

the uncertainty in the new $\mathbf{X}$ (i.e., $\mathbf{X}_{new}$), updated by the new measurements. Generally speaking, as information is gained through an increasing number of measurements, the corresponding uncertainty gradually reduces (i.e., $H(\mathbf{X}_{new}) < H(\mathbf{X})$). The difference Δ between $H(\mathbf{X})$ and $H(\mathbf{X}_{new})$ quantifies the information gain or equivalently the uncertainty reduction in $\mathbf{X}$ by the new measurements. Note that different sampling locations for measurements may lead to different $H(\mathbf{X}_{new})$ and Δ . Therefore, to gain as much as possible information (i.e., to reduce as much as possible the uncertainty) from the new measurements, the sampling locations for the new measurements are proposed to be those that maximize the difference Δ between $H(\mathbf{X})$ and $H(\mathbf{X}_{new})$:

$$(2) \quad \Delta = H(\mathbf{X}) - H(\mathbf{X}_{new})$$

In the context of geotechnical site characterization, soil property varies spatially (e.g., along depths or a horizontal direction), and spatial variability of soil property (e.g., a soil property profile along depth or a horizontal direction) may be modeled as a random process (e.g., Webster 2000; Baecher and Christian 2003; Vanmarcke 2010; Wang et al. 2010), especially when there are no measurement data available or when the number of measurements is sparse and limited (e.g., Stuedlein et al. 2012; Wang and Cao 2013b; Li et al. 2016). Therefore, the information entropy-based method described by Eqs. (1) and (2) may be used to determine the efficient sampling locations in geotechnical site characterization, as discussed in the following two sections. It is worth noting that the principle of maximum entropy has been successfully used in geotechnical engineering before, such as identification of underground soil stratification using CPT data (Wang et al. 2013).

Efficient sampling locations for a given number of measurements

In this section, the information entropy is used to determine the efficient sampling locations for a given number, M , of measurements without pre-existing measurement data. As discussed in the previous section, a soil property profile along depth may be modeled as a random process, such as a Gaussian random process, with an auto-correlation structure, such as an exponential correlation function. A single exponential correlation function has been widely used in geotechnical engineering (e.g., Cao and Wang 2013; Zhu and Zhang 2013; Fenton et al. 2018), and it is written as:

$$(3) \quad \rho_{i,j} = \exp\left[-\frac{2|z_i - z_j|}{\lambda_c}\right] = \exp\left[-\frac{2d_{i,j}}{\lambda_c}\right]$$

where $\rho_{i,j}$ represents the correlation between soil properties at different locations along a spatial direction, such as a horizontal direction or at depths z_i and z_j and $0 \leq \rho_{i,j} \leq 1$; $d_{i,j}$ is the distance between z_i and z_j ; λ_c is the correlation length within which soil properties are strongly correlated.

As discussed in the previous section, the efficient sampling locations for a given number of measurements are those that maximize the difference between the current information entropy $H(\mathbf{X})$ of the random process (e.g., the spatially varying soil property profile) and the new entropy $H(\mathbf{X}_{new})$ after the measurements. Both $H(\mathbf{X})$ and $H(\mathbf{X}_{new})$ are unknown when there are no pre-existing measurement data. In this case, the difference in information entropy (i.e., Eq. (2)) cannot be evaluated directly. However, the difference can be approximated as the information entropy associated with the given number, M , of measurements $\mathbf{y}$, i.e., $\Delta = H(\mathbf{X}) - H(\mathbf{X}_{new}) \approx H(\mathbf{y})$. Note that the measurements $\mathbf{y}$ may also be modeled as an M -length Gaussian random process before measuring them (e.g., Webster

2000), as described in the previous paragraph. Then, the information entropy of $\mathbf{y}$, i.e., $H(\mathbf{y})$, can be expressed as (e.g., MacKay 1992):

$$(4) \quad H(\mathbf{y}) = \frac{1}{2} \ln[\det(\mathbf{C}_y)] + \frac{M}{2} [1 + \ln(2\pi)]$$

where $\mathbf{C}_y$ is an $M \times M$ correlation matrix, with the (i, j) element $C_y(i, j) = \rho_{i,j}$ ($i, j = 1, 2, \dots, M$). $\rho_{i,j}$ represents the correlation among elements of $\mathbf{y}$ and it is calculated from Eq. (3), and $\rho_{i,j} = 1$ when $i = j$ ($i, j = 1, 2, \dots, M$). Therefore, the efficient locations for a given number of measurements are those that maximize $H(\mathbf{y})$ calculated by Eq. (4). Note that $H(\mathbf{y})$ in the left-hand side of Eq. (4) is maximized when the determinant of $\mathbf{C}_y$ (i.e., $\det(\mathbf{C}_y)$) in the right-hand side of Eq. (4) reaches its maxima, because $\frac{M}{2} [1 + \ln(2\pi)]$ is a constant for a given M value (e.g., Ko et al. 1995). $\det(\mathbf{C}_y)$ may be written as (e.g., Strang 2006):

$$(5) \quad \det(\mathbf{C}_y) = \det \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,M} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,1} & \rho_{M,2} & \cdots & 1 \end{bmatrix} = \sum_{i=1}^M (-1)^{i+j} \rho_{i,j} \det(\mathbf{C}_y^{-i,-j})$$

where $\mathbf{C}_y^{-i,-j}$ represents a reduced matrix by eliminating the i -th row and the j -th column of $\mathbf{C}_y$. The following paragraphs will show that $\det(\mathbf{C}_y)$ and $H(\mathbf{y})$ reach their maxima when the M measurements are taken with an equal interval throughout the whole layer of interest, i.e., $d_{1,2} = d_{2,3} = \dots = d_{M-1,M} = h/(M-1)$, where h is thickness of the concerned soil layer.

Consider, for example, $M = 4$. Then, $\rho_{i,j}$ ($i, j = 1, 2, \dots, 4$) may be calculated from Eq. (3), and they are substituted into Eq. (5). Rearranging the terms of Eq. (5) leads to:

$$(6) \quad \det(\mathbf{C}_y) = 1 + \exp[-4(d_{1,2} + d_{2,3})/\lambda_c] + \exp[-4(d_{1,2} + d_{3,4})/\lambda_c] + \exp[-4(d_{2,3} + d_{3,4})/\lambda_c] \\ - \exp[-4d_{1,2}/\lambda_c] - \exp[-4d_{2,3}/\lambda_c] - \exp[-4d_{3,4}/\lambda_c] - \exp[-4(d_{1,2} + d_{2,3} + d_{3,4})/\lambda_c]$$

It is obvious from Eq. (6) that $\det(\mathbf{C}_y)$ is a function of the variables $d_{1,2}$, $d_{2,3}$ and $d_{3,4}$. Therefore, finding the maxima of $\det(\mathbf{C}_y)$ may be formulated as an optimization problem with an objective of maximizing $\det(\mathbf{C}_y)$, or for numerical convenience, minimizing the negative logarithm of $\det(\mathbf{C}_y)$, i.e., $-\ln[\det(\mathbf{C}_y)]$. The optimization problem is expressed as:

Optimization objective: Minimize $-\ln[\det(\mathbf{C}_y)]$

Optimization variables: $d_{1,2}$, $d_{2,3}$ and $d_{3,4}$

$$\text{Constraint: } \begin{cases} 0 < d_{1,2} < h \\ 0 < d_{2,3} < h \\ 0 < d_{3,4} < h \\ 0 < d_{1,2} + d_{2,3} + d_{3,4} \leq h \end{cases}$$

Note that $d_{1,2}$, $d_{2,3}$ and $d_{3,4}$ are all greater than zero and less than h due to the physical meaning of distance between locations of different measurements.

The optimization problem mentioned above can be readily solved by an optimization function in some commercial software, such as “fmincon” in MATLAB, once the h and λ_c are specified. For example, Figure 1 shows the maximum determinant obtained from numerical optimization using the MATLAB “fmincon” function (MathWorks 2018), denoted as $\det(\mathbf{C}_y)_N$, for the $M = 4$ cases by open circles. Various combinations of soil layer thickness h and correlation length λ_c are adopted for the results shown in Figure 1, including $h = 1\text{m}, 3\text{m}, 5\text{m}, 7\text{m}, 10\text{m}, 20\text{m}, 35\text{m}$ and 50m and $\lambda_c = 0.5\text{m}, 1\text{m}, 2\text{m}, 3\text{m}, 5\text{m}, 10\text{m}, 15\text{m}$ and 20m . The range of the soil layer thickness h and vertical correlation length λ_c adopted in Figure 1 are consistent with those commonly encountered in engineering practice and reported in literature (e.g., Phoon and Kulhawy 1999). Figure 1 also includes the $\det(\mathbf{C}_y)$ calculated analytically for the case of measurements with an equal sampling interval, denoted as $\det(\mathbf{C}_y)_E$, for $M = 4$ (i.e., 4 measurements with an equal sampling interval of $h/(M-1) = h/3$)

and various combinations of h and λ_c , by a solid line. It is evident in Figure 1 that all open circles are overlapped with the solid line, indicating that $\det(\mathbf{C}_y)_N = \det(\mathbf{C}_y)_E$. This suggests that information entropy $H(\mathbf{y})$ and $\det(\mathbf{C}_y)$ are maximized when the given number of measurements are taken with an equal sampling interval of $h/3$ for $M=4$.

The optimization formulation for the $M = 4$ case described above is also repeated for many different M values, such as $M = 1, 4, 7, 10, 15, 20, 25, 30, 35, 40, 45$ and 50 . Both the numerically optimized determinant $\det(\mathbf{C}_y)_N$ and the $\det(\mathbf{C}_y)_E$ obtained analytically for measurements with an equal sampling interval of $h/(M-1)$ are calculated for each M value. Figure 2 shows a ratio $r = [\det(\mathbf{C}_y)_N]/[\det(\mathbf{C}_y)_E]$ calculated for all combinations of M , h and λ_c by solid balls. It is evident in Figure 2 that all solid balls fall on a horizontal surface of $r=1$. This suggests that $\det(\mathbf{C}_y)_N = \det(\mathbf{C}_y)_E$ for a large number of combinations of M , h and λ_c and that the $\det(\mathbf{C}_y)$ reaches its maxima when the given M number of measurements are taken with an equal sampling interval of $h/(M-1)$, at least for the typical ranges of M , h and λ_c in engineering practice of geotechnical site characterization. In addition, note that the optimization formulation described above is only applicable for $M > 2$. For $M = 1$, the distance between measurements (e.g., $d_{1,2}$) does not exist; and for $M = 2$, there is only one distance (i.e., $d_{1,2}$). To use the optimization formulation described above, two virtual points at the top and bottom boundaries of the soil layer (i.e., depth $z = 0$ and $z = h$) are added when $M = 1$ or 2 . The efficient sampling location obtained is the mid-depth (i.e., $z = h/2$) for $M=1$; while $z_1 = h/3$ and $z_2 = 2h/3$ are the efficient sampling locations for $M = 2$. The results for $M = 1$ or 2 are consistent with those for $M > 2$. In other words, the efficient sampling locations for a given M number of measurements correspond to measurements with an equal sampling interval through the whole soil layer. This conclusion is slightly different from the previous study by Degroot and Baecher (1993), where “nested sampling scheme” was proposed as the

most efficient one, because of the different objectives of the two studies. In Degroot and Baecher (1993), the objective is to achieve the minimum variances in the estimated random field parameters, particularly correlation length; whereas this study aims to maximize information gain on the variation of soil property along a spatial direction (e.g., depth or a horizontal direction), and hence, provide better prediction and less uncertainty at un-sampled locations.

In summary, the discussion in this section numerically demonstrates that the determinant $\det(\mathbf{C}_y)$ and information entropy $H(y)$ obtain their maxima when the the given M number of measurements are taken with an equal sampling interval throughout the whole soil layer of interest. This is consistent with the engineer experience and judgment in geotechnical engineering practice, as mentioned under the “Introduction” section. The formulation and sensitivity study that covers the typical λ_c range in geotechnical practice in this section provide quantitative results and scientific evidence to support the engineering practice of taking measurements with an equal sampling interval in geotechnical site characterization. After the efficient sampling locations for measurements y are determined, measurements are then taken at these locations and the numerical vaules of y are obtained after the measurements (e.g., Atterberg’s limit test or triaxial test). If additional measurements are needed in a later stage of site characterization, the measurement data y (i.e., both measurement locations and measured values) may be used to determine efficient sampling locations for additional measurements, as discussed in the next section.

Efficient sampling locations for additional measurements

In this section, the information entropy and pre-existing measurements y (i.e., both sampling locations and measured values) are used to determine the efficient sampling locations for additional measurements. No assumed correlation structure is needed in the proposed method.

Note that with the pre-existing measurement data $\mathbf{y}$ on soil properties, the PDF of the soil property profile concerned can be obtained using BCS (e.g., Wang and Zhao 2017), and the information entropy of the soil property profile concerned can be calculated directly (See Eq. (1)). A brief review on BCS is provided in the following subsection.

Bayesian compressive sampling (BCS)

Bayesian compressive sampling or sensing (BCS) is a probabilistic method to reconstruct a signal from partial measurements on that signal (e.g., Ji et al. 2008; Ji et al. 2009; Wang and Zhao 2017). A signal is defined as a variation of a quantity with time or space, e.g., a soil property profile, and it is denoted as a real-valued column vector $\mathbf{f}$ with a length of N . In the context of BCS, a signal $\mathbf{f}$ can be concisely represented by a weighted summation of a limited number of basis functions (e.g., Daubechies 16 wavelet basis functions). Mathematically, $\mathbf{f} = \mathbf{B}\boldsymbol{\omega}$, where $\mathbf{B}$ is an $N \times N$ orthonormal matrix, each column of which represents the basis function with different orders; and $\boldsymbol{\omega}$ is an N -length column vector, representing N weights corresponding to N columns of $\mathbf{B}$. The matrix $\mathbf{B}$ is independent of measurement data and can be readily constructed using some online freeware, such as Wavelab850 (Donoho et al. 2006). In the context of compressive sensing (e.g., Wang and Zhao 2016; Ji et al. 2008), most elements of $\boldsymbol{\omega}$ are very close to zero except several non-trivial ones. Therefore, the soil property profile $\mathbf{f}$ can be reconstructed from a limited number of measurements $\mathbf{y}$ if the non-trivial elements of $\boldsymbol{\omega}$ can be estimated using $\mathbf{y}$ through $\mathbf{y} = \boldsymbol{\Psi}\mathbf{f} = \boldsymbol{\Psi}\mathbf{B}\boldsymbol{\omega} = \mathbf{A}\boldsymbol{\omega}$, where $\mathbf{A} = \boldsymbol{\Psi}\mathbf{B}$; and $\boldsymbol{\Psi}$ is a measurement matrix that represents the locations of all components of $\mathbf{y}$ in $\mathbf{f}$. Using $\mathbf{y} = \mathbf{A}\boldsymbol{\omega}$, $\boldsymbol{\omega}$ can be estimated and denoted as $\boldsymbol{\omega}_s$ deterministically (e.g., Wang and Zhao 2016; Zhao and Wang 2018b) or probabilistically (e.g., Ji et al. 2008; Ji et al. 2009; Wang and Zhao 2017). $\boldsymbol{\omega}_s$ has the same length as $\boldsymbol{\omega}$, and all elements of $\boldsymbol{\omega}_s$ are set to zero, except the non-trivial ones. Wang and Zhao (2017) estimated $\boldsymbol{\omega}_s$ probabilistically under a

Bayesian framework, and they showed that $\boldsymbol{\omega}_s$ follows a multivariate Student's t distribution, with a degree of freedom of $2c_n$. The mean $\boldsymbol{\mu}_{\omega_s}$ and covariance matrix $\mathbf{COV}_{\omega_s}$ of $\boldsymbol{\omega}_s$ are expressed as (Wang and Zhao 2017):

$$(7) \quad \begin{aligned} \boldsymbol{\mu}_{\omega_s} &= \mathbf{H}\mathbf{A}^T \mathbf{y} = (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{A}^T \mathbf{y} \\ \mathbf{COV}_{\omega_s} &= \frac{d_n \mathbf{H}}{c_n - 1} = \frac{d_n (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1}}{c_n - 1} \end{aligned}$$

where the superscript “T” represents transpose operation; $\mathbf{H} = (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1}$, $c_n = M/2 + c$; $d_n = d + (\mathbf{y}^T \mathbf{y} - \boldsymbol{\mu}_{\omega_s}^T \mathbf{H}^{-1} \boldsymbol{\mu}_{\omega_s})/2$; c and d are small non-negative constants (e.g., $c = d = 10^{-4}$); $\mathbf{D}$ is a diagonal matrix with $D_{i,i} = \alpha_i$ ($i = 1, 2, \dots, N$), where α_i ($i = 1, 2, \dots, N$) are unknown non-negative parameters, which can be estimated by maximizing the likelihood of observing $\mathbf{y}$ (e.g., Tipping 2001; Ji et al. 2008; Wang and Zhao 2017). Because $\boldsymbol{\omega}_s$ follows a multivariate Student's t distribution, the reconstructed signal $\hat{\mathbf{f}} = \mathbf{B}\boldsymbol{\omega}_s$ also follows a multivariate Student's t distribution (e.g., Taboga 2012; Wang et al. 2018; Zhao et al. 2018) with a degree of freedom of $2c_n$, mean $\boldsymbol{\mu}_{\hat{\mathbf{f}}} = \mathbf{B}\boldsymbol{\mu}_{\omega_s}$ and covariance matrix $\mathbf{COV}_{\hat{\mathbf{f}}} = \mathbf{B}\mathbf{COV}_{\omega_s}\mathbf{B}^T = d_n/(c_n - 1)\mathbf{B}\mathbf{H}\mathbf{B}^T$. Therefore, PDF of $\hat{\mathbf{f}}$ is expressed as:

$$(8) \quad p(\hat{\mathbf{f}}) = \frac{\Gamma\left(\frac{2c_n + N}{2}\right)}{\Gamma(c_n)\pi^{N/2}\left(\det\left[\frac{d_n}{c_n}\mathbf{B}\mathbf{H}\mathbf{B}^T\right]\right)^{1/2}} (2c_n)^{N/2} \left(1 + \frac{1}{2c_n}(\hat{\mathbf{f}} - \boldsymbol{\mu}_{\hat{\mathbf{f}}})^T \left(\frac{d_n}{c_n}\mathbf{B}\mathbf{H}\mathbf{B}^T\right)^{-1} (\hat{\mathbf{f}} - \boldsymbol{\mu}_{\hat{\mathbf{f}}})\right)^{\frac{2c_n + N}{2}}$$

Using the PDF of $\hat{\mathbf{f}}$ (i.e., $p(\hat{\mathbf{f}})$) calculated from Eq. (8), information entropy of $\hat{\mathbf{f}}$ can be calculated using Eq. (1). Then Eq. (2) can be used directly to determine the efficient sampling locations for additional measurements, as discussed in the following two subsections for one additional measurement and multiple additional measurements, respectively.

Efficient sampling location for one additional measurement

In this subsection, formulation is developed to objectively determine the efficient sampling location for one additional measurement given pre-existing measurements $\mathbf{y}$ using information entropy. In accordance with the definition of information entropy in Eq. (1) and the PDF in Eq. (8), the information entropy of $\hat{\mathbf{f}}$ can be derived as (e.g., Kotz and Nadarajah 2004):

$$(9) \quad H(\hat{\mathbf{f}}) = -\int p(\hat{\mathbf{f}}) \ln[p(\hat{\mathbf{f}})] d\hat{\mathbf{f}} = \frac{1}{2} \ln(\det(\frac{d_n}{c_n} \mathbf{BHB}^T)) + \text{const}I$$

where “constI” is a function of $2c_n$ and N , and “constI” is independent of locations of additional measurements. Note that $H(\hat{\mathbf{f}})$ in Eq. (9) is expressed as two terms in the right hand side of the equation: $\frac{1}{2} \ln(\det(\frac{d_n}{c_n} \mathbf{BHB}^T))$ that is a function of measurement locations and “constI” that is a constant and independent of locations of additional measurements. It shall be pointed out that, during the planning of additional measurements (e.g., determination of efficient locations for additional measurements), the measured values at the locations of additional measurements are not available because the measurements (e.g., Atterberg’s limit test or triaxial test) have not been performed. Therefore, only the pre-existing measured values $\mathbf{y}$ are used in the calculation of $H(\hat{\mathbf{f}})$ (i.e., information entropy without additional measurements) and $H(\hat{\mathbf{f}}_{\text{new}})$ (i.e., information entropy with additional measurements). In other words, the term “constI” in Eq. (9) remains constant for both $H(\hat{\mathbf{f}})$ and $H(\hat{\mathbf{f}}_{\text{new}})$. In addition, note that the term “ d_n/c_n ” in Eq. (9) is also a constant when only the pre-existing measured values $\mathbf{y}$ are used in the calculation of $H(\hat{\mathbf{f}})$ and $H(\hat{\mathbf{f}}_{\text{new}})$. Therefore, Eq. (9) may be rewritten as Eq. (10) below when $\mathbf{H} = (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1}$ (see Eq. (7)) is substituted into Eq. (9):

$$(10) \quad H(\hat{\mathbf{f}}) = \frac{1}{2} \ln[\det(\mathbf{B}) \det(\mathbf{H}) \det(\mathbf{B}^T)] + \text{const2} = -\frac{1}{2} \ln[\det(\mathbf{A}^T \mathbf{A} + \mathbf{D})] + \text{const2}$$

where “const2” is a constant that is a function of “const1” and “ d_n/c_n ”. Note that $\mathbf{B}$ and $\mathbf{B}^T$ are orthonormal matrix. Therefore, $\det(\mathbf{B})=1$ and $\det(\mathbf{B}^T)=1$, and they disappear in the right-hand side of Eq. (10).

When the information entropy with additional measurements $H(\hat{\mathbf{f}}_{\text{new}})$ is calculated during the determination of efficient sampling locations, the term “const2” in Eq. (10) remains constant because only the pre-existing measured values $\mathbf{y}$ are used in the calculation. The difference between $H(\hat{\mathbf{f}})$ and $H(\hat{\mathbf{f}}_{\text{new}})$ is reflected by the term “ $-\frac{1}{2} \ln[\det(\mathbf{A}^T \mathbf{A} + \mathbf{D})]$ ” in Eq. (10). Suppose that one (i.e., $N_m = 1$) additional measurement on $\mathbf{f}$ is planned. Then, a new row of $\mathbf{B}$, denoted as $\mathbf{r}_{N_m}^T$ ($N_m = 1$) with a length of N , shall be augmented to the matrix $\mathbf{A}$, since each measurement corresponds to a row of matrix $\mathbf{B}$. The new matrix of $\mathbf{A}$ is denoted as $\mathbf{A}_{N_m}^{\text{new}}$, which has a dimension of $(M+1) \times N$. In linear algebra, $(\mathbf{A}_{N_m}^{\text{new}})^T (\mathbf{A}_{N_m}^{\text{new}})$ may be rewritten as $(\mathbf{A}_{N_m}^{\text{new}})^T (\mathbf{A}_{N_m}^{\text{new}}) = \mathbf{A}^T \mathbf{A} + \mathbf{r}_{N_m} \mathbf{r}_{N_m}^T$ (e.g., Strang 2006). Then, the information entropy $H(\hat{\mathbf{f}}_{\text{new}})$ with an additional measurement is expressed as:

$$(11) \quad H(\hat{\mathbf{f}}_{\text{new}}) = -\frac{1}{2} \ln[\det(\mathbf{A}^T \mathbf{A} + \mathbf{D} + \mathbf{r}_{N_m} \mathbf{r}_{N_m}^T)] + \text{const2}$$

In accordance with matrix determinant lemma (e.g., Brookes 2005), $\det(\mathbf{A}^T \mathbf{A} + \mathbf{D} + \mathbf{r}_{N_m} \mathbf{r}_{N_m}^T)$ in Eq. (11) may be rewritten as $[1 + \mathbf{r}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{r}_{N_m}] \det(\mathbf{A}^T \mathbf{A} + \mathbf{D})$, and Eq. (11) is rewritten as:

$$(12) \quad \begin{aligned} H(\hat{\mathbf{f}}_{\text{new}}) &= -\frac{1}{2} \ln[1 + \mathbf{r}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{r}_{N_m}] - \frac{1}{2} \ln[\det(\mathbf{A}^T \mathbf{A} + \mathbf{D})] + \text{const2} \\ &= -\frac{1}{2} \ln[1 + \mathbf{r}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{r}_{N_m}] + H(\hat{\mathbf{f}}) \end{aligned}$$

When Eq. (12) is rearranged and combined with Eq. (2), the difference Δ between $H(\hat{\mathbf{f}})$ and $H(\hat{\mathbf{f}}_{\text{new}})$ is expressed as:

$$(13) \quad \Delta = H(\hat{\mathbf{f}}) - H(\hat{\mathbf{f}}_{\text{new}}) = \frac{1}{2} \ln[1 + \mathbf{r}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{r}_{N_m}]$$

As discussed in the section entitled “Information entropy”, the efficient sampling location for the additional measurement is the location that maximizes the information gain calculated by Eq. (13). Eq. (13) suggests that the maximum Δ occurs when the term “ $\mathbf{r}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{r}_{N_m} = \mathbf{r}_{N_m}^T \mathbf{H} \mathbf{r}_{N_m}$ ” is maximized because a logarithm function is monotonically increasing function. Because $\mathbf{r}_{N_m}^T$ is a row of the matrix $\mathbf{B}$, maximization of Δ reduces to finding the row of the matrix $\mathbf{B}$ that maximizes $\mathbf{r}_{N_m}^T \mathbf{H} \mathbf{r}_{N_m}$, except the M rows of the matrix $\mathbf{B}$ that correspond to the M data points in pre-existing measurements $\mathbf{y}$.

Note that no additional calculation of $\mathbf{r}_{N_m}^T \mathbf{H} \mathbf{r}_{N_m}$ is needed because they have been calculated already in the BCS results. Recall that $\mathbf{COV}_{\hat{\mathbf{f}}} = d_n / (c_n - 1) \mathbf{B} \mathbf{H} \mathbf{B}^T$ and the variance of the i -th element of $\hat{\mathbf{f}}$ (i.e., the i -th diagonal term of $\mathbf{COV}_{\hat{\mathbf{f}}}$) is $d_n / (c_n - 1) \mathbf{r}_i^T \mathbf{H} \mathbf{r}_i$, where $\mathbf{r}_i^T$ is the i -th row of the matrix $\mathbf{B}$ (Wang et al. 2017). Similar to the term “ d_n / c_n ” discussed before, “ $d_n / (c_n - 1)$ ” is a constant for variance of all i -th element of $\hat{\mathbf{f}}$ (i.e., all diagonal terms of $\mathbf{COV}_{\hat{\mathbf{f}}}$). Therefore, when the location of the maximum variance of $\hat{\mathbf{f}}$ (i.e., the maximum of $d_n / (c_n - 1) \mathbf{r}_i^T \mathbf{H} \mathbf{r}_i$ or $\mathbf{r}_i^T \mathbf{H} \mathbf{r}_i$, $i = 1, 2, 3, \dots, N$) is taken as the location for additional measurement (i.e., the maximum of $\mathbf{r}_i^T \mathbf{H} \mathbf{r}_i$ is taken as $\mathbf{r}_{N_m}^T \mathbf{H} \mathbf{r}_{N_m}$), $\mathbf{r}_{N_m}^T \mathbf{H} \mathbf{r}_{N_m}$ and the entropy gain in Eq. (13) obtain their maxima. In other words, the efficient location for one additional measurement is the location of maximum variance of $\hat{\mathbf{f}}$ that has been obtained from the

BCS interpretation of the pre-existing measurements $\mathbf{y}$. The efficient sampling location for the additional measurement corresponds to the location where the interpreted soil property has the largest uncertainty. Once sample is taken at this location and measurement is obtained, the uncertainty in the interpreted soil property profile is efficiently reduced and the maximum information gain is achieved. This is consistent with those proposed for signal processing in electronic engineering literature (e.g., MacKay 1992; Ji et al. 2008). In addition, note that the efficient sampling locations for additional measurements should not include locations of the M pre-existing measurements $\mathbf{y}$ because measurements have already been taken at these locations.

Efficient sampling locations for multiple additional measurements

In this subsection, formulation is developed to simultaneously determine efficient sampling locations for multiple ($N_m > 1$) additional measurements using $\mathbf{y}$. Suppose that N_m additional measurements on $\mathbf{f}$ are planned. Then, with the N_m additional measurements, matrix $\mathbf{A}$ becomes $\mathbf{A}_{N_m}^{\text{new}}$, which is a matrix with a dimension of $(M+N_m) \times N$. $\mathbf{A}_{N_m}^{\text{new}}$ contains N_m ($N_m > 1$) new rows of the matrix $\mathbf{B}$, denoted as $\mathbf{R}_{N_m}^T$. $\mathbf{R}_{N_m}^T$ is a matrix with a dimension of $N_m \times N$. Subsequently, $(\mathbf{A}_{N_m}^{\text{new}})^T \mathbf{A}_{N_m}^{\text{new}}$ may be expressed as $(\mathbf{A}_{N_m}^{\text{new}})^T \mathbf{A}_{N_m}^{\text{new}} = \mathbf{A}^T \mathbf{A} + \mathbf{R}_{N_m} (\mathbf{R}_{N_m}^T)$. Then, the information entropy of the newly reconstructed signal $\hat{\mathbf{f}}_{\text{new}}$ with the N_m additional measurements is expressed as:

$$(14) \quad H(\hat{\mathbf{f}}_{\text{new}}) = -\frac{1}{2} \ln [\det(\mathbf{A}^T \mathbf{A} + \mathbf{D} + \mathbf{R}_{N_m} \mathbf{R}_{N_m}^T)] + \text{const}2$$

Similar to the formulation in the previous subsection, $\det(\mathbf{A}^T \mathbf{A} + \mathbf{D} + \mathbf{R}_{N_m} \mathbf{R}_{N_m}^T)$ in Eq. (14) can be expanded as $\det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}] \det(\mathbf{A}^T \mathbf{A} + \mathbf{D})$, in accordance with

the matrix determinant lemma (e.g., Brookes 2005). $\mathbf{I}_{N_m \times N_m}$ is an identity matrix with a dimension of $N_m \times N_m$. Then Eq. (14) is rewritten as:

$$(15) \quad \begin{aligned} H(\hat{\mathbf{f}}_{\text{new}}) &= -\frac{1}{2} \ln \left\{ \det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}] \right\} - \frac{1}{2} \ln [\det(\mathbf{A}^T \mathbf{A} + \mathbf{D})] + \text{const2} \\ &= -\frac{1}{2} \ln \left\{ \det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}] \right\} + H(\hat{\mathbf{f}}) \end{aligned}$$

and the difference between $H(\hat{\mathbf{f}})$ and $H(\hat{\mathbf{f}}_{\text{new}})$ is:

$$(16) \quad \Delta = H(\hat{\mathbf{f}}) - H(\hat{\mathbf{f}}_{\text{new}}) = \frac{1}{2} \ln \left\{ \det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}] \right\}$$

The efficient sampling locations for multiple additional measurements are determined when the Δ calculated by Eq. (16) is maximized. Based on Eq. (16), maximizing Δ is equivalent to finding the $\mathbf{R}_{N_m}^T$ (i.e., a combination of N_m rows of the matrix $\mathbf{B}$, except the M rows corresponding to the pre-existing measurements $\mathbf{y}$) that maximizes the term “ $\det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}]$ ”. The process of finding the $\mathbf{R}_{N_m}^T$ can be formulated as an optimization problem and solved by an optimization function in some commercial software, such as MATLAB (MathWorks 2018). The N_m rows of the matrix $\mathbf{B}$ that maximize $\det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}]$ correspond to the efficient sampling locations for the N_m additional measurements. It is worth noting that the $N_m = 1$ case in the previous subsection is a special case of the $N_m > 1$ case in this subsection. When $N_m = 1$, Eqs. (14) to (16) reduces to Eqs. (11) to (13), respectively.

Note that determination of the efficient sampling locations in this study is not related to engineering response of a geotechnical structure (e.g., stability of a slope). In geotechnical engineering practice, details of the intended geotechnical structures (e.g., exact geometry, locations, and loading, as well as hydraulic conditions and spatial variability of soil properties) might not be available during preliminary/planning stage of a geotechnical site

characterization. Therefore, it might be difficult to analyse engineering response of a geotechnical structure during preliminary/planning stage of a site characterization. This is consistent with engineering practice that geotechnical site characterization is usually performed before detailed geotechnical analysis and design.

Illustrative example I: measurements along vertical direction

In this section, the proposed method is illustrated using a set of real CPT tip resistance, q_c , data, as shown in Figure 3a by a solid line. The test was carried out vertically into a sand layer in Christchurch, New Zealand and the data were downloaded from the website of New Zealand Geotechnical Database (<https://www.nzgd.org.nz/>) with a CPT No. of 10126 in the database. The sand layer is around 13.60m thick, ranging from a depth of 6.00m to a depth of around 19.60m. For illustration, the q_c data within the depths of [6.02m, 16.24m] are used to demonstrate the method proposed in this paper. The q_c data were recorded with an interval of 0.02m and have 512 data points in total. Therefore, the original q_c data are almost continuous, and can be used to validate the results obtained from the BCS method using sparse measurements as input and to evaluate the performance of the sampling locations determined in characterizing soil property profile. Note that CPT data are adopted in this example only for validation because they are almost continuous. In geotechnical engineering practice, additional measurements cannot be taken in a vertical CPT sounding.

Efficient sampling locations for a given number M of measurements

Suppose that M (e.g., $M = 18$) measurements on the q_c profile are to be taken first to characterize the underlying q_c profile. In accordance with the results shown in the Section entitled “Efficient sampling locations for a given number of measurements”, the $M = 18$

measurements should be taken with an equal sampling interval of $h/(M-1) = (16.24-6.02)/17 \approx 0.6\text{m}$, as shown in Figure 3a by open circles. After the efficient sampling locations are determined, measurements are performed to obtain the numerical values of q_c at these locations. In this example, the q_c values from the original q_c profile are taken as the measured values to mimic the measurement procedure. Once these $M=18$ measurements are taken at different depths, the whole q_c profile can be interpreted from 18 measured data points using BCS (e.g., Wang and Zhao 2017). The BCS interpretation results are shown in Figure 3b, including the best estimate profile by a dashed line and the 95% confidence interval (CI) by two dotted lines. The 95% CI is approximately equal to $\mu_{\hat{f}} \pm 1.96 \text{diag} \sqrt{\text{COV}_{\hat{f}}}$, where $\text{diag} \sqrt{\text{COV}_{\hat{f}}}$ is a column vector, representing the square root of diagonal elements of $\text{COV}_{\hat{f}}$. Figure 3b also includes the original q_c profile by a solid line. It shows that the dashed line follows a trend similar to that of the solid line, and most variation of the original q_c profile falls within the 95% confidence interval. These agreements imply that the measurements with an equal sampling interval in engineering practice provide reasonable results.

To evaluate the performance of the $M = 18$ measurements taken with an equal sampling interval throughout the soil layer, the difference between $\mu_{\hat{f}}$ and the original q_c profile is evaluated using a commonly used estimator, such as mean squared error (MSE). The MSE is expressed as:

$$(17) \quad \text{MSE} = \frac{1}{N} \sum_{i=1}^N (\mu_{\hat{f}_i} - f_i)^2$$

where $\mu_{\hat{f}_i}$ and f_i represent the i -th element of $\mu_{\hat{f}}$ and $\mathbf{f}$, respectively; and $\mathbf{f}$ represents the original q_c profile. Note that the smaller the MSE between $\mu_{\hat{f}}$ and $\mathbf{f}$, the better estimate of $\mu_{\hat{f}}$.

In this example, the MSE is around 3.38 for the $M = 18$ measurements with an equal sampling interval.

For comparison, random samples of the $M = 18$ measurements are performed within the soil layer with depths ranging from 6.02m to 16.24m. BCS interpretation is performed using the random samples as input to estimate the μ_f , and the corresponding MSE between μ_f and f is calculated using Eq. (17). 1000 sets of the random samples of $M = 18$ measurements are taken and the procedure is repeated 1000 times, leading to 1000 MSEs. Subsequently, a box-plot of MSE is constructed using the 1000 MSEs. Figure 4 plots the 1000 MSEs by open circles together with a box-plot of the statistics, e.g., maxima, minima, 25th percentile, 75th percentile, median and mean. Figure 4 also includes the MSE corresponding to the $M = 18$ measurements with an equal sampling interval shown in Figure 3 by an open diamond. The MSE obtained from random samples exhibits significant variability, ranging from around $\text{MSE} = 3.04$ to $\text{MSE} = 9.54 \times 10^3$. In contrast, the open diamond is much smaller than the mean (i.e., the cross) and the 25th percentile of the 1000 MSEs obtained from random samples, and it plots very close to the minima of the 1000 MSEs. These observations demonstrate that the equal interval sampling strategy performs well.

After a given number of measurements is performed, their measured values are obtained, and they may be used as pre-existing measurements in the method proposed in the previous section for determining the efficient sampling locations of additional measurements, if necessary. Based on the number N_m of additional measurements planned, two cases are investigated respectively in the following two subsections, including (1) only one additional measurement (i.e., $N_m = 1$), and (2) multiple additional measurements (e.g., $N_m > 1$).

Efficient sampling location for one additional measurement

In this subsection, the proposed method is used to determine the efficient sampling location for one additional measurement (i.e., $N_m = 1$) using the pre-existing measurements $\mathbf{y}$, i.e., the open circles in Figure 3. The covariance matrix $\mathbf{COV}_{\hat{\mathbf{f}}} = \mathbf{B}\mathbf{COV}_w\mathbf{B}^T = d_n/(c_n-1)\mathbf{B}\mathbf{H}\mathbf{B}^T$ is calculated accordingly, and the efficient sampling location for one additional measurement is the location corresponding to the element of $\hat{\mathbf{f}}$ with the largest variance obtained. Figure 5b shows the variance associated with each element of $\hat{\mathbf{f}}$ by a dotted line. The element of $\hat{\mathbf{f}}$ at the location marked by an open diamond in Figure 5 has the largest variance, and it corresponds to the efficient sampling location for one additional measurement in this example.

To test the performance of the location determined, the q_c value at the location determined in Figure 5 is obtained from the original q_c profile, and it is used together with the $M = 18$ equal interval measurement points (i.e., the open circles in Figure 5a) as input to interpret a complete q_c profile by BCS (i.e., $M = 19$ in this case). The BCS results are summarized in Figure 6, including the best estimate profile and 95% CI represented by a dashed line and a pair of two dotted lines, respectively. For comparison, Figure 6 also includes the original q_c profile by a solid line. It shows that the dashed line follows a trend quite similar to that of the solid line, and it is more consistent than that shown in Figure 5a (i.e., the $M = 18$ case), especially at the depths of around 9-11m.

For a quantitative comparison, MSE between $\mu_{\hat{\mathbf{f}}}$ and the original q_c profile is evaluated for both the $M = 18$ and $M = 19$ cases. For the $M = 18$ case, the MSE is around 3.38, as reported previously. The MSE value reduces to around 1.94 when $M = 19$. A decrease of around 42.6% is obtained when one additional measurement is taken at the proposed sampling location. This substantial reduction in MSE demonstrates that the proposed method is effective in selecting efficient sampling locations. To further explore the

effectiveness of the proposed method, the location for the additional measurement is randomly selected from the remaining $N-M = 512-18 = 494$ locations. Then, the q_c value at each randomly selected location is obtained from the original q_c profile and used together with the 18 pre-existing measurements (i.e., the open circles in Figure 3a) to interpret a complete q_c profile by BCS. The MSE for each random sample is calculated. The procedure is repeated 1000 times, leading to 1000 MSEs. Figure 7 shows the 1000 MSEs, together with statistics of the 1000 MSEs using similar legends as those in Figure 5. Figure 7 also includes the MSE obtained from the proposed method by an open diamond.

Figure 7 shows that the open diamond from the proposed method is much smaller than the mean and 25th percentile of the 1000 MSEs obtained from random sampling. It is very close to the minima of the 1000 random sampling results. Because the total number of possible sampling location is small for one additional measurement in this example, the true minimum MSE may be obtained through exhaustive enumeration of $N-M = 512-18 = 494$ locations, followed by BCS interpretation. All 494 possible sampling locations are tested one by one, and the minimum MSE is shown by the open square in Figure 7. In this example, the true minimum MSE value is overlapped with the minima of the 1000 random samples, and it is around 1.80. The absolute error between the minimum MSE (i.e., 1.80) and the one obtained by the proposed method is only 0.14, with a relative difference of around 7.8%. This small difference suggests that the location determined by the proposed method is very effective in characterization of spatially varying soil properties.

Efficient sampling locations for multiple additional measurements

In this subsection, the proposed method is used to simultaneously determine efficient sampling locations for multiple additional measurements using pre-existing measurements y . Consider, for example, two additional measurements, $N_m = 2$. The efficient sampling

locations for two additional measurements correspond to the combination of two rows $\mathbf{R}_2^T$ of the matrix $\mathbf{B}$ that maximizes $\det[\mathbf{I}_{2 \times 2} + \mathbf{R}_2^T(\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_2]$. The combination of two rows $\mathbf{R}_2^T$ that maximizes $\det[\mathbf{I}_{2 \times 2} + \mathbf{R}_2^T(\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_2]$ is obtained using an optimization function in MATLAB (MathWorks 2018). The maximum value of $\det[\mathbf{I}_{2 \times 2} + \mathbf{R}_2^T(\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_2]$ obtained is shown by an open diamond in Figure 8b, and the corresponding efficient sampling locations are shown by the two open diamonds in Figure 8a. To validate the numerical optimization results from MATLAB, $\det[\mathbf{I}_{2 \times 2} + \mathbf{R}_2^T(\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_2]$ for all combinations of any two rows of the matrix $\mathbf{B}$ are calculated analytically and shown as a contour map Figure 8b. The maximum value of $\det[\mathbf{I}_{2 \times 2} + \mathbf{R}_2^T(\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_2]$ and their corresponding locations obtained from the analytical results are identical to those from MATLAB numerical optimization.

To evaluate the performance of the two locations (i.e., the open diamonds in Figure 8a) determined by the proposed method, the q_c values at the two selected location are obtained from the original q_c profile and used together with the 18 pre-existing measurements (i.e., the open circles in Figure 3a) to interpret a complete q_c profile by BCS. Then, the corresponding MSE between $\mu_{\hat{f}}$ and f is calculated and represented by an open diamond in Figure 7 (See the results corresponding to $N_m = 2$). Results similar to those for $N_m = 1$ are observed for $N_m = 2$ in Figure 7: the MSE corresponding to the efficient sampling locations (i.e., the open diamonds in Figure 8a) determined by the proposed method is much smaller than the mean and 25th percentile of 1000 MSEs obtained from random samples. The open diamond plots very close to the minima of the 1000 MSEs and the minimum MSE determined through exhaustive enumeration (see the open square in Figure 7). These agreements once again demonstrate that the proposed method is very effective in selecting efficient sampling locations. Similar analyses are also performed for $N_m = 3$ and 6, and their results of MSE and

its statistics are also shown in Figure 7. Observations similar to those for $N_m = 1$ and 2 are also obtained for $N_m = 3$ and 6. The proposed method is able to effectively select highly efficient sampling locations for multiple additional measurements.

It is clarified that determination of the minimum MSE in this example involves exhaustive enumeration of the remaining (N-M) locations of the original q_c profile. In other words, the minimum MSE can be obtained only if the original q_c profile is completely known (i.e., every q_c data point in the profile is measured). Such information is of course not available during the planning of site investigation or determination of sampling locations. Therefore, the minimum MSE is only used for comparison in this study. In a geotechnical site characterization, the objective of which is to characterize the whole soil property profile, the whole soil property profile is unknown, and hence, the minimum MSE and corresponding locations are unknown, too. Therefore, the sampling locations determined by the proposed method might not correspond to the minimum MSE. However, in accordance with the discussions and observations shown in this section, the locations determined by the proposed method with sparse measurements are very efficient because they provide results that are much smaller than the mean and median of MSE and approach to the minimum MSE. The method has been tested using a series of CPT data from different sites, and it is robust and performs reasonably well for all the test cases. Because of page limit, the results are not reported herein. In addition, note that the method proposed in this study is only illustrated using q_c data along the vertical direction (i.e., depth) in this section; whereas in real applications, determination of sampling locations for additional soil samples along horizontal directions may be encountered, as discussed in the next section.

Illustrative example II: measurements along a horizontal direction

In this section, the proposed method is applied to select efficient sampling locations for additional measurements along a horizontal direction using multiple sets of CPT data. Jaksa (1995) reported a large number of CPTs performed vertically into a Keswick clay layer at the South Parklands site in Adelaide. The CPTs were aligned along a horizontal line with a horizontal spacing of 0.5m. In this example, 32 sets of CPT tip resistance q_c data, with CPT ID ranging from CD1 to CD32, are adopted to illustrate the proposed method along a horizontal direction. The CPT data are downloaded from the webpage of TC304 database (http://140.112.12.21/issmge/Database_2011.htm). The CPT data generally vary from a depth of around 2m to a depth of around 5m. For a given depth (e.g., the depth of 2.62m), there are 32 q_c data points with a horizontal measurement interval of 0.5m, which are used to test the performance of the proposed method in determining sampling locations along a horizontal line.

Suppose that $M = 10$ out of 32 measurements of q_c are available during the preliminary stage of a geotechnical site characterization and they are shown in Figure 9a by open circles. Then, the 10 pre-existing measurements can be used to statistically reconstruct the variations of q_c along a horizontal direction (see Figure 9b) and determine the efficient sampling locations for additional measurements of q_c , when necessary. For example, the efficient sampling location for $N_m = 1$ additional measurement of q_c is selected and marked by an open triangle in Figure 10. It is determined by finding the element of $\hat{f}$ with the largest variance, following the procedure illustrated in Figure 5. If multiple, e.g., $N_m = 2, 3$ and 6, additional measurements of q_c are required, the sampling locations along the horizontal direction can also be determined by the proposed method, as shown in Figure 10 by open squares, crosses and open diamonds, respectively. These locations are determined by finding

the combination of rows $\mathbf{R}_{N_m}^T$ of $\mathbf{B}$ that maximizes $\det[\mathbf{I}_{N_m \times N_m} + \mathbf{R}_{N_m}^T (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1} \mathbf{R}_{N_m}]$, as illustrated in the previous section.

To evaluate the efficiency of the N_m sampling locations determined, the best estimate q_c profile is reconstructed using the BCS method with and without the N_m additional measurements at the proposed locations and compared with the original q_c profile in terms of MSE (i.e., Eq. (17)). This leads to five MSE values, including $\text{MSE} = 0.11$ corresponding to the $M = 10$ measurements scenario (i.e., before using the additional measurements at the proposed locations) and the four MSEs corresponding to $N_m = 1, 2, 3$ and 6 scenarios (i.e., after using the N_m additional measurements at the proposed locations), which are shown in Figure 11 by open diamonds. Figure 11 shows that the open diamond in each N_m scenario is much smaller than the $\text{MSE} = 0.11$ for $M = 10$. This observation indicates that a dramatic reduction in MSE or significant increase in accuracy in the interpreted q_c profile is obtained using the additional measurements at the locations determined by the proposed method. To validate the performance of the proposed locations, additional measurements at the size of N_m are enumerated exhaustively from the $(N-M)$ locations in the original q_c profile and used together with the $M = 10$ pre-existing measurements (i.e., the open circles in Figure 9) for reconstruction. For each case, a MSE is calculated between the best estimate and the original q_c profile. This leads to a large number of MSEs and the corresponding statistics for each N_m scenario, as shown in Figure 11. Figure 11 also plots the minimum MSE from exhaustive enumeration for each N_m scenario. Similar to the observation in Figure 7, the open diamond approaches the minima of the MSEs, and is much smaller than the mean and the 25th percentile of the MSEs determined through exhaustive enumeration in each N_m scenario. These observations suggest that the sampling locations for additional measurements along a horizontal direction determined by the proposed method are effective.

Note that Figures 9 - 11 are for a given depth (i.e., 2.62m). To systematically test the proposed method, the CPT data at 100 different depths ranging from 2.35m to 2.85m are used. For each given depth, the procedure used in the above example is repeated for $N_m = 2$ and 6 scenarios, and the MSE from the proposed method, the mean MSE and minimum MSE determined from exhaustive enumeration are obtained accordingly. This leads to 100 MSEs for the proposed method, 100 mean MSEs and minimum MSEs determined from exhaustive enumeration, respectively. The results are shown as histograms in Figure 12a&b for the $N_m = 2$ and 6 scenarios, respectively. In Figure 12, the histograms of MSEs from the proposed method, the histogram of the minimum and mean MSE determined from exhaustive enumeration are represented by solid columns, open columns with solid lines, and open columns with dashed lines, respectively. It is seen from Figure 12 that the histograms of the MSEs obtained from the proposed method approaches histograms of the minimum MSE determined from exhaustive enumeration and that they plot to the left-hand side of the histograms of the mean MSE determined from exhaustive enumeration. This observation shows that the MSE determined from the proposed method is statistically smaller than the mean of MSE and approaches the minimum MSE determined from exhaustive enumeration, which demonstrates that the proposed method is robust and efficient in determining sampling locations for additional measurements along a horizontal direction.

Summary and conclusions

This paper proposed an objective and quantitative method to determine efficient sampling locations in geotechnical site characterization under two scenarios: (1) a given number of measurements without pre-existing measurements, and (2) additional measurements with a limited number of pre-existing measurement data. For the first scenario, it was shown that the efficient sampling strategy is to take measurements with an equal sampling interval

throughout the whole soil layer of interest. This is quite consistent with the engineering experience and judgment in geotechnical engineering practice. The formulation and analysis in the manuscript provides scientific evidence to support this commonly used engineering experience and judgment.

For the second scenario, efficient sampling locations for additional measurements are objectively determined by integrating the concept of information entropy and Bayesian compressive sampling (BCS) method. The efficient sampling location for one additional measurement is the location with maximum variance in BCS interpretation. Equations were also derived for determining the efficient sampling locations of multiple additional measurements. The proposed methods effectively tackle the problem faced by engineers of where the additional measurements should be taken such that the pre-existing measurements can be best used and the most information on spatially varying soil property profiles can be obtained. The proposed methods were illustrated and validated using real CPT q_c data along both vertical and horizontal directions. The results showed that the proposed methods are effective and robust in determining sampling locations for measurements along the vertical or horizontal directions.

Acknowledgements

The work described in this paper was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 9042516 (CityU 11213117) and Project No. 8779012 (T22-603/15N)). The financial support is gratefully acknowledged. The authors would also like to thank the members of the TC304 Committee on Engineering Practice of Risk Assessment & Management of the International Society of Soil Mechanics and Geotechnical Engineering for developing the database 304dB used in this

study and making it available for scientific inquiry. We also wish to thank Jaksa, M. for contributing this database to the TC304 compendium of databases.

List of Symbols

$\mathbf{A}$	an $M \times N$ matrix and $\mathbf{A} = \Psi \mathbf{B}$
$\mathbf{A}_{N_m}^{\text{new}}$	an $(M+N_m) \times N$ matrix composed of matrix $\mathbf{A}$ and N_m rows of $\mathbf{B}$
$\mathbf{B}$	a prescribed orthonormal basis function matrix
COV_{ω_s}	derived covariance matrix of ω_s from sparse measurements $\mathbf{y}$
$\text{COV}_{\hat{f}}$	derived covariance matrix of $\hat{f}$ from sparse measurements $\mathbf{y}$
$\mathbf{C}_y$	an $M \times M$ correlation matrix
$\mathbf{C}_y^{-i,-j}$	a reduced matrix by eliminating the i -th row and the j -th column of $\mathbf{C}_y$
c, d	nonnegative parameters of a Gamma distribution
c_n, d_n	nonnegative parameters of a Gamma distribution
$\mathbf{D}$	an $N \times N$ diagonal matrix containing hyper-parameters α_i
$\det(\cdot)$	determinant of a matrix
$d_{i,j}$	distance between depths z_i and z_j
$\mathbf{f}$	a spatially varying signal (e.g., soil property profile)
$\hat{\mathbf{f}}$	approximation of the $\mathbf{f}$
$\hat{\mathbf{f}}_{\text{new}}$	reconstructed $\mathbf{f}$ using $\mathbf{y}$ together with new measurements
$\mathbf{H}$	an $N \times N$ matrix and $\mathbf{H} = (\mathbf{A}^T \mathbf{A} + \mathbf{D})^{-1}$
$H(\cdot)$	information entropy of a random process/vector
h	the thickness of a soil layer
$\mathbf{I}_{N_m \times N_m}$	an identity matrix with a dimension of $N_m \times N_m$

M	the number of pre-existing measurements $\mathbf{y}$, and $M \ll N$
N	the number of data points in $\hat{\mathbf{f}}$ and $\mathbf{f}$
N_m	the number of additional measurements
$p(\cdot)$	the probability density function (PDF) of a random variable(s)
q_c	CPT tip resistance data
$\mathbf{r}_{N_m}^T$	a row vector composed of $N_m = 1$ row of the matrix $\mathbf{B}$
$\mathbf{R}_{N_m}^T$	an $N_m \times N$ matrix composed of N_m rows of the matrix $\mathbf{B}$
r	the ratio between the numerically optimized determinant $\det(\mathbf{C}_y)_N$ and $\det(\mathbf{C}_y)_E$ obtained analytically for measurements with an equal sampling interval, i.e., $r = [\det(\mathbf{C}_y)_N]/[\det(\mathbf{C}_y)_E]$
$(\cdot)^T$	transpose operation in linear algebra
$\mathbf{X}$	a random vector with a length of N
$\mathbf{y}$	sparse measurements with a length of M
z_i	depth location
α_i	hyper-parameters when estimating ω_s in a Bayesian framework
μ_{ω_s}	estimated mean of vector ω_s
$\mu_{\hat{\mathbf{f}}}$	estimated mean of vector $\hat{\mathbf{f}}$
ϕ	effective friction angle of soil
Ψ	a problem-specific measurement matrix reflecting positions of $\mathbf{y}$ in $\mathbf{f}$
ω	weight vector, elements of which correspond to columns of matrix $\mathbf{B}$
$\rho_{i,j}$	correlation coefficient between soil properties at depth z_i and z_j and $0 \leq \rho_{i,j} \leq 1$
$\Gamma(\cdot)$	a Gamma function

References

- Back, P. 2000. Sampling Strategies and Data Worth Analysis for Contaminated Land. CLR Report No. 04, Swedish Geotechnical Institute, Sweden.
- Baecher, G.B., and Christian, J.T. 2003. Reliability and statistics in geotechnical engineering. John Wiley & Sons, New York.
- Brookes, M. 2005. The matrix reference manual. Available from <http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html> [accessed 24 Feb 2018].
- Cao, Z., and Wang, Y. 2013. Bayesian approach for probabilistic site characterization using cone penetration tests. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(2): 267-276. doi: doi:10.1061/(ASCE)GT.1943-5606.0000765.
- Cox, L.A. 1999. Adaptive spatial sampling of contaminated soil. *Risk Anal.* 19(6): 1059-1069.
- Cressie, N. 1991. Statistics for spatial data. John Wiley & Sons, New York, USA.
- DeGroot, D.J. and Baecher, G.B. 1993. Estimating autocovariance of in-situ soil properties. *Journal of Geotechnical Engineering*, 119 (1), 147-166. [https://doi.org/10.1061/\(ASCE\)0733-9410\(1993\)119:1\(147\)](https://doi.org/10.1061/(ASCE)0733-9410(1993)119:1(147))
- Donoho, D., Maleki, A., and Shahram, M. 2006. Wavelab 850. Software toolkit for time-frequency analysis. Available from http://statweb.stanford.edu/~wavelab/Wavelab_850/index_wavelab850.html.
- Fenton, G.A. 1999. Estimation for stochastic soil models. *Journal of Geotechnical and Geoenvironmental Engineering*, 125(6): 470-485. [https://doi.org/10.1061/\(ASCE\)1090-0241\(1999\)125:6\(470\)](https://doi.org/10.1061/(ASCE)1090-0241(1999)125:6(470))
- Fenton, G.A., Naghibi, F., and Hicks, M.A. 2018. Effect of sampling plan and trend removal on residual uncertainty. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, <https://doi.org/10.1080/17499518.2018.1455106>.

- Halim, I.S., and Tang, W.H. 1993. Site Exploration Strategy for Geologic Anomaly Characterization. *Journal of Geotechnical Engineering*, 119(2): 195-213. doi: doi:10.1061/(ASCE)0733-9410(1993)119:2(195).
- Hennig, P., and Schuler, C.J. 2012. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun): 1809-1837.
- Jaksa M.B. 1995. The influence of spatial variability on the geotechnical design properties of a stiff, overconsolidated clay: University of Adelaide.
- Ji, S., Dunson, D., and Carin, L. 2009. Multitask compressive sensing. *IEEE Transactions on Signal Processing*, 57(1): 92-106. doi: 10.1109/TSP.2008.2005866.
- Ji, S., Xue, Y., and Carin, L. 2008. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6): 2346-2356. doi: 10.1109/TSP.2007.914345.
- Jiang, S.-H., Papaioannou, I., and Straub, D. 2017. Optimizing Borehole Locations for Slope Reliability Assessment. In *Geo-Risk 2017: Geotechnical Risk from Theory to Practice*. Edited by J. Huang and G.A. Fenton and L. Zhang and D.V. Griffiths. ASCE, Denver. pp. 420-430.
- Jiang, S.H., Papaioannou, I., and Straub, D. 2018. Bayesian updating of slope reliability in spatially variable soils with in-situ measurements. *Engineering Geology*, 239: 310-320. <https://doi.org/10.1016/j.enggeo.2018.03.021>
- Johnson, R., LePoire, D., Huttenga, A., and Quinn, J. 2005. Bayesian approaches for adaptive spatial sampling: an example application. No. ANL/EAD/TM/05-1. Argonne National Lab.(ANL), Argonne, IL, USA.
- Ko, C.W., Lee, J., and Queyranne, M. 1995. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4), 684-691.
- Kotz, S., and Nadarajah, S. 2004. *Multivariate t distributions and their applications*. Cambridge University Press, New York, USA.

- Lacasse, S., and Nadim, F. 1996. Uncertainties in characterizing soil properties. In: *Uncertainties in the geologic environment: from theory to practice (GSP 58)*, New York.
- Lark, R. 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1-2): 49-80. [https://doi.org/10.1016/S0016-7061\(01\)00092-1](https://doi.org/10.1016/S0016-7061(01)00092-1).
- Li, J., Cassidy, M.J., Huang, J., Zhang, L., and Kelly, R. 2016. Probabilistic identification of soil stratification. *Géotechnique* 66(1): 16-26. doi: doi:10.1680/jgeot.14.P.242.
- Lo, M., and Leung, Y. 2018. Reliability Assessment of Slopes Considering Sampling Influence and Spatial Variability by Sobol' Sensitivity Index. *Journal of Geotechnical and Geoenvironmental Engineering*, 144(4): 04018010. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0001852](https://doi.org/10.1061/(ASCE)GT.1943-5606.0001852)
- MacKay, D.J. 1992. Information-based objective functions for active data selection. *Neural Computation*, 4(4): 590-604. Doi:10.1162/neco.1992.4.4.590
- MacKay, D.J. 2003. *Information theory, inference and learning algorithms*. Cambridge university press, USA.
- MathWorks, I. 2018. MATLAB: the language of technical computing. (<http://www.mathworks.com/products/matlab/>). [accessed 24 Feb 2018].
- Mayne, P.W., Christopher, B.R., and DeJong, J. 2002. *Subsurface investigations—geotechnical site characterization*. FHWA NHI-01-031. Federal Highway Administration, DOT, Washington, DC.
- McBratney, A., and Webster, R. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables—II: program and examples. *Computers & Geosciences*, 7(4): 335-365. [https://doi.org/10.1016/0098-3004\(81\)90077-7](https://doi.org/10.1016/0098-3004(81)90077-7).

- McBratney, A., Webster, R., and Burgess, T. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables—I: Theory and method. *Computers & Geosciences*, 7(4): 331-334. [https://doi.org/10.1016/0098-3004\(81\)90077-7](https://doi.org/10.1016/0098-3004(81)90077-7).
- Phoon, K.-K. 2017. Role of reliability calculations in geotechnical design. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 11(1), 4-21. <https://doi.org/10.1080/17499518.2016.1265653>.
- Phoon, K.-K., and Kulhawy, F.H. 1999. Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36(4): 612-624. doi: 10.1139/t99-038.
- Stein, A., and Ettema, C. 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agriculture, Ecosystems & Environment*, 94(1): 31-47. [https://doi.org/10.1016/S0167-8809\(02\)00013-0](https://doi.org/10.1016/S0167-8809(02)00013-0).
- Strang, G. 2006. *Linear algebra and its applications*. Academic, New York, USA.
- Stuedlein, A.W., Kramer, S.L., Arduino, P., and Holtz, R.D. 2012. Geotechnical characterization and random field modeling of desiccated clay. *Journal of Geotechnical and Geoenvironmental Engineering*, 138(11): 1301-1313. [https://doi.org/10.1061/\(ASCE\)GT.1943-5606.0000723](https://doi.org/10.1061/(ASCE)GT.1943-5606.0000723)
- Taboga, M. 2012. *Lectures on probability theory and mathematical statistics*. CreateSpace Independent Pub., San Bernardino, CA, USA.
- Tipping, M.E. 2001. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1: 211-244. doi: 10.1162/15324430152748236.
- van Groenigen, J.W., Siderius, W., and Stein, A. 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* 87(3): 239-259. doi: [https://doi.org/10.1016/S0016-7061\(98\)00056-1](https://doi.org/10.1016/S0016-7061(98)00056-1).

- Vanmarcke, E.H. 1977. Probabilistic modeling of soil profiles. *Journal of the geotechnical engineering division*, 103(11): 1227-1246.
- Vanmarcke, E. 2010. *Random fields: analysis and synthesis*. World Scientific, Singapore.
- Wang, Y. and Akeju, O.V. 2016. Quantifying the cross-correlation between effective cohesion and friction angle of soil from limited site-specific data. *Soils and Foundations*, 56(6):1057–1072. <https://doi.org/10.1016/j.sandf.2016.11.009>
- Wang, Y., Akeju, O.V., and Zhao, T. 2017. Interpolation of spatially varying but sparsely measured geo-data: A comparative study. *Engineering Geology*, 231: 200-217. doi: <https://doi.org/10.1016/j.enggeo.2017.10.019>.
- Wang, Y., Au, S.-K., and Cao, Z. 2010. Bayesian approach for probabilistic characterization of sand friction angles. *Engineering Geology*, 114(3–4): 354-363. doi: <http://dx.doi.org/10.1016/j.enggeo.2010.05.013>.
- Wang, Y., and Cao, Z. 2013a. Probabilistic characterization of Young's modulus of soil using equivalent samples. *Engineering Geology*, 159:106-118. <https://doi.org/10.1016/j.enggeo.2013.03.017>
- Wang, Y., and Cao, Z. 2013b. Expanded reliability-based design of piles in spatially variable soil using efficient Monte Carlo simulations. *Soils and Foundations*, 53(6):820–834. <https://doi.org/10.1016/j.sandf.2013.10.002>
- Wang, Y., Cao, Z., and Li, D. 2016. Bayesian perspective on geotechnical variability and site characterization. *Engineering Geology*, 203: 117-125. <https://doi.org/10.1016/j.enggeo.2015.08.017>
- Wang, Y., Huang, K., and Cao, Z. 2013. Probabilistic identification of underground soil stratification using cone penetration tests. *Canadian Geotechnical Journal*, 50(7): 766-776. <https://doi.org/10.1139/cgj-2013-0004>

- Wang, Y., and Zhao, T. 2016. Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. *Canadian Geotechnical Journal*, 53(9): 1547-1559. doi: 10.1139/cgj-2015-0545.
- Wang, Y., and Zhao, T. 2017. Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique* 67(6): 523-536. doi: 10.1680/jgeot.16.P.143.
- Wang, Y., Zhao, T., and Phoon, K.K. 2018. Direct Simulation of Random Field Samples from Sparsely Measured Geotechnical Data with Consideration of Uncertainty in Interpretation. *Canadian Geotechnical Journal*, 55(6): 862-880, <https://doi.org/10.1139/cgj-2017-0254>.
- Webster, R. 2000. Is soil variation random? *Geoderma* 97(3-4): 149-163. [https://doi.org/10.1016/S0016-7061\(00\)00036-7](https://doi.org/10.1016/S0016-7061(00)00036-7).
- Zhao, T., Montoya-Noguera, S., Phoon, K.K., and Wang, Y. 2018. Interpolating spatially varying soil property values from sparse data for facilitating characteristic value selection. *Canadian Geotechnical Journal*, 55(2): 171-181. doi: 10.1139/cgj-2017-0219.
- Zhao, T., and Wang, Y. 2018a. Simulation of cross-correlated random field samples from sparse measurements using Bayesian compressive sensing, *Mechanical Systems and Signal Processing*, 112: 384–400, <https://doi.org/10.1016/j.ymssp.2018.04.042>
- Zhao, T., and Wang, Y. 2018b. Interpretation of pile lateral response from deflection measurement data: A compressive sampling-based method. *Soils and Foundations*, 58(4): 957-971, <https://doi.org/10.1016/j.sandf.2018.05.002>.
- Zhao, T., Hu, Y., and Wang, Y. 2018. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. *Engineering Geology*, 246, 162-175. <https://doi.org/10.1016/j.enggeo.2018.09.022>

Zhu, H., and Zhang, L.M. 2013. Characterizing geotechnical anisotropic spatial variations using random field theory. *Canadian Geotechnical Journal*, 50(7): 723-734. doi: 10.1139/cgj-2012-0345.

Draft

Figure Captions

- Figure 1. Comparison between numerically optimized $\det(\mathbf{C}_y)_N$ and $\det(\mathbf{C}_y)_E$ for measurements with an equal sampling interval for the $M = 4$ case
- Figure 2. Comparison between numerically optimized $\det(\mathbf{C}_y)_N$ and $\det(\mathbf{C}_y)_E$ for measurements with an equal sampling interval for different M , h and λ_c
- Figure 3. Original profile, measurement data and Bayesian compressive sampling results
- Figure 4. Performance of $M = 18$ measurements with an equal sampling interval
- Figure 5. Efficient sampling location for one additional measurement ($N_m = 1$)
- Figure 6. Bayesian compressive sampling results for $M = 18$ equal interval measurement points plus one additional measurement point determined by the proposed method
- Figure 7. Performance of the efficient sampling locations for N_m additional measurements determined by the proposed method
- Figure 8. Efficient sampling locations for two additional measurements
- Figure 9. Original q_c profile (along horizontal direction), sparse measurements and Bayesian compressive sampling results
- Figure 10. Efficient sampling locations along a horizontal direction for N_m additional measurements determined by the proposed method
- Figure 11. Performance of the efficient sampling locations along a horizontal direction for N_m additional measurements determined by the proposed method
- Figure 12. Performance of the efficient sampling locations along a horizontal direction for $N_m = 2$ and 6 additional measurements determined by the proposed method under 100 different depths

Figure 1 Comparison between numerically optimized $\det(\mathbf{C}_y)_N$ and $\det(\mathbf{C}_y)_E$ for measurements with an equal sampling interval for the $M = 4$ case

Figure 2 Comparison between numerically optimized $\det(\mathbf{C}_y)_N$ and $\det(\mathbf{C}_y)_E$ for measurements with an equal sampling interval for different M , h and λ_c

Figure 3 Original profile, measurement data and Bayesian compressive sampling results

Figure 4 Performance of M = 18 measurements with an equal sampling interval

Figure 5 Efficient sampling location for one additional measurement ($N_m = 1$)

Figure 6 Bayesian compressive sampling results for $M = 18$ equal interval measurement points plus one additional measurement point determined by the proposed method

Figure 7 Performance of the efficient sampling locations for N_m additional measurements determined by the proposed method

Figure 8 Efficient sampling locations for two additional measurements

Figure 9 Original q_c profile (along horizontal direction), sparse measurements and Bayesian compressive sampling results

Figure 10 Efficient sampling locations along a horizontal direction for N_m additional measurements determined by the proposed method

Figure 11 Performance of the efficient sampling locations along a horizontal direction for N_m additional measurements determined by the proposed method

Figure 12 Performance of the efficient sampling locations along a horizontal direction for $N_m = 2$ and $N_m = 6$ additional measurements determined by the proposed method under 100 different depths