Journal of GeoEngineering, Vol. 16, No. 2, pp. 061-073, June 2021
http://dx.doi.org/10.6310/jog.202106_16(2).2

PROJECT DEEPGEO — DATA-DRIVEN 3D SUBSURFACE MAPPING

Kok-Kwang Phoon [1*] and Jianye Ching [2]

ABSTRACT

Data-driven site characterization (DDSC) is defined as any site characterization methodology that relies solely on measured data, both site-specific data collected for the current project and existing data of any type collected from past stages of the same project or past projects at the same site, neighboring sites, or beyond. One key complication is that real data is "ugly". A useful mnemonic is MUSIC-3X (Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted with "3X" denoting three dimensional spatial variations). It is an open question whether DDSC can solve real world subsurface mapping problems based on real world MUSIC-3X data from routine projects with minimum ad-hoc assumptions. The computational challenges are very significant, but some reasonable partial solutions have been obtained recently. One promising solution is Sparse Bayesian Learning (SBL). It is nearly data-driven and it can handle a large scale 3D problem without incurring excessive cost. However, it can only handle one type of field test data. Nonetheless, it is already useful for practice. A 3D SBL version would be made available in Rocscience's Settle3 (three-dimensional soil settlement analysis) in the near future to generate subsurface maps based on cone penetration test data. The second solution is based on a variant of the Gaussian Process Regression (GPR-MUSIC-3X). It can handle multiple field test data by learning the cross-correlation behavior among different soil parameters at a single site of interest. GPR-MUSIC-3X can be enhanced to learn cross-correlation behaviors at multiple sites and thus bring information from "similar" sites in a larger generic database to bear on improving predictions at a single site. Both 3D SBL and GPR-MUSIC-3X are cross validated using a 2D virtual ground and an actual 3D site in Texas. The hunt is on for a "holy grail" mapping approach that is fully data-driven, MUSIC-3X compliant, and is able to exploit all available data including data from similar sites. This is Project DeepGeo (inspired by DeepMind that produces AlphaGo), which constitutes one major research effort in the emerging field of data-centric geotechnics.

Key words: Data-driven site characterization (DDSC), MUSIC-3X, Sparse Bayesian Learning (SBL), Gaussian process regression, data-centric geotechnics.

1. INTRODUCTION

Site characterization is a cornerstone of geotechnical and rock engineering. It is not possible to derive the characteristics of a *specific* site (stratification, discontinuities, anomalies, spatial variation of physical/mechanical properties, ground water flow, *etc.*) from first principles. A broad appreciation of local geology and experience from similar sites can inform site characterization, but only an interpretation of data collected from a site investigation programme can provide detailed *quantitative* information of the ground conditions at a specific location. It is not surprising that a minimum site investigation programme is mandated in building regulations. This is a tacit acknowledgement that each site is unique to some degree.

It is well known that site-specific data alone are not sufficient for making design decisions. One example is the estimation of a design property from one or more field test parameters. A local correlation between the undrained shear strength and the cone tip resistance is preferred, but there are frequently insufficient undisturbed samples or field vane shear tests to support a reasonably accurate and precise correlation. Generic correlations supported by data from multiple sites are much more commonly used in practice (Kulhawy and Mayne 1990; Ching and Phoon 2012; Ching *et al.* 2014). However, a generic correlation is not directly applicable to a given site. For example, the empirical cone factor relating the undrained shear strength to the cone tip resistance is known to be site-specific. The generic average cone factor can under- or over-estimate the actual value relevant to a given site (Phoon *et al.* 2003). Many generic correlations are global, rather than regional or municipal, in data coverage. Therefore, the estimation of a design property is frequently guided by both classical statistics (generic correlation) and human judgment founded on local experience. An experienced engineer is cognizant of the site effect and will adjust for the under- or over-estimation issue in the design. The practice of geotechnical and rock engineering is perceived to be an art as much as a science for this reason.

Although human judgment remains indispensable to decision making in the foreseeable future, there is significant room to improve the fairly simple statistical models that are widely used in geotechnical and rock engineering practice. The National Research Council (1995) astutely observed that "in the process of applying probabilistic methods to geotechnical engineering, the problems tend to be oversimplified, thus the results achieved do not reflect the real issues at the specific site". Phoon (2017) also exhorted "the geotechnical reliability community to evolve beyond overly simplistic assumptions and methods that are incongruent with the established body of geotechnical knowledge,

Manuscript received April 21, 2021; revised May 10, 2021; accepted May 11, 2021.

[1*] Professor (corresponding author), Singapore University of Technology and Design, No. 8, Somapah Road, Singapore 487372 (e-mail: kkphoon@sutd.edu.sg).

[2] Professor, Dept of Civil Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.).

principles, and experience". One major limitation of prevailing statistical models is the gap between the assumed and the actual attributes of geotechnical data. Many classical statistical models assume that data is homogeneous, abundant, independent, and normally distributed. These models have been criticized as unrealistic by practitioners for many years. Over-simplification arguably constitutes one major roadblock in the adoption of data-driven methods in geotechnical engineering (Phoon 2017). Phoon *et al.* (2021) opined that the "ugly data" challenge lies at the heart of any data-driven site characterization (DDSC). The authors defined DDSC as any site characterization methodology that relies solely on measured data, both site-specific data collected for the current project and existing data of any type collected from past stages of the same project or past projects at the same site, neighboring sites, or beyond. One example of ugly data is MUSIC-3X (Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted with "3X" denoting three dimensional spatial variations). It is a useful mnemonic to highlight the attributes of real site data and to contrast with the highly idealized assumptions underlying classical statistics. Ideal data is "beautiful". Real world data is "ugly". It is premature to say that ugly data does not contribute to decision making. The central tenet in data-centric geotechnics is that data has value as long as it is not fake. The challenge is to draw useful inferences from ugly data (Phoon *et al.* 2021). It is accurate to say that data-centric geotechnics is ugly-data-centric geotechnics. Ideal data and methods drawing insights from ideal data are not part of the agenda of this emerging field.

It is an open question what data-driven site characterization (DDSC) can achieve and how useful are the outcomes for practice, but this "value of data" question is of major interest given the rapid pace of digital transformation in many industries. The scientific aspects of this question are presented as three challenges by Phoon *et al.* (2021): (1) ugly data, (2) site recognition, and (3) stratification. These challenges are inter-related. The role of human judgment is expected to be sharpened with the advent of DDSC. This evolution is similar to the advent of powerful 3D finite element software that has largely relieved human judgment from bridging the gap between highly simplified back of the envelope calculations and complex real-world soil-structure interactions. The potential difference worth highlighting here is that DDSC may one day evolve into an artificial intelligence (AI) that can emulate human learning and experience building. Phoon (2020) called this longer term project AlphaGeo. Phoon *et al.* (2021) imagined an artificial intelligence that can mimic human learning to be equivalent to a "super engineer" that has gained and continues to gain from the pooled experiences of all human engineers around the world. Since it is already known that an experienced engineer makes better judgment than a novice with no experience, one can speculate that such a "super engineer" with access to extensive databases, capable of detecting site differences through data-driven methods (as demonstrated in this paper), and capable of moderating its predictions by drawing upon relevant past human experiences will be making better site-specific predictions than those from classical statistical methods moderated by "reality checks" from a single engineer.

It is important to show that the solution can be extended in reality (not in principle) to 3D at reasonable computational cost on a desktop computer. Engineers are looking for effective solutions that they can apply in practice based on available data in *routine* projects. Without paying attention to the value of data in solving real world problems, it is difficult to see how any data-driven research agenda can progress because engineers own almost all data and they have yet to be convinced what data, big or small, could do to transform current practice. Although the computational challenges are very significant for 3D subsurface mapping, some reasonable partial solutions have been obtained recently. The purpose of this paper is to present two promising advances in Project DeepGeo, a research effort to bring data-driven 3D subsurface mapping to practice, specifically to routine projects in practice. A subsurface mapping problem can include identifying geometrical features (stratification, discontinuities, anomalies, *etc.*), evaluating material behaviors (*e.g.,* physical and mechanical properties) and their spatial distributions, and characterizing geoenvironmenal processes (*e.g.,* ground water flow). The first solution is Sparse Bayesian Learning (SBL). This approach focuses on simulating a site-specific subsurface map containing stratification and mechanical properties conditioned on one type of field test data, for example cone tip resistance soundings. Notwithstanding that it is a partial DDSC solution, it is already useful for practice. A 3D SBL version would be made available in Rocscience's Settle3 (three-dimensional soil settlement analysis) in the near future to generate subsurface maps based on cone penetration test data. The second solution is based on a variant of the Gaussian process regression (GPR-MUSIC-3X) (*e.g.,* Neal 1998; Rasmussen and Williams 2016). It can handle multiple field test data by learning the cross-correlation behavior among different soil parameters at a single site of interest. GPR-MUSIC-3X can be enhanced to learn cross-correlation behaviors at multiple sites and thus bring information from "similar" sites in a larger generic database to bear on improving predictions at a single site. The performance of 3D SBL and GPR-MUSIC-3X will be illustrated and cross-validated using a 2D virtual ground and an actual 3D site at Baytown, Texas, USA. Project DeepGeo focuses more narrowly on data-driven 3D subsurface mapping. It can be regarded as one key building block that may one day bring a geotechnical AI (AlphaGeo) to fruition.

2. 1D SPARSE BAYESIAN LEARNING

Let us denote $Y = (y_1, y_2, \ldots, y_n)$ as the site investigation data observed at depths $(z_1, z_2, \ldots, z_n)$. For instance, y_i can be the corrected cone tip resistance (q_t) in the cone penetration test (CPT) at depth z_i. It can also be a transformed observation, *e.g.*, $y_i = \ln[q_t(z_i)]$. It is common to *model* the observation y_i as a summation of a "trend" $t(z_i)$ and a "spatial variation" $\varepsilon(z_i)$:

$$y_i = t(z_i) + \varepsilon(z_i) \tag{1}$$

The trend function is parameterized by expressing it as a linear combination of basis functions (BFs):

$$t(z) = \sum_{k=0}^{m} w_k \, \phi_k(z) \tag{2}$$

where $\phi_k(z)$ is the k-th BF, and w_k is the unknown weight. The basis functions are generally non-linear. Hence, Eq. (2) is non-linear. The simplest trend function is a constant = w_0. The most widely adopted is arguably a linear trend = $w_0 + w_1 \times z$. The spatial variation $\varepsilon(z)$ is *assumed* to follow a zero-mean stationary Gaussian random field with standard deviation = σ and a Whittle-Matérn (WM) auto-correlation model (Ching and Phoon 2019; Ching *et al.* 2019):

$$\rho(\Delta)=\frac{2}{\Gamma(\nu)}\cdot\left(\frac{\sqrt{\pi}\cdot\Gamma(\nu+0.5)\cdot|\Delta|}{\Gamma(\nu)\cdot\delta}\right)^{\nu}K_{\nu}\left(\frac{2\sqrt{\pi}\cdot\Gamma(\nu+0.5)\cdot|\Delta|}{\Gamma(\nu)\cdot\delta}\right) \quad (3)$$

where $\rho(\Delta)$ is the auto-correlation between two points with separation distance = $\Delta = |z_i - z_j|$; ν is the smoothness parameter; δ is the scale of fluctuation (SOF); Γ is the Gamma function; and K_ν is the modified Bessel function of the second kind with order ν. Cami *et al.* (2020) noted that two commonly adopted autocorrelation models are special cases of Eq. (3): (1) $\nu = 0.5$ produces the Markovian or single exponential model and (2) $\nu = \infty$ produces the Gaussian or squared exponential model. In practice, the case of $\nu > 3.5$ is practically indistinguishable from the Gaussian model ($\nu = \infty$) (Cami *et al.* 2020). Chang *et al.* (2021) proposed a cosine WM model to handle auto-correlations that fluctuate rather than decrease monotonically with Δ. The cosine WM model is parameterized by δ, ν, and a third parameter that controls the wavelength of the cosine function. Cami *et al.* (2020) noted that the cosine exponential (special case of cosine WM) model is adopted 10% of the time among the papers reviewed. In contrast, the single and squared exponential models are adopted more than 60% of the time.

It is important to emphasize here that only y_i is real. The "trend" and "spatial variation" are mathematical constructs or models. The weights of the basis functions w_k and (σ, δ, ν) are parameters of the trend and random field models, respectively. There is actually no satisfactory solution for estimating these parameters from a purely data-driven perspective. The current state-of-the-practice is estimate these parameters using ad-hoc approaches that may not be sufficiently robust (Jaksa *et al.* 1999; Phoon *et al.* 2003). For example, the trend is typically estimated using linear regression by assuming the residuals ($y_i - t_i$) are uncorrelated (Lumb 1966). However, this violates the reality that these residuals [termed as spatial variation in Eq. (1)] are correlated. Spatial interpolation methods such as kriging depend on these autocorrelations to name one important application in practice. A linear trend is prescribed based in part on judgment (soil properties tend to increase with depth or overburden pressure) and possibly popularity of linear regression. The random field parameters are commonly estimated using the method of moments (Cami *et al.* 2020). The statistical uncertainties associated with w_k and (σ, δ, ν) are ignored, although they are important in several fundamental aspects. For example, different combinations of w_k and (σ, δ, ν) are possible for the same set of data and these combinations exhibit correlations. At present, it is widely known that the simplicity of Eq. (1) is misleading. Equation (1) requires the "trend" to be separated from the "spatial variation". This "detrending" problem is known to be exceedingly challenging under some conditions (Ching *et al.* 2016, 2017).

The data-driven approach is to start with $Y = (y_1, y_2, \ldots, y_n)$ and estimate w_k and (σ, δ, ν) (including the functional form of the trend) with as few ad-hoc assumptions as possible. In this paper, we say an assumption is ad-hoc when it is not founded on physics and/or informed by data. Given that subsurface mapping is primarily based on site investigation data, an ad-hoc assumption is most likely one that is not related to real world data. Needless to say, the approach should ideally apply under MUSIC-3X or less general but nonetheless realistic conditions. The performance of any proposed approach must be examined using leave-one-out (LOO) or k-fold cross validation. This is demonstrated in the examples presented below. This kind of validation exercise is crucial particularly in the presence of ad-hoc assumptions, but seldom emphasized in the current state-of-the-practice, although it is common in the machine learning research community. Finally, the approach should be reasonably computable in 3D.

As noted above, no satisfactory approach exists currently. One promising candidate approach is a two-step Bayesian framework proposed by Ching and Phoon (2017). In Step 1, a set of suitable basis functions that parameterizes the trend function (Eq. (2)) is selected using Sparse Bayesian Learning (SBL) (Tipping 2001). In this way, the functional form of the trend is "learned" from data rather than prescribed with no relation to data. In Step 2, an advanced Markov chain Monte Carlo method (Ching and Chen 2007) is adopted to draw posterior samples of (W, $\ln\sigma$, $\ln\delta$) conditioning on the Y data. Note that $W = (w_0, w_1, \ldots, w_m)^T$. In this way, the trend and random field parameters, or more precisely their posterior distributions, are "learned" from the data as well. The key ad-hoc assumption in this approach is that (W, $\ln\sigma$, $\ln\delta$) follows non-informative flat prior distributions. Nonetheless, one can argue that this SBL approach is largely or nearly data-driven.

3. GAUSSIAN PROCESS REGRESSION (GPR)

SBL only involves a single soil parameter such as the cone tip resistance. Ching *et al.* (2021a) proposed a GPR-MUSIC-3X approach that can handle multiple parameters that vary in 3D. This approach is based on a variant of the Gaussian process regression (GPR) (*e.g.,* Neal 1998; Rasmussen and Williams 2016). It builds on past research on GPR-MUSIC (no vertical spatial correlation and perfect horizontal spatial correlation) by Ching and Phoon (2019) and GPR-MUSIC-X (vertical spatial correlation and perfect horizontal spatial correlation) by Ching and Phoon (2020a).

Similar to SBL, physical soil parameters Y_i are assumed to vary spatially, say with depth (z) and distance (h) for a 2D profile as shown in Fig. 1(a). $Y_i(z, h)$ can be transformed to a vector Gaussian process $X_i(z, h)$ (Ching and Phoon 2015). The vector Gaussian process $X_i(z, h)$ can be further regarded as a scalar Gaussian process $X(z, h, p)$ in which $X(z, h, 1) = X_1(z, h)$, $X(z, h, 2) = X_2(z, h)$, and $X(z, h, 3) = X_3(z, h)$. The parameter p can be regarded as continuous although we are only interested in $p = 1, 2, 3$. Let $v = (z, h, p)$. A Gaussian process regression requires a Gaussian prior with mean function $m(v)$ and a covariance function $K(v, v')$.

Ching and Phoon (2019) were the first to propose a feasible Gibbs sampler (GS) that can construct the PDF of MUSIC data. Their approach can be regarded as GPR with the following priors: (1) mean function $m(v) = m(p)$ and $m(1) = c_1$, $m(2) = c_2$ and $m(3) = c_3$, in which c_1, c_2, and c_3 are constant means of property 1, 2, and 3, respectively and (2) covariance function $K(v, v') = 0$ if $z \neq z'$ and $K(v, v') = C(p, p')$ if $z = z'$ in which $C(p, p')$ is cross-covariance between property p and property p'. This GPR-MUSIC is extended to GPR-MUSIC-X by Ching and Phoon (2020a) using the following priors: (1) $m(v) = m(p)$ and (2) $K(v, v') = C(p, p') \times R_z(|z - z'|)$, in which R_z is the vertical autocorrelation function associated with the $m(v) = m(p)$ prior (this is distinct from the vertical autocorrelation function for the detrended residuals adopted in SBL). In GPR-MUSIC-X, C is updated by data (*i.e.,* it has a prior distribution that is updated to a posterior distribution by data), but R_z is prescribed and not updated by data. As noted above, GPR-

MUSIC-X has since been further extended to GPR-MUSIC-3X using the following priors: (1) $m(v) = m(p)$ and (2) $K(v, v') = C(p, p') \times R_z(|z - z'|)) \times R_h(|h - h'|)$, in which R_z and R_h are respectively the vertical and horizontal autocorrelation function associated with the $m(v) = m(p)$ prior (this is distinct from the vertical and horizontal autocorrelation function for the detrended residuals). In GPR-MUSIC-3X, C is updated by data (*i.e.,* it has a prior distribution that is updated to a posterior distribution by data), but R_z and R_h are prescribed and not updated by data. For the incomplete data shown in Fig. 1(b), a variant of the Gaussian process defined by (m, K, X^u) is adopted (Ching and Phoon 2019), in which X^u is a matrix of unobserved measurements (open markers in Fig. 1(b)). Both m and X^u are updated. The covariance function K is only partially updated (only the cross-covariance C is updated). The prior mean function $m(v) = m(p)$ is a typical assumption for GPR, although more complicated priors such as a linear combination of weighted basis functions with weights treated as additional hyperparameters that can be updated are possible. In the same vein, the vertical scale of fluctuation in R_z and the horizontal scale of fluctuation in R_h can be regarded as hyperparameters. GPR-MUSIC-3X did not consider more complicated GPR, because conjugate priors are not known to exist for these additional hyperparameters. The application of Gaussian process regression as a learning framework is powerful, because it could learn from similar sites through a hierarchical Bayesian model (HBM) (Ching *et al.* 2021b, 2021c). This enhanced method is called HBM-MUSIC-3X.

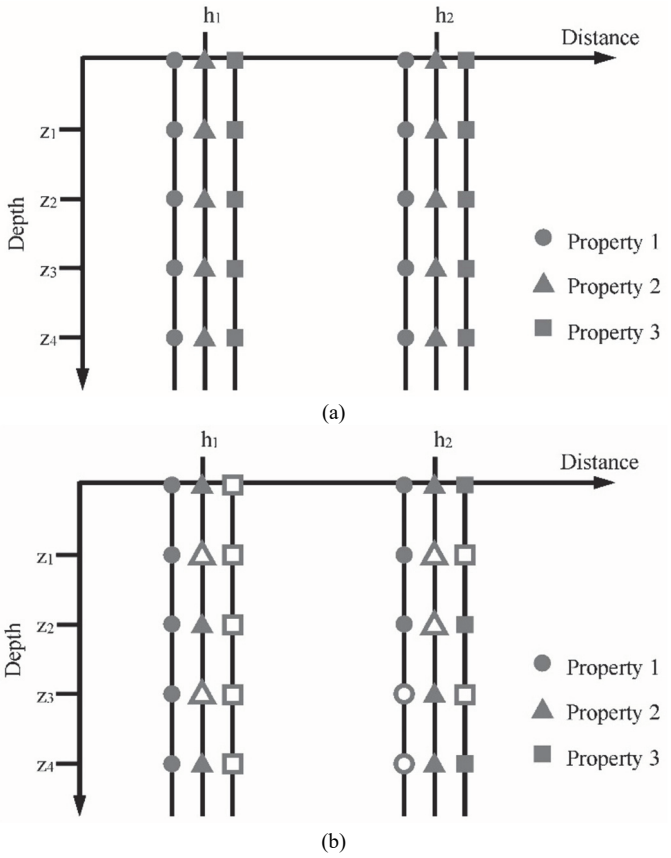


Fig. 1 Multivariate 2D spatially varying data: (a) complete and (b) incomplete (solid marker: observed location; open marker: unobserved location)

4. 2D VIRTUAL GROUND

To illustrate the performance of SBL and GPR-MUSIC-3X, the undrained shear strength (s_u) of a simple 2D virtual ground is created using a 2D quadratic trend in the depth (z) direction and a 1D linear trend in the horizontal (x) direction. The spatial trend for the cone tip resistance (q_t) has the same form but 10 times larger in magnitude, $t_{qt}(x, z) = 10 \times t_{su}(x, z)$, where $t(x, z)$ stands for the trend function. The undrained shear strength of the virtual ground is simulated by $s_u(x, z) = t_{su}(x, z) + w_{su}(x, z)$, where $w(x, z)$ stands for the spatial variability. Similarly, the cone tip resistance is simulated by $q_t(x, z) = t_{qt}(x, z) + w_{qt}(x, z)$. The two random fields $w_{su}(x, z)$ and $w_{qt}(x, z)$ are assumed to follow zero-mean Gaussian with σ = 10 and 100 kPa, respectively, with identical δ_v (vertical scale of fluctuation) = 0.5 m and identical δ_h (horizontal scale of fluctuation) = 5 m. The autocorrelation model is assumed to be single exponential. The two random fields $w_{su}(x, z)$ and $w_{qt}(x, z)$ have a positive cross correlation of 0.6. Figures 2 and 3 show the q_t and s_u data for the virtual ground. The virtual ground is "tested" at 5 locations along the x-axis (x = 0, 0.5, 1.5, 5, and 10 m). At each x location, both s_u and q_t data are available at a depth interval of 0.1 m. The SBL approach can only analyze one type of data (either s_u or q_t), so only the s_u data in Fig. 3 are analyzed. The SBL is trained by the s_u data at the 2 soundings at the horizontal coordinates x = 0 m and x = 10 m (dark lines in Fig. 3) and validated by the s_u data at the 3 sounding at x = 0.5 m, 1.5 m, and 5 m (red lines). The random field parameters (σ, δ_v, δ_h) as well as the trend function are treated as unknown during the SBL training.

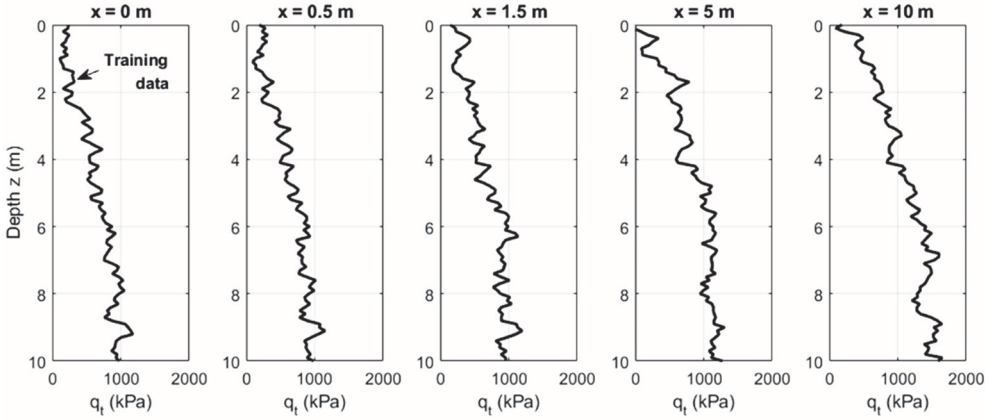


Fig. 2 q_t data for virtual ground

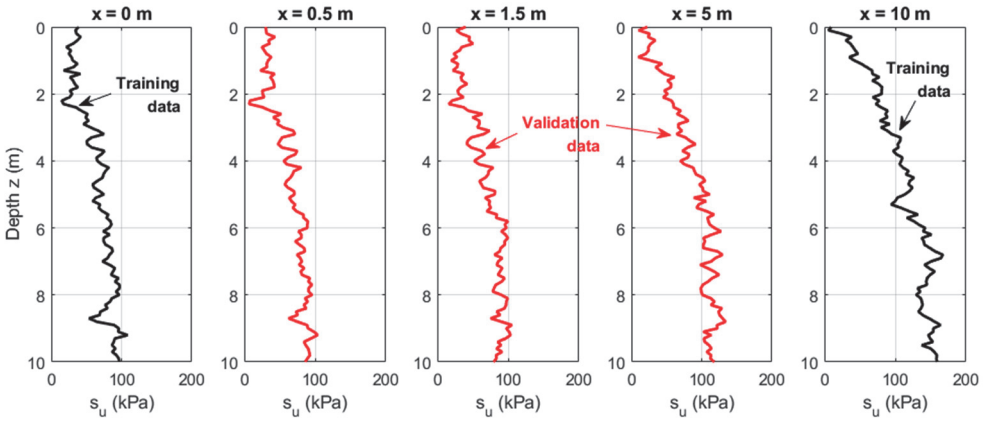


Fig. 3 s_u data for virtual ground

Given the training data (dark lines in Fig. 3), the SBL approach first identifies (σ, δ_v, δ_h) as well as the trend, then it simulates the conditional random fields of s_u. It is noteworthy that during this training and conditional simulation process, there is no need to prescribe the functional form of the trend (*e.g.,* linear or quadratic) or to estimate the coefficients of the trend separately using regression. The SBL can automatically detect the optimal form and establish the coefficients consistent with Eq. (1). Figure 4 shows the samples and histogram for the identified (σ, δ_v, δ_h). The red marker and line in the figure indicate the actual values of the random field parameters (σ, δ_v, δ_h) used to define the virtual ground. Figure 5 shows one realization of the conditional random field of s_u. It is remarkable that the conditional random field always passes through the training data.

Figure 5 only shows one realization of the 2D conditional random field of s_u. One thousand such random field samples are obtained, and Fig. 6 shows the resulting median s_u profiles at the 3 validation soundings (x = 0.5 m, 1.5 m, and 5 m) as the dark lines and the 95% Bayesian confidence intervals (CIs) as the dashed dark lines. The green lines represent one realization of the conditional random field sample. The red lines in the figure are the actual s_u data at the 3 validation soundings. Note that these data are treated as unknown during the SBL training. The two-step SBL approach is shown to be consistent in the well-defined sense that the resulting 95% Bayesian CIs contain the actual validation sounding data with a large chance (close to 0.95, as reported in Ching and Phoon 2017).

To illustrate GPR-MUSIC-3X, consider that a clay layer has s_u and q_t with clear depth trends shown in Fig. 7(a) and 7(b). Moreover, the trends at x = 0 m and 10 m are distinct. Although both s_u and q_t have clear depth trends, if they are plotted in a s_u-q_t plot (Fig. 7(c)), a unique bi-variate correlation exists. Therefore, the depth trends of both s_u and q_t can be explained away by the s_u-q_t correlation. Hence, it suffices to construct the s_u-q_t correlation. The GPR-MUSIC-3X approach is adopted to analyze the observed s_u data at the 2 soundings with x = 0 m and x = 10 m in Fig. 3 as well as the q_t data at the 5 soundings with x = 0, 0.5, 1.5, 5, and 10 m in Fig. 2. The GPR-MUSIC-3X approach will construct the s_u-q_t correlation using the s_u-q_t data. The approach can also further simulate the conditional random field of s_u at the 3 validation soundings (x = 0.5 m, 1.5 m, and 5 m), and Fig. 8 shows the resulting 95% Bayesian CIs at the 3 validation soundings. The 95% CIs are narrower than those in Fig. 6. This is because the q_t data at the 3 soundings are known, so the s_u values can be estimated more accurately by the s_u-q_t correlation constructed within GPR-MUSIC-3X. However, if the q_t data at the 3 soundings are unknown, the 95% Bayesian CIs for s_u become wide (Fig. 9), seemingly even wider than those shown in Fig. 6 produced by SBL. Note that GPR-MUSIC-3X adopts prescribed vertical and horizontal autocorrelation functions. It is uncertain if Fig. 9 will improve significantly when the vertical and horizontal scales of fluctuation are updated using data. Research is in progress to develop a full GPR-MUSIC-3X or FGPR-MUSIC-3X that allows the trend and both auto- and cross-correlations to be updated by data.

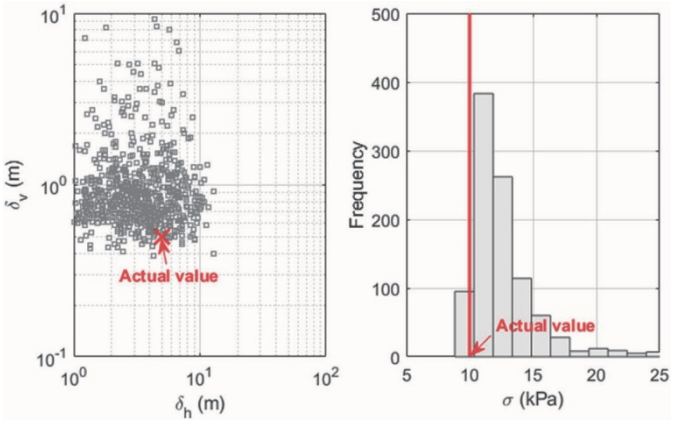


Fig. 4 Samples and histogram for identified random field parameters (σ, δ_v, δ_h) for s_u

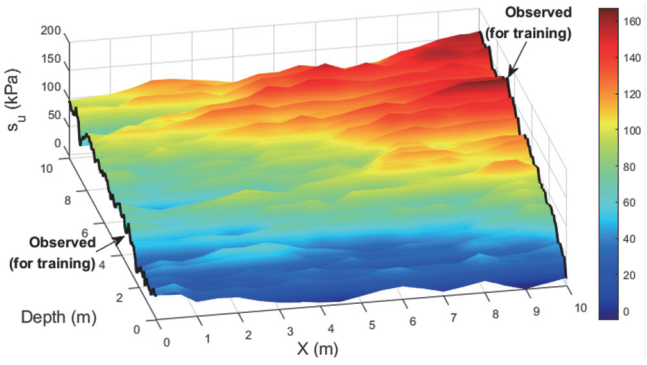


Fig. 5 One 2D realization of conditional random field of s_u

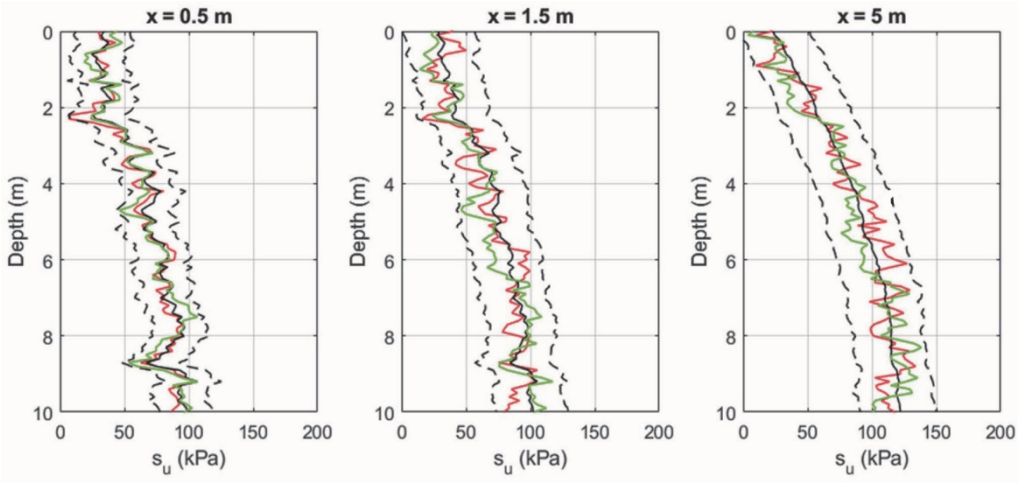


Fig. 6 95% Bayesian confidence intervals for 3 validation soundings produced by 3D SBL (dark lines are median, dark dashed lines are 95% CIs, green lines are random field realizations lines, and red lines are validation data)

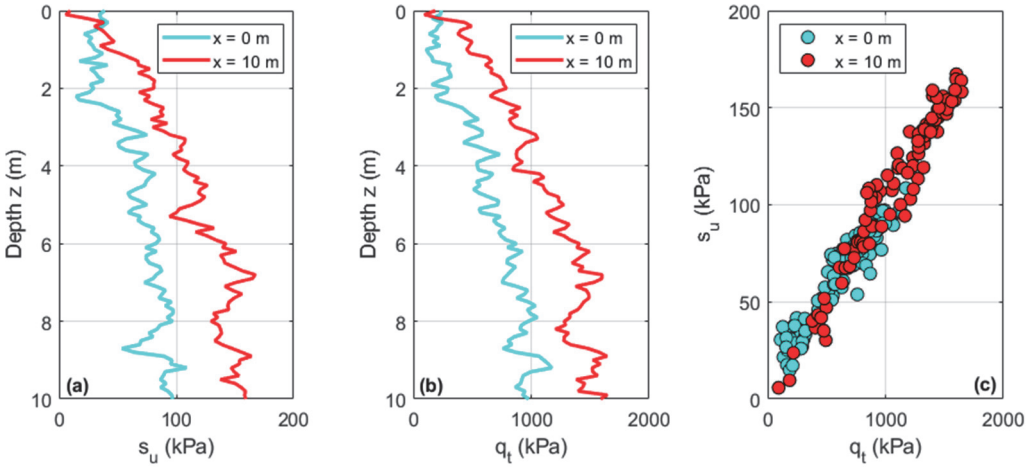


Fig. 7 Sounding data at x = 0 and 10 m: (a) s_u; (b) q_t; (c) s_u-q_t relation

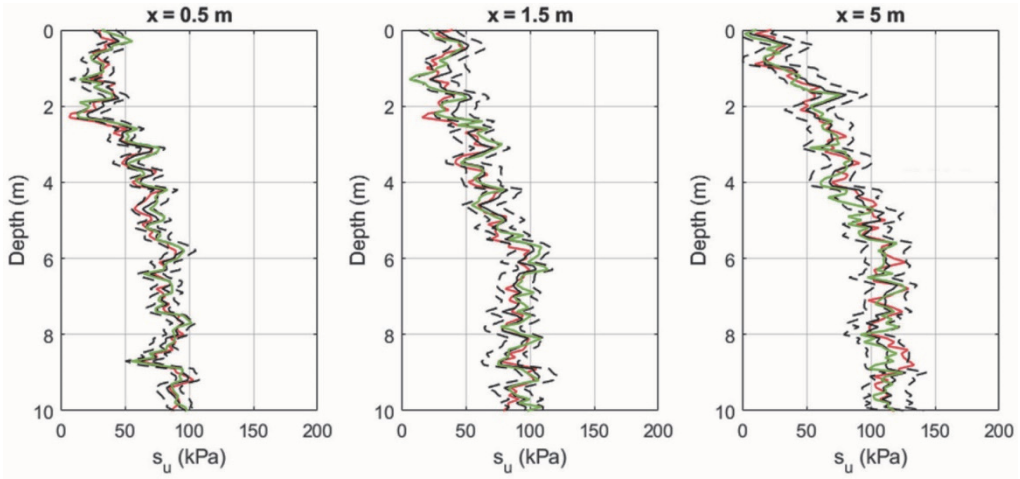


Fig. 8 95% Bayesian confidence intervals of s_u for 3 validation soundings produced by GPR-MUSIC-3X (dark lines are median, dark dashed lines are 95% CIs, green lines are random field realizations, and red lines are validation data)

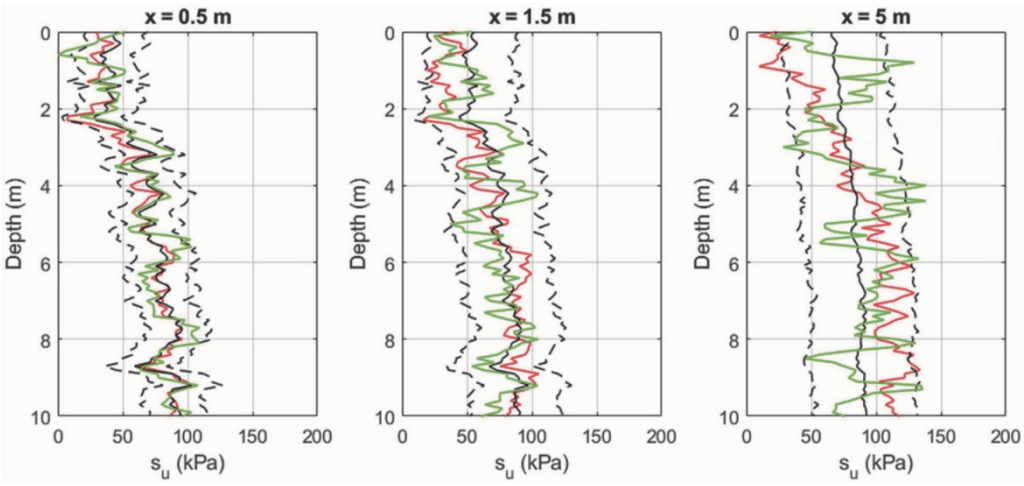


Fig. 9 95% Bayesian confidence intervals of s_u for 3 validation soundings produced by GPR-MUSIC-3X assuming q_t data at 3 soundings are unknown (dark lines are median, dark dashed lines are 95% CIs, green lines are random field realizations, and red lines are validation data)

5. 3D SPARSE BAYESIAN LEARNING

The SBL approach proposed by Ching and Phoon (2017) is applicable to 1D spatial variability only. Although it laid the theoretical basis for a nearly data-driven subsurface mapping approach, its value to practice is limited. Phoon *et al.* (2021) opined that "spatial variation is only meaningful to practice when expressed in 3D. A 1D spatially varying profile with constant properties in the horizontal plane is a possible simplification of a real profile. It can be viewed as a more realistic representation of the classical layered soil profile. Alternately, a layered profile can be viewed as an extremely coarse discretization of a spatially varying profile in the depth direction. However, a 2D spatial varying profile is an impossibility in practice. It is absurd to say that a randomly chosen 2D section is spatially variable while the out of plane dimension is not varying to preserve the plane strain assumption. There is no consistent 3D spatially varying soil mass that can produce such a 2D section with the possible exception of an engineered structure such as a levee."

Direct extension to 3D is non-trivial. The main challenge is computational, *i.e.,* 3D problems require numerical manipulations of very large matrices. This aspect deserves much more attention in the literature. There appears to be a widespread misconception that developing a new method is more challenging that making the method work in practice which would include making the method reasonably computable in 3D. This is not true (Ching *et al.* 2020, 2021d; Shuku and Phoon 2021). Consider a 3D example with 20 CPT soundings, and suppose that there are 500 data points for each sounding. If the maximum likelihood method is adopted, the computation may require repeated calculations of the inverse of a $(10{,}000 \times 10{,}000)$ matrix, which is costly and the matrix determinant is error-prone. Ching *et al.* (2020) shows that under the "separability" assumption between the z direction and (x, y) directions in the auto-correlation structure, it only requires inversions and Cholesky decompositions for two significantly smaller (500×500) and (20×20) matrices. Therefore, the computational cost and numerical errors for 3D probabilistic site characterization are significantly reduced. The "separability" assumption is widely adopted in the literature, although the authors are not aware of studies establishing its veracity. In addition to this assumption, a second "vertically-dense-lattice" assumption is needed for conditional random field simulation. "Vertically-dense" means the sampling interval in the depth direction should be smaller than the vertical scale of fluctuation. The definition of "lattice" data is shown in Fig. 10. The layout for the soundings can be arbitrary as shown in Fig. 10(a). This "vertically-dense-lattice" assumption can be satisfied by CPT soundings of equal lengths taken from a horizontal ground with no missing data as shown in Fig. 10(b). Conditional simulation is necessary, because a single most likely subsurface map will not alert the engineer to the presence of less likely maps that can be critical to the design. For example, a slope could be unstable if a thin horizontal weak layer were to exist below its toe. A range of simulated maps consistent with the observed soundings is a more appropriate representation of the underlying epistemic (statistical) uncertainties that can be significant under MUSIC-3X conditions.

6. FUTURE WORK

The 3D SBL approach proposed by Ching *et al.* (2020) requires lattice data (all soundings are carried out to the same depth with no missing data) to take advantage of the Kronecker-product derivations. In practice, non-lattice data are more common because soundings are carried out to different depths to cater to geologic variations and/or geotechnical engineering needs. The lattice assumption implies that all CPT soundings have to be truncated to match the length of the shortest sounding. This defeats the purposes of carrying out deeper soundings, which must be important to justify the additional costs. It is also possible for the soundings to be incomplete in the sense that some sections are not recorded. These soundings do not constitute "lattice" data as well (Fig. 10(d)). The lattice assumption severely restricts the value of 3D SBL in practice. Fortunately, this assumption was recently relaxed by Ching *et al.* (2021d). In summary, research in SBL has progressed to the following stage:

1. Incorporation of the Whittle-Matérn (WM) autocorrelation model parameterized by the scale of fluctuation (δ) and smoothness parameter (ν) — this is arguably the most general monotonic autocorrelation model that includes the common single and squared exponential models as special cases.

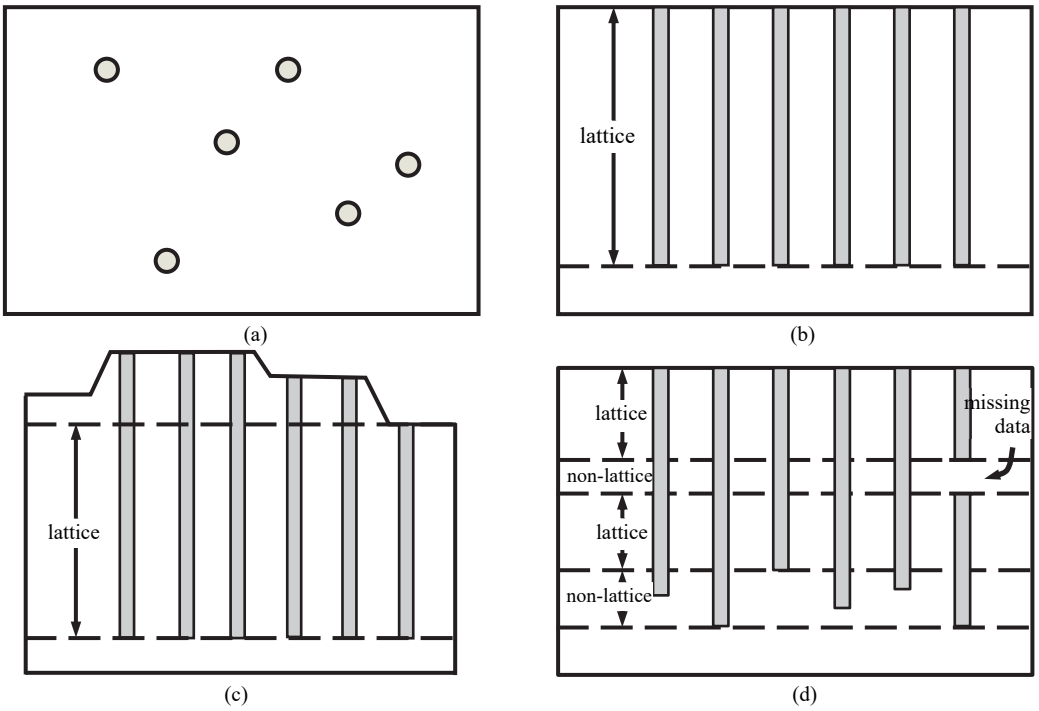


Fig. 10 Illustration of (a) layout of soundings (plan view), (b) lattice data (horizontal ground), (c) lattice data (non-horizontal ground), and (d) non-lattice data with unequal soundings and missing data

2. Fully consistent characterization of statistical uncertainties for the: (a) functional form of the trend, (b) coefficients of the trend function, and (c) random field parameters: standard deviation, scale of fluctuation, and smoothness parameter — this obviates the need to specify a minimum sample size. A small sample size will result in larger statistical uncertainties leading to more conservative designs to achieve the same target reliability index (Ching *et al.* 2014). It is not possible for an engineer to know how much and where data should be collected to lead to sufficiently precise solutions for deterministic analysis. The authors submit that this common "minimum sample size" question imposed by a deterministic paradigm is not useful. It is more useful for a data-driven method to inform the engineer what is the precision of the solution based on available data and to guide the engineer to collect more data to achieve a desired precision level. The engineer is well placed to decide the level of precision achievable based on the project budget.
3. Conditional simulation of the subsurface map — this is necessary to represent the effect of the statistical uncertainties on the map correctly. A less likely map is not a less important map in terms of design consequences. Hence, in the opinion of the authors, the inability to simulate a range of maps is a major problem, although it has rarely been emphasized in the literature.
4. Stratification — when Y denotes the soil behavior type index (I_c), the subsurface map becomes a stratigraphy. This is a special application of SBL. The soil behavior type index is defined in Table 1.
5. 3D mapping — this allows all CPT soundings in a site to be used as inputs directly and equally importantly, 3D SBL is reasonably computable. There is no need to draw 2D sections. This is difficult to do in practice, because a typical CPT layout does not follow a regular rectangular grid.

Table 1 CPT-based soil behaviour type index (Robertson 2016; Robertson and Wride 1998; Robertson 1990)

Soil behavior type index, I_c	Zone	Soil behavior type (SBT)
–	9	Very stiff fine-grained
–	8	Very stiff sand to clayey sands
$I_c < 1.31$	7	Gravelly sand to dense sand
$1.31 < I_c < 2.05$	6	Sands: clean sand to silty sand
$2.05 < I_c < 2.60$	5	Sand mixtures: silty sand to sandy silt
$2.60 < I_c < 2.95$	4	Silt mixtures: clayey silt to silty clay
$2.95 < I_c < 3.60$	3	Clays: silty clay to clay
$I_c > 3.60$	2	Organic soils: peats
–	1	Sensitive fine-grained

6. Non-lattice data — this allows CPT soundings of unequal lengths with missing sections to be used as inputs directly. 3D SBL is less computationally efficient in the presence of non-lattice data, but it is a major step forward as it addresses the "Incompleteness" (I) attribute in "MUSIC-3X" and brings 3D SBL closer to full MUSIC-3X compliance.

The current 3D SBL approach is not applicable under the following conditions:

1. It does not apply to multivariate data at the current stage of development.
2. Autocorrelation models that do not fall under the WM model. One example is a non-monotonic autocorrelation model such as the cosine WM model (Chang *et al.* 2021). It is straightforward to replace WM by cosine WM.
3. Non-separable autocorrelation models. No solution is available at present.
4. Trend functions that do not allow a sparse representation under the chosen basis functions. One example is a 3D trend function with rapid changes. No solution is available at present.

5. Spatial variation is non-stationary. It is possible to transform the observations Y so that it is stationary in some cases.
6. Test data that do not satisfy the "vertically-dense-lattice" assumption. It is possible to relax this assumption into the "lattice" assumption if a different conditional random field simulation algorithm is adopted. Research is in progress on this front.
7. Stratigraphy is restricted to the soil behavior type index (I_c) only. Two limitations are evident. One, only CPT data can be used to inform stratigraphy analysis. Second, the transformation model relating I_c to soil behavior type is generic. The "Multivariate" (M) and "Uniqueness" (U) attributes in MUSIC-3X are not addressed.

The site "Uniqueness" attribute (U) and its relation to 3D SBL deserves more elaboration here. All sites are unique to some degree, but they are not completely unique. Phoon (2018) posed the site challenge in an editorial for a special collection on probabilistic site characterization. The challenge is to *quantify* "site uniqueness", directly or indirectly, so that big indirect databases (BIDs) can be combined with sparse site-specific (local) data in a manner sensitive to site differences. This idea is not new as geotechnical and rock engineers have been relying on data from similar sites to inform their understanding of a current site. Phoon *et al.* (2021) named this challenge as a "site recognition" challenge. Thus far, research is focused on developing quasi-site-specific transformation models from local data and BIDs (Ching and Phoon 2020b; Ching *et al.* 2021b, 2021c). Phoon (2020) provided a useful overview of BIDs for soil/rock properties. The databases are labelled as (geo-material type)/(number of parameters of interest)/(number of data points). For example, the CLAY/10/7490 database consists of 7490 records from 251 studies carried out in 30 countries (Ching and Phoon 2014). Each record contains ten clay parameters measured at roughly the same depth, although some may be missing. However, these BIDs do not contain sufficient information for subsurface mapping. Thus far, to the authors' knowledge, no research has been conducted to quantify "site uniqueness" for a subsurface map. This is entirely possible, because sites containing the same geology should be more "similar" (or less "unique") and combining subsurface maps from similar sites should reduce statistical uncertainties at any one site. It is possible to use information from other sites to define the prior distributions for the weights of the trend function (w_k) and the random field parameters (σ, δ, ν), but a better approach is to identify "similar" sites and only use the information from these sites as prior. The key question here is whether identification of "similar" sites is best carried out by looking at the prior distributions of the model parameters (w_k, σ, δ, ν), the original observation data Y, the subsurface maps produced by conditional simulation, or others. Ching *et al.* (2021a) has generalized GPR-MUSIC-3X to HBM-MUSIC-3X recently to allow cross-correlations at a specific site to be updated by information from similar sites in a generic database. In contrast to 3D SBL, this HBM-MUSIC-3X can address the "Multivariate" (M) and "Uniqueness" (U) attributes in MUSIC-3X. For stratigraphic mapping using CPT soundings, Phoon *et al.* (2021) noted that the soil behavior type index (I_c) is a generic transformation model. Hence, one aspect of stratigraphic mapping is not site-specific as noted above.

7. CASE STUDY

The 2D virtual ground example in the previous section was analyzed using 3D SBL and GPR-MUSIC-3X (Ching *et al.* 2021a). An actual case study is presented next to demonstrate the applicability of 3D SBL to practice. This test site is located at Baytown, Texas, USA (Stuedlein *et al.* 2012). The site exploration plan is shown in Fig. 11, and soil profile at the A-A′ section is shown in Fig. 12. The soil at the test site is mainly clay with occasional layers of silt and fine sand. The test site was characterized by 9 CPTs in an area of 15 m × 30 m. The cone tip resistance (q_t) and sleeve friction (f_s) data are converted to the soil behavior type index (I_c) proposed by Robertson and Wride (1998) (Table 1). The sampling depth interval for the original data is 0.02 m, but the data are resampled by a sampling interval of 0.05 m to reduce computation time. Among the 9 soundings, 3 of them (CPT-1 to CPT-3) are deep soundings, whereas the remaining 6 soundings (CPT-F1 to CPT-F6) are relatively shallow. The I_c profiles of these soundings are shown in Fig. 13. Robertson's soil behavior types (SBT = 2 to 7) are also annotated in the figure.

Note that the I_c data do not satisfy the lattice data requirement, so the algorithm proposed by Ching *et al.* (2021d) is adopted. The WM model is adopted as the auto-correlation model, so there are in total 5 parameters for the auto-covariance: (σ, δ_v, δ_h, ν_v, ν_h), where ν_v and ν_h are the vertical and horizontal smoothness parameters, respectively. The CPT-F3 sounding is taken to be the validation sounding. It is regarded as unknown during 3D SBL. The SBL approach is trained using the remaining 8 soundings. Figure 14 shows the identification results for (σ, δ_v, δ_h, ν_v, ν_h) by 3D SBL. Figure 15 shows the resulting 95% Bayesian confidence intervals at the validation sounding (CPT-F3). It is clear that the resulting 95% Bayesian confidence interval mostly contains the actual validation sounding data. Figure 16(a) shows the conditional random field for I_c on the A-A′ cross section in Fig. 12. It is clear that the conditional random field sample coincides with the observed I_c values at the sounding locations (CPT-1 to CPT-3). One thousand such conditional random field samples are obtained by 3D

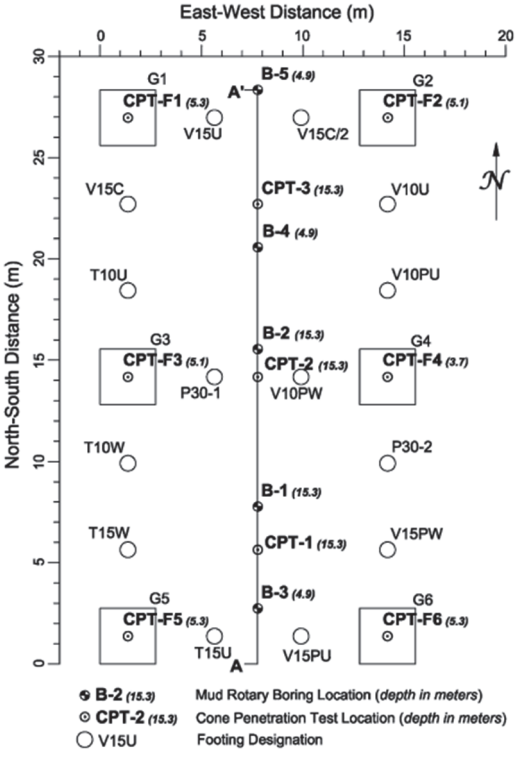


Fig. 11 Exploration plan for Baytown site (Stuedlein *et al.* 2012)

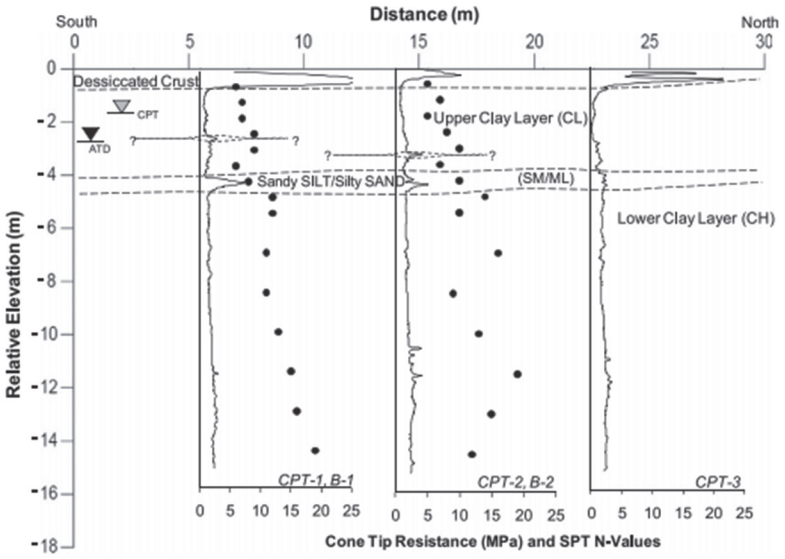


Fig. 12 Soil profile at A-A′ section (Stuedlein *et al.* 2012)

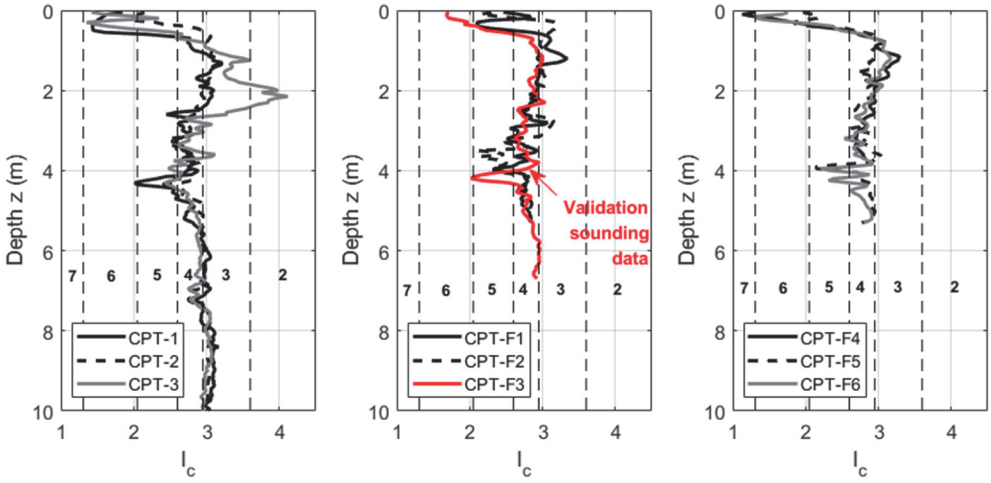


Fig.13 I_c data for Baytown site (labels 2 to 7 denote Robertson's SBTs)

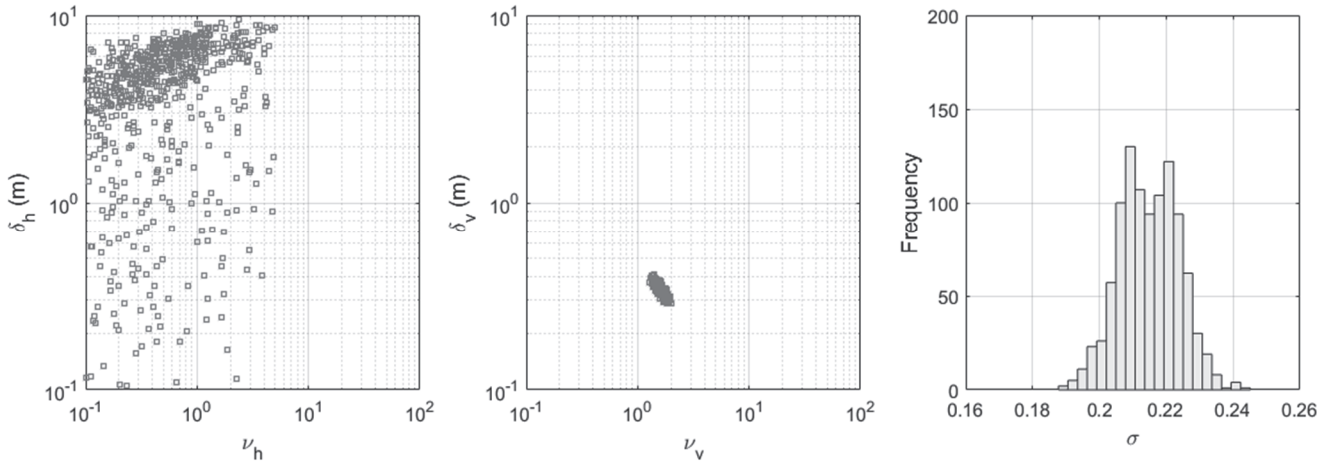


Fig. 14 Samples and histogram for identified random field parameters (σ, δ_v, δ_h, ν_v, ν_h)

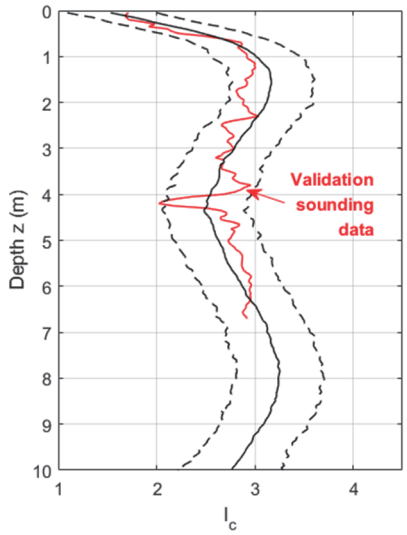
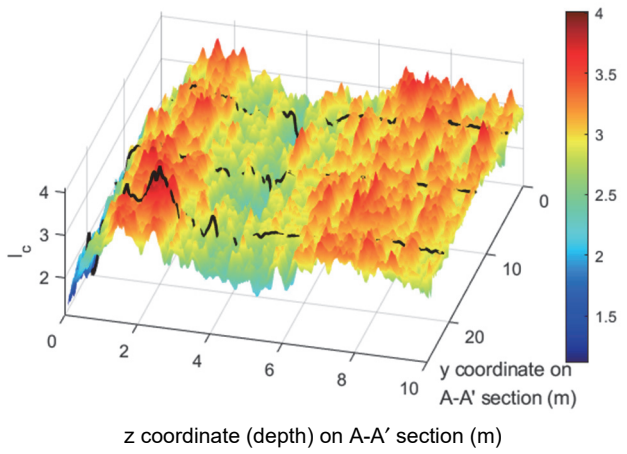


Fig. 15 95% Bayesian confidence interval for validation sounding (CPT-F3)

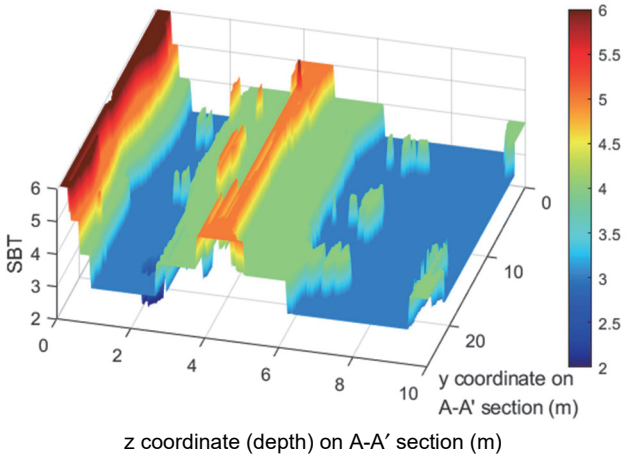
SBL, and they can be converted to 1000 SBT samples. Based on the SBT samples, the most probable SBT (SBT = 2 to 7) at any location can be determined. Figure 16(b) shows the most probable SBT on the A-A′ section. It is clear that the soil is mostly clay (SBT = 3 "Clays: silty clay to clay") to silty clay (SBT = 4 "Silt mixtures: clayey silt to silty clay"). There seems to be a thin layer of silt (SBT = 5 "Sand mixtures: silty sand to sandy silt") at the depth of about 4 m and a crustal layer of sand (SBT = 6 "Sands: clean sand to silty sand") near the ground surface. This is consistent with Fig. 12. There may be a lens of peat at the depth of about 2 m at the north of the site where the north-south distance is large. Ching *et al.* (2021a) analyzed the same site using GPR-MUSIC-3X and HBM-MUSIC-3X.

8. CONCLUSIONS

A subsurface map seeks to quantify some geometrical features, material behaviors, and geoenvironmental processes at a specific site. The current practice is based on an assortment of prior knowledge (regional geology), observations (neighbouring sites, open cuts, borelogs), test data (CPT, geophysical), geostatistics (kriging), and engineering judgment. Phoon *et al.* (2021) defined data-driven site characterization (DDSC) as any site characterization methodology that relies solely on measured data, both site-specific data collected for the current project and existing data of any type collected from past stages of the same project or past projects at the same site, neighboring sites, or beyond. One key complication is that real data is "ugly". A useful mnemonic is MUSIC-3X (Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted with "3X" denoting three dimensional spatial variations). It is an open question whether DDSC can solve real world 3D subsurface mapping problems based on real world MUSIC-3X data with minimum ad-hoc assumptions. The computational challenges are very significant, but some reasonable partial solutions have been obtained recently in Project DeepGeo (inspired by DeepMind that produces AlphaGo), which constitutes one major research effort in the emerging field of data-centric geotechnics.



(a)



(b)

Fig. 16 (a) Conditional random field sample for I_c on the A-A′ cross section; (b) most probable SBT on the A-A′ section

The first solution is Sparse Bayesian Learning (SBL). It has the potential to simulate subsurface maps containing stratification and mechanical properties conditioned on one type of field test data, for example cone tip resistance soundings. SBL is considered to be "nearly data driven", because the following features are "learned" from data: (1) functional form of the trend, (b) coefficients of the trend function, and (c) random field parameters: standard deviation, scale of fluctuation, and smoothness parameter. In contrast, the current practice is to assume a trend function, say a linear function, compute the coefficients of the trend function using regression that contradicts spatial correlations, and characterize the random field parameters using the method of moments. Statistical uncertainties are not considered, although they are significant for MUSIC-3X data. In addition, statistical uncertainties are crucial for decision making. It is not appropriate to impose on an engineer the popular "minimum sample size" question — how much data is needed for a solution to be useful? The onus is exactly opposite. The burden is on an analysis to inform the engineer what is the solution precision associated with the data on hand. The appropriate decision to impose on an engineer is to judge whether this precision is sufficient and if not, what additional measures should be undertaken to address this issue.

The SBL has recently evolved into 3D SBL, which is a significant step forward because large scale calculations are usually too costly. Another recent breakthrough is the relaxation of the "lattice data" assumption. Research is underway to relax the "vertically dense" assumption, which restricts 3D SBL to CPT data. Overall, more research is needed, because 3D SBL is not fully MUSIC-3X compliant, although it can handle sparse, incomplete, and spatially varying CPT data. In particular, it can only handle one type of field test data. It is also not fully data-driven, because ad-hoc assumptions such as a separable autocorrelation model are embedded in the current version. Nonetheless, it is already useful for practice. A 3D SBL version would be made available in Rocscience's Settle3 (three-dimensional soil settlement analysis) in the near future to generate subsurface maps based on CPT data.

A second solution is based on a variant of the Gaussian process regression (GPR-MUSIC-3X) (Ching *et al.* 2021a). It can handle multiple field test data by learning the cross-correlation behavior among different soil parameters at a single site of interest. GPR-MUSIC-3X can be enhanced to learn cross-correlation behaviors at multiple sites and thus bring information from "similar" sites in a larger generic database to bear on improving predictions at a single site (HBM-MUSIC-3X). In contrast, 3D SBL can only learn from site-specific data — it is unable to benefit from site data found in other sites. The results produced by 3D SBL and GPR-MUSIC-3X are shown to be reasonable using cross validation for the case of a 2D virtual ground and an actual 3D site at Baytown, Texas, USA. Table 2 presents a list of available methods in Project DeepGeo and the capabilities/limitations of each method.

FUNDING

The second author would like to thank the gracious support from the Ministry of Science and Technology of Taiwan (106-2221-E-002-084-MY3, 107-2221-E-002-053-MY3 & 109-2221-E-002-029-MY3).

DATA AVAILABILITY

The computer codes used in this study are available from the corresponding author on reasonable request.

CONFLICT OF INTEREST STATEMENT

The authors certify that there is no conflict of interest.

ACKNOWLEDGMENTS

The authors thank Dr. Thamer Yacoub, CEO & President of Rocscience, for his kind invitation to present this keynote paper in the Rocscience International Conference in 2021 to mark the occasion of Rocscience's 25th anniversary. In addition, the authors had frutitful exchanges with Dr. Thamer Yacoub and Dr. Sina Javankhoshdel that resulted in an ongoing collaboration to bring 3D SBL to practice using Settle3. This paper is an expanded version of a keynote paper published in "Phoon, K.K. and Ching, J. (2021). 'Advances in data-driven subsurface mapping'. In Reginald Hammah, Thamer Yacoub, John Curran and Alison McQuillan, Eds., *The Evolution of Geotech: 25 Years of Innovation; Proceedings of Rocscience International Conference 2021*, Toronto, 20-21 April 2021. Rotterdam: Balkema"

Table 2 Data-driven methods for site characterization in Project DeepGeo

Method	No. of parameters	Spatial variability		Use generic database?	Other limitations	Reference
		Trend	Auto-correlation			
Sparse Bayesian Learning (SBL)	Single	Yes	1X	No	Stationary autocorrelation	Ching and Phoon (2017)
3D Sparse Bayesian Learning (3D SBL)	Single	Yes	3X	No	Stationary separable autocorrelation; vertically dense; lattice	Ching *et al.* (2020)
3D Sparse Bayesian Learning (3D SBL)	Single	Yes	3X	No	Stationary separable autocorrelation; vertically dense	Ching *et al.* (2021d)
Gaussian Process Regression (GPR-MUSIC)	Multiple	–	–	No	No vertical auto correlation; perfect horizontal autocorrelation	Ching and Phoon (2019)
Gaussian Process Regression (GPR-MUSIC-X)	Multiple	No	1X	No	Prescribed stationary vertical autocorrelation; perfect horizontal autocorrelation	Ching and Phoon (2020a)
Gaussian Process Regression (GPR-MUSIC-3X)	Multiple	No	3X	No	Prescribed stationary separable autocorrelation	Ching *et al.* (2021a)
Gaussian Process Regression + Hierarchical Bayesian Model (HBM-MUSIC)	Multiple	–	–	Yes	No vertical auto correlation; perfect horizontal autocorrelation	Ching *et al.* (2021b, 2021c)
Gaussian Process Regression + Hierarchical Bayesian Model (HBM-MUSIC-3X)	Multiple	No	3X	Yes	Prescribed stationary separable autocorrelation	Ching *et al.* (2021a)
Full Gaussian Process Regression (FGPR-MUSIC-3X)	Multiple	Yes	3X	No	Stationary separable autocorrelation	In progress

REFERENCES

Cami, B., Javankhoshdel, S., Phoon, K.K., and Ching, J. (2020). "Scale of fluctuation for spatially varying soils: estimation methods and values." ASCE-*ASME Journal of Risk and Uncertainty in Engineering Systems*, Part A: Civil Engineering, **6**(4), 03120002. https://doi.org/10.1061/AJRUA6.0001083

Ching, J. and Chen, Y.C. (2007). "Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection and model averaging." *Journal of Engineering Mechanics*, ASCE, **133**(7), 816-832. https://doi.org/10.1061/(ASCE)0733-9399(2007)133:7(816)

Ching, J. and Phoon, K.K. (2012). "Establishment of generic transformations for geotechnical design parameters." *Structural Safety*, **35**, 52-62. https://doi.org/10.1016/j.strusafe.2011.12.003

Ching, J., Phoon, K.K., and Yu, J.W. (2014). "Linking site investigation efforts to final design savings with simplified reliability-based design methods." *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE, **140**(3), 04013032. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001049

Ching, J. and Phoon, K.K. (2014). "Correlations among some clay parameters – the multivariate distribution." *Canadian Geotechnical Journal*, **51**(6), 686-704. https://doi.org/10.1139/cgj-2013-0353

Ching, J. and Phoon, K.K. (2015). "Constructing multivariate distribution for soil parameters." Chapter 1, *Risk and Reliability in Geotechnical Engineering*, 3-76. CRC Press/Balkema.

Ching, J. and Phoon, K.K. (2019). "Constructing site-specific multivariate probability distribution model using Bayesian machine learning." *Journal of Engineering Mechanics*, ASCE, **145**(1), 04018126. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001537

Ching, J., Phoon, K.K., and Chen, C.H. (2014). "Modeling piezocone cone penetration (CPTU) parameters of clays as a multivariate normal distribution." *Canadian Geotechnical Journal*, **51**(1), 77-91. https://doi.org/10.1139/cgj-2012-0259

Ching, J. and Phoon, K.K. (2017). "Characterizing uncertain site-specific trend function by Sparse Bayesian Learning." *Journal of Engineering Mechanics*, ASCE, **143**(7), 04017028. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001240

Ching, J. and Phoon, K.K. (2019). "Impact of auto-correlation function model on the probability of failure." *Journal of Engineering Mechanics*, ASCE, **145**(1), 04018123. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001549

Ching, J. and Phoon, K.K. (2020a). "Constructing a site-specific multivariate probability distribution using sparse, incomplete, and spatially variable (MUSIC-X) data." *Journal of Engineering Mechanics*, ASCE, **146**(7), 04020061. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001779

Ching, J. and Phoon, K.K. (2020b). "Measuring similarity between site-specific data and records from other sites." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems*, Part A: Civil Engineering, **6**(2), 04020011. https://doi.org/10.1061/AJRUA6.0001046

Ching J., Wu, S.S., and Phoon, K.K. (2016). "Statistical characterization of random field parameters using frequentist and Bayesian approaches." *Canadian Geotechnical Journal*, **53**(2), 285-298. https://doi.org/10.1139/cgj-2015-0094

Ching, J., Phoon, K.K., Beck, J.L., and Huang, Y. (2017). "Identifiability of geotechnical site-specific trend functions." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems*, Part A: Civil Engineering, **3**(4), 04017021. https://doi.org/10.1061/AJRUA6.0000926

Ching, J., Phoon, K.K., Stuedlein, A.W., and Jaksa, M. (2019). "Identification of sample path smoothness in soil spatial variability." *Structural Safety*, **81**, 101870. https://doi.org/10.1016/j.strusafe.2019.101870

Ching, J., Huang, W.H., and Phoon, K.K. (2020). "3D probabilistic site characterization by Sparse Bayesian Learning." *Journal of Engineering Mechanics*, ASCE, **146**(12), 04020134. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001859

Ching, J., Phoon, K.K., Yang, Z., and Stuedlein, A.W. (2021a). "Constructing a quasi-site-specific multivariate probability distribution model for soil properties using sparse, incomplete, and three-dimensional (MUSIC-3X) data." *Georisk*, in review.

Ching, J., Wu, S., and Phoon, K.K. (2021b). "Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model." *Journal of Engineering Mechanics*, ASCE, in press.

Ching, J., Phoon, K.K., Ho, Y.H., and Weng, M.C. (2021c). "Quasi-site-specific prediction for deformation modulus of rock mass." *Canadian Geotechnical Journal*, in press. https://doi.org/10.1139/cgj-2020-0168

Ching, J., Yang, Z.Y., and Phoon, K.K. (2021d). "Dealing with non-lattice data in three-dimensional probabilistic site characterization." *Journal of Engineering Mechanics*, ASCE, **147**(5), 06021003. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001907

Jaksa, M.B., Kaggwa, W.S., and Brooker, P.I. (1999). "Experimental evaluation of the scale of fluctuation of a stiff clay." *Proceedings, 8th International Conference on Application of Statistics and Probability*, 415-422.

Kulhawy, F.H. and Mayne, P.W. (1990). "Manual on estimating soil properties for foundation design." Report EL-6800, *Electric Power Research Institute*, Palo Alto, California.

Lumb, P. (1966). "Variability of natural soils." *Canadian Geotechnical Journal*, **3**(2), 74-97. https://doi.org/10.1139/t66-009

National Research Council (1995). *Probabilistic Methods in Geotechnical Engineering*. National Academies Press, Washington, DC.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. In J. M. Bernardo, *et al*., Eds., *Bayesian Statistics 6*, Oxford University Press, 475-501.

Phoon, K.K. (2017). "Role of reliability calculations in geotechnical design." *Georisk*, **11**(1), 4-21. https://doi.org/10.1080/17499518.2016.1265653

Phoon, K.K. (2018). "Editorial for special collection on probabilistic site characterization." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems*, Part A: Civil Engineering, **4**(4), 02018002. https://doi.org/10.1061/AJRUA6.0000992

Phoon, K.K., (2020). "The story of statistics in geotechnical engineering." *Georisk*, **14**(1), 3-25. https://doi.org/10.1080/17499518.2019.1700423

Phoon, K.K., Quek, S.T., and An, P. (2003). "Identification of statistically homogeneous soil layers using modified Bartlett statistics." *Journal of Geotechnical and Geoenvironmental Engineering*, ASCE, **129**(7), 649-659. https://doi.org/10.1061/(ASCE)1090-0241(2003)129:7(649)

Phoon, K. K., Low, H. E., and Tan, T. S. (2003). "Estimation of overconsolidation ratio from piezocone tests in Singapore Marine Clays." *Proceedings of the Sixth International Symposium on Field Measurements in Geomechanics*, Oslo, September 15-18, 2003, 287-292.

Phoon, K.K., Ching, J., and Shuku, T. (2021). "Challenges in data-driven site characterization." *Georisk*, in press. https://doi.org/10.1080/17499518.2021.1896005

Rasmussen, C.E. and Williams, C.K.I. (2016). *Gaussian Processes for Machine Learning*. The MIT Press.

Robertson, P.K. (2016). "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — an update." *Canadian Geotechnical Journal*, **53**(12), 1910-1927. https://doi.org/10.1139/cgj-2016-0044

Robertson, P.K. (1990). "Soil classification using the cone penetration test." *Canadian Geotechnical Journal*, **27**, 151-158. https://doi.org/10.1139/t90-014

Robertson, P.K. and Wride, C.E. (1998). "Evaluating cyclic liquefaction potential using the cone penetration test." *Canadian Geotechnical Journal*, **35**, 442-459. https://doi.org/10.1139/t98-017

Shuku, T. and Phoon, K. K. (2021). "Three-dimensional subsurface modeling using geotechnical Lasso." *Computers and Geotechnics*, **133**(1), 104068. https://doi.org/10.1016/j.compgeo.2021.104068

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*, ASCE, **138**(11), 1301-1313. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000723

Tipping, M.E. (2001). "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, **1**, 211-244.