Recent advancements in data-driven site characterization

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

[1]*National Taiwan University, Department of Civil Engineering, Taipei, Taiwan*
[2]*Singapore University of Technology and Design, Singapore*
[#]*Corresponding author: jyching@ntu.edu.tw*

ABSTRACT

This paper reviews some recent advancements that address the challenges faced by the broad application area of data-driven site characterization (DDSC). The challenges include the ugly-data challenge, site-recognition challenge, and stratification challenge. The ugly-data challenge is about the MUSIC-3X attributes of the site investigation data, where MUSIC-3X stands for multivariate, uncertain and unique, sparse, incomplete, possibly corrupted, and 3D spatial variability (3X). The site-recognition challenge is about the site-uniqueness feature of the site investigation data. The stratification challenge is about the task of layer delineation in soil profiling. In recent years, some studies have been conducted to address these challenges with an encouraging degree of success, which are briefly reviewed in this paper. However, there are still unresolved issues yet to be addressed, which are briefly summarized in this paper as well.

Keywords: Data-driven site characterization; MUSIC-3X; site-uniqueness; stratification.

1. Introduction

Phoon et al. (2022a) published a paper entitled "Challenges in data-driven site characterization (DDSC)". In this paper, three challenges are stated:

- Ugly-data challenge: The geotechnical site investigation data are "ugly" in the sense that they have some realistic attributes that are not "ideal" for geotechnical analysis. Ideal data are abundant, complete, certain, free of outliers, and statistically independent. However, realistic site investigation data are not ideal but rather "ugly". They are MUSIC-3X, which stands for multivariate (multiple types of tests are conducted), uncertain and unique (quantification of site-specificity in the face of uncertainties), sparse (only limited boreholes and soundings are conducted), incomplete (not all multivariate soil parameters are observed at each location), possibly corrupted (presence of outliers), and spatially variable in the three-dimensional space (3X). DDSC methods need to address this ugly-data challenge.
- Site-recognition challenge: Site-uniqueness is a well-known feature in geotechnical engineering, which means that the site investigation data from site A cannot be directly adopted by site B. This is the "U" aspect in MUSIC-3X. This leads to a difficulty in practice: the site-specific data are sparse (the "S" aspect in MUSIC-3X), and these sparse site-specific data alone are typically insufficient to support reliable decision making. Conceptually, it is possible to find sites or data "similar to" the target site to support decision making or to transfer the knowledge learned from other sites to the target site (transfer learning). However, what are these "similar" sites or data? How to define the similarity? How to do the transfer learning?
- Stratification challenge: One important step in site characterization is to delineate soil layers based on the site-specific data. The core of the challenge lies in two aspects: (a) the significant complexity of the geological formation of the ground; (b) the MUSIC-3X nature of the site-specific data, especially the multivariate, sparse, and incomplete attributes.

The purpose of the current paper is two-fold: (a) review recent advancements that address these challenges; (b) outline unresolved issues related to these challenges and discuss possible future directions.

2. Ugly-data challenge

2.1. MUSIC-3X site-specific data

To showcase the ugly-data challenge, the MUSIC-3X attributes of typical site-specific data are illustrated by a real case history of a test site at Baytown, Texas, USA (Stuedlein et al. 2012). Figure 1 shows the site investigation plan: it consists of 5 boreholes (B-1 to B-5) and 9 cone penetration tests (CPTs). At B-1 and B-2, Atterberg limits (LL and PL) and water content (w) are available (Figures 2a and 2b). At B-3 to B-5, preconsolidation stress (σ'_p) (Figure 2c) and undrained shear strength (s_u) (Figure 2d) data are available at some depths. Figure 3 shows the CPT data. This site is mainly a clay site, with a thin silty-sand layer from the depth of 3.4 to 4.7 m. It is clear that the site investigation data are multivariate (Atterberg limits, water content, preconsolidation stress, undrained shear strength, CPT data, etc.) and sparse (limited boreholes and CPTs). The incomplete attribute can be seen more clearly from Table 1, the data at borehole B-1. It is clear that at B-1 locations,

only LL, PL, and w are known, but σ'_p, s_u, and CPT parameters are missing (incompleteness).

2.2. Cross-correlation and auto-correlation

In the opinion of the authors, the data-driven site characterization can be achieved if the soil parameters at locations unexplored by the boreholes and CPTs can be simulated by conditioning on the site-specific borehole and CPT data. This requires the following items:

- Estimation of the site-specific cross-correlation based on the MUSIC-3X data (cross-correlation denotes the correlation relationship between soil parameters).
- Estimation of the site-specific auto-correlation based on the MUSIC-3X data (auto-correlation denotes the spatial correlation between different locations in space).
- Ability to simulate conditional random fields (CRFs) conditioning on the site-specific MUSIC-3X data based on the site-specific cross-correlation and auto-correlation.

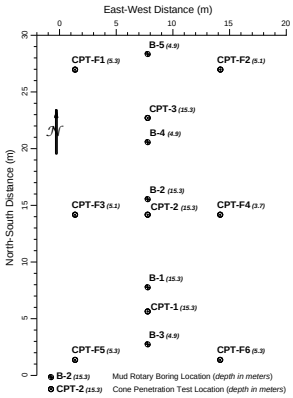


Figure 1. Site investigation plan for Baytown site (from Stuedlein et al. 2012).

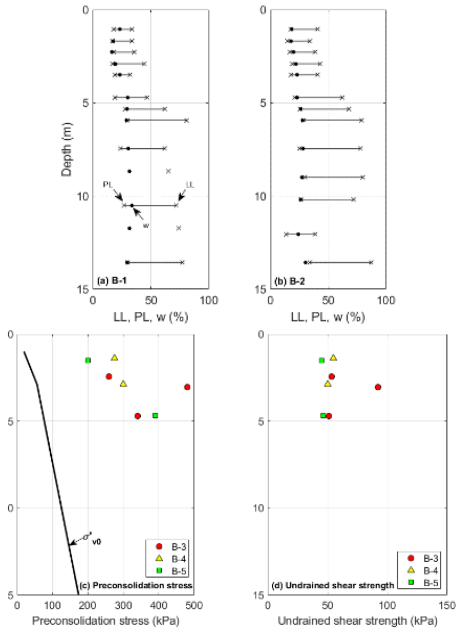


Figure 2. Borehole data for Baytown site (source: Stuedlein et al. 2012).

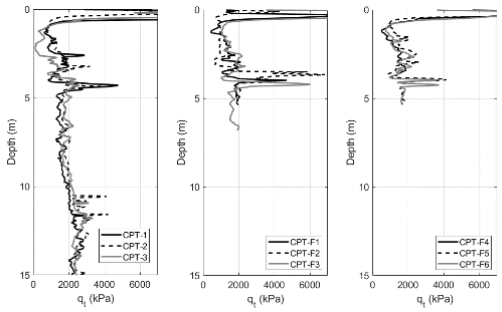


Figure 3. CPTs for Baytown site (source: Stuedlein et al. 2012).

Table 1. Data at B-1 (source: Stuedlein et al. 2012).

Depth (m)	LL (%)	PL (%)	w (%)	σ'_p (kPa)	s_u (kPa)	q_t (MPa)
1.06	34	18	23.4	-	-	-
1.68	34	17	17.8	-	-	-
2.28	36	18	16.6	-	-	-
2.90	44	17	19.7	-	-	-
4.72	47	19	29.9	-	-	-
5.34	62	28	29.5	-	-	-
5.94	81	30	29.3	-	-	-
7.46	62	24	30.5	-	-	-
8.68	65	-	31.5	-	-	-
10.52	72	27	33.9	-	-	-
11.72	74	-	31.5	-	-	-
13.56	77	30	29.5	-	-	-
15.08	81	30	26.7	-	-	-

2.2.1. Estimation of site-specific cross-correlation

In the literature, the term "transformation model" is adopted (Phoon and Kulhawy 1999) to denote the cross-correlation. Kulhawy and Mayne (1990) compiled many transformation models (or cross-correlations) between different soil parameters. However, the cross-correlations compiled in Kulhawy and Mayne (1990) are "generic" ones. Generic cross-correlation has a different meaning from site-specific cross-correlation, as discussed below. Consider an example where there are many sites whose soil parameters (X_1, X_2) exhibit zero site-specific cross-correlation (each local X_1-X_2 correlation forms a circle, as shown in Figure 4). Suppose that the circles have different centers due to site-uniqueness, and suppose that the centers lie on a line with a positive gradient. As seen in Figure 4, the generic X_1-X_2 data exhibit positive generic cross-correlation, although the X_1-X_2 data of each site exhibit a zero site-specific cross-correlation. Therefore, it is inappropriate to adopt generic cross-correlation in Kulhawy and Mayne (1990) as a replacement of site-specific cross-correlation.

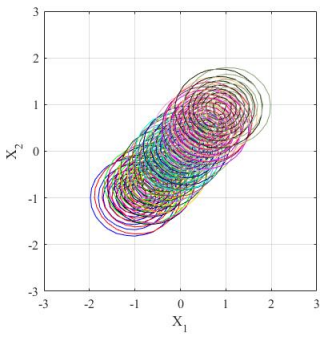


Figure 4. Generic vs. site-specific cross-correlations.

In principle, the estimation of site-specific cross-correlation requires site-specific data, not generic data. If the site-specific data are "ideal" (e.g., abundant,

complete, and independent), standard statistical methods can be readily adopted to estimate the site-specific cross-correlation (e.g., the Matlab command "cov" that computes the sample covariance matrix based on a set of multivariate data can be adopted). For the data in Table 1, abundance, completeness, and independence mean that there are numerous investigated depths (abundance) that are far apart so that there is no spatial correlation (independence) and that at each depth all soil parameters are measured (completeness, i.e., no empty entries in Table 1). However, realistic site investigation data are never ideal: they are MUSIC-3X (sparse, incomplete, and spatially correlated).

To our best knowledge, Ching and Phoon (2020a) is the first study that addresses the estimation of site-specific cross-correlations based on MUSIC-3X data. The spatial variation is limited to 1D (the depth direction) in Ching and Phoon (2020a), but this limitation is relaxed by Ching et al. (2022) by considering spatial variation in 3D. To address the "M" aspect (multivariate), these two studies are fully compatible to multivariate data because they adopt multivariate models. To address the 3X aspect, the spatial correlation in the site-specific data is modelled by a stationary random field model whose site-specific auto-correlation is estimated beforehand (the estimation of site-specific auto-correlation is discussed in Section 2.2.2). To address the "I" aspect (incompleteness), the Gibbs sampler (GS) algorithm (Geman and Geman 1984), a special instance of the Markov chain Monte Carlo (MCMC) methods in Bayesian analysis (Gilk et al. 1986), is adopted to deal the incomplete data: the GS can simulate the missing soil parameters while estimating the cross-correlation matrix at the same time. The "U" aspect (uncertainty) is also addressed because the two studies adopted the Bayesian analysis, which quantifies uncertainties with a probabilistic manner. Moreover, these two studies can further simulate conditional random fields (CRFs) for all soil parameters at unexplored locations. To make the Bayesian derivations and 3D computation/simulation tractable, two crucial assumptions are made in the two studies:

- (Assumption #1) Ching and Phoon (2020a) assumed that the cross-correlation and auto-correlation are separable, meaning that all soil parameters share the same auto-correlation parameters (such as the scale of fluctuation).
- (Assumption #2) Ching et al (2022) further assumed that the horizontal and vertical auto-correlations are separable, meaning that the 3D auto-correlation can be expressed as the product between the horizontal and vertical auto-correlations.

The above method of estimating site-specific cross-correlation and simulating CRFs was named as the "MUSIC-3X method" by Ching et al. (2022). In Ching et al. (2022), the Baytown site was analyzed by the MUSIC-3X method. The CRF simulation results for the B-3 borehole location of the Baytown site are shown in Figure 5. Note that the CRF can be simulated at other unexplored locations, but Figure 5 deliberately shows the CRF simulation results at the B-3 location to showcase the significance of CRF. The dark solid and dashed lines in Figure 5 show the median and 95% confidence interval (CI) profiles of the CRF, respectively, whereas the yellow dots show the observed data at B-3. It is reassuring that all simulated CRFs pass through the observed data.

Now consider a scenario where site-specific data are sparse by leaving B-3 data out of the analysis. By leaving B-3 data out, the number of (σ'_p, s_u) data points reduces from 7 to 4 (see the right-hand-side two plots in Figure 2). The CRF simulation results for this scenario of sparse data are shown in Figure 5 as the magenta lines. It is clear that now the 95% CIs for (σ'_p, s_u) are wide, suggesting that the uncertainty is high when the site-specific data are sparse. Note that in practice, it is common to have 4 or less (σ'_p, s_u) data points for a construction project of small to medium size. The magenta lines illustrate the severity of the challenge in the "S" aspect (data sparsity). The challenge for data sparsity will be addressed in Section 3 (site-recognition challenge).

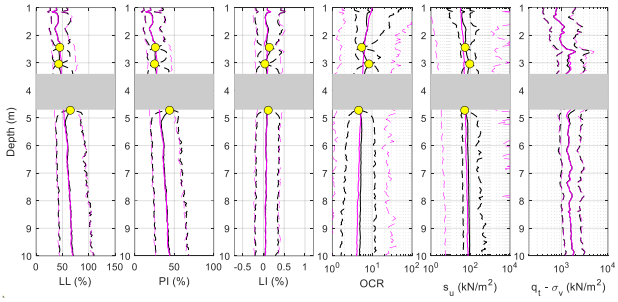


Figure 5. CRF simulation results for the B-3 borehole location of the Baytown site (magenta lines are the CRF results by leaving B-3 data out; the grey-out depth from 3.4 to 4.7 m is for the silty-sand layer) (source: Ching et al. 2022).

2.2.2. Estimation of site-specific auto-correlation

In the literature, the spatial variation of a soil parameter profile is usually modelled by two components: trend and spatial variability. The spatial variability is defined as the residual of the soil parameter profile with respect to its trend. The stationary random field model (Vanmarcke 1977) is the common way of modelling the spatial variability. A stationary random field is characterized by its auto-correlation. Classical auto-correlation models [such as the single-exponential (SExp) and squared-exponential (QExp) models] are governed by a single auto-correlation parameter called the scale of fluctuation (SOF). The SOF can be regarded as the correlation distance within which the detrended soil parameters of two locations are noticeably correlated. Besides the SOF, a non-classical auto-correlation model called the Whittle-Matérn (WM) auto-correlation model (Guttorp and Gneiting 2006; Liu et al. 2017; Ching and Phoon 2018) considers the smoothness as another important auto-correlation parameter. The smoothness parameter (ν) characterizes the degree of differentiability of the sample path of a random field. Ching and Phoon (2018) showed that ν has a significant impact on the failure probability of a geotechnical structure constructed in a spatially variable soil modelled by the stationary random field. The WM model is considered as a more general auto-correlation model because many classical models are special instances of the WM model (e.g., WM reduces to SExp and QExp, respectively, when ν = 0.5 and ∞).

In order to estimate site-specific SOF and smoothness, it is required to have a sufficiently small sampling interval and also sufficiently long total data length. For instance, if the actual SOF is 0.2 m, Ching and Phoon (2016) showed that the sampling interval has to be less than 0.2m/4 = 0.05m and the total data length has to be greater than 0.2m×5 = 1m in order to consistently identify the SOF. Among common site investigation tests, probably only CPT has sufficiently small vertical sampling interval. The vertical sampling interval for other tests such as borehole and vane shear tests may be too large to identify the vertical SOF and smoothness. For non-CPT soil parameters (e.g., Atterberg limits, water content, undrained shear strength, friction angle, modulus, etc.), a common assumption is that their auto-correlation parameters are the same as those of the CPT parameters. Note that this assumption is identical to Assumption #1 in Section 2.2.1. This assumption is based on the argument that the spatial correlation of a soil is governed largely by the spatial variability in its source materials, weathering patterns, stress, and formation history, etc. so that one would expect that all the soil parameters will vary similarly between the two points (Fenton and Griffiths 2003; Fenton et al. 2005). There is no evidence in the literature thus far to support or reject this assumption. Many studies in the literature have adopted Assumption #1 to model the auto-correlation parameters of non-CPT parameters. With Assumption #1, the estimation of site-specific auto-correlation can be conducted by analyzing site-specific CPT data.

The method of moments (MM) (e.g., Uzielli et al. 2005; Firouzianbandpey et al. 2014; Lloret-Cabot et al. 2014) is probably the most popular method of estimating the site-specific vertical auto-correlation parameters based on detrended CPT data, although the maximum likelihood (ML) method (DeGroot and Baecher 1993; Liu et al. 2016; Xiao et al. 2018) is more rigorous. Ching et al. (2019) investigated various estimation methods for vertical auto-correlation parameters, and they obtained the following conclusions (if the sampling interval is sufficiently small and if the total data length is sufficiently long):

- Regardless of the estimation method (MM or ML), classical auto-correlation models such as SExp and QExp cannot identify vertical smoothness because their smoothness parameter is fixed (ν = 0.5 for SExp and = ∞ for QExp).
- The MM method is effective in identifying vertical SOF, but it is ineffective in identifying vertical smoothness (even if the WM model is adopted).
- The ML method with the WM model can effectively identify vertical SOF and smoothness.

The above conclusions are for the estimation of site-specific vertical auto-correlation parameters. For the estimation of site-specific horizontal ones, our recent investigations (not published yet) showed that site-specific horizontal SOF and smoothness can be identified only if there are several pairs of closely spacing CPTs (the horizontal distance between the CPTs need to be at least less than the actual horizontal SOF). In practice, this condition (several pairs of closely spacing CPTs) may be hard to achieve because CPTs spread far apart to maximize the explored region. In our opinion, a reliable and practical method of estimating site-specific horizontal auto-correlation parameters is still lacking.

For the scenario where site-specific horizontal auto-correlation parameters cannot be estimated, the concept of "worst-case" values may be useful. For instance, the worst-case horizontal SOF for a footing can be defined as the conservative horizontal SOF value that maximizes its bearing capacity failure probability. Such worse-case auto-correlation parameters can be found by numerical simulation. In the literature, the worst-case SOFs for various design problems have been investigated using random finite element analyses (e.g., Fenton and Griffiths 2003; Jaksa et al. 2005; Fenton et al. 2005; Breysse et al. 2005; Griffiths et al. 2006; Soubra et al. 2008; Vessia et al. 2009; Ahmed and Soubra 2014; Ching et al. 2017). A state-of-the-art review of works related to worst-case SOFs has been conducted by Chapter 7 of ISSMGE-TC304 (2021).

As mentioned earlier, the spatial variation of a soil parameter profile consists of trend and spatial variability. Detrending is a standard pre-processing procedure for spatial data (e.g., Fenton 1999; Jaksa et al. 1999; Uzielli et al. 2005). In the past, the site-specific trend is usually determined using regression based on the spatial profile, and the site-specific SOF and smoothness are then estimated based on the detrended profile. A polynomial trend with a prescribed order (e.g., a linear or quadratic trend) is frequently adopted in the literature. Nonetheless, recent investigations showed that the modelling of the site-specific trend has profound impacts. Ching et al. (2020) showed that the assumption in the trend order (constant, linear, quadratic, etc.) has significant impacts. They analyzed the CPT data at a test site in Hollywood, South Carolina, USA (Stuedlein et al. 2016). This test site consists of 25 CPTs of 5 clusters, as shown in Figure 6. It is assumed that the trend follows a (k-1)-th order polynomial (e.g., k = 1 means that a constant trend is adopted). Figure 7 shows the site-specific auto-correlation parameters estimated based on the detrended CPT data for different assumed trend orders (k = 1, 2, …, 5). Figure 8 shows the median and 95% CI of the CRF simulation results at the central CPT location of Cluster #1 (this central CPT data are left out of the analysis and serve as validation data; see the red lines in Figure 8). It is evident that the assumed trend order has significant impacts on the estimated site-specific auto-correlation parameters and CRF simulation results.

Due to its significant impact, the modelling of the site-specific trend deserves further attention. Two recent developments have addressed the modelling of the trend with acceptable computations:

- Ching and Phoon (2017) and Ching et al. (2020) adopted the Sparse Bayesian Learning (SBL) method (Tipping 2001) to model the trend. Unlike the traditional way of modelling the trend as a polynomial with a prescribed order, the SBL method selects an optimal set of basis functions (BFs) such that the trend can be represented as the weighted sum of the BFs. The BF selection is adaptive: a simple trend requires very few BFs, and a complicated trend requires more BFs. For real CPT data, the optimal set of BFs is usually sparse: a small set of BFs can

usually well represent the trend. For the Hollywood site, the grey dots in Figure 7 show the estimation results (posterior samples) of the site-specific SOF and smoothness for the SBL method.

- Yoshida et al. (2021) adopted the Gaussian Process Regression (GPR) method (Rasmussen and Williams 2006) to model the trend. For the GPR method, the trend is modelled as a zero-mean stationary random field. This random field for trend is independent of the random field for spatial variability. The zero-mean random field for trend is the prior model, and this prior model is updated into the posterior model by the site-specific CPT data. The posterior random field for the trend is no longer zero-mean; instead, its mean follows the general trend of the CPT data.

More recently, Ching et al. (2023) compared these two methods and found that the SBL method usually prevails for 1D real cases (data from a single CPT) but the GPR method usually prevails for 2D or 3D real cases (data from multiple CPTs).

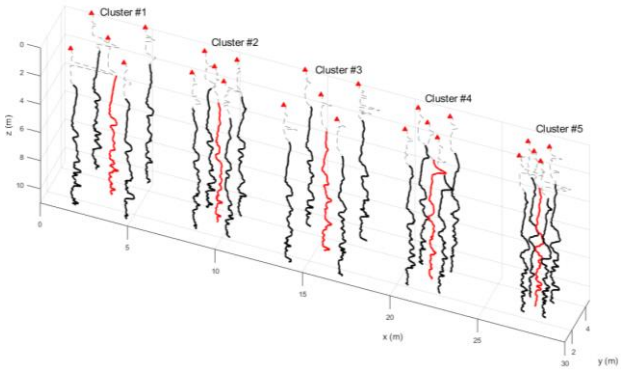


Figure 6. CPT data (con tip resistance) in the 3D underground space of the Hollywood test site (source: Stuedlein et al. 2016).

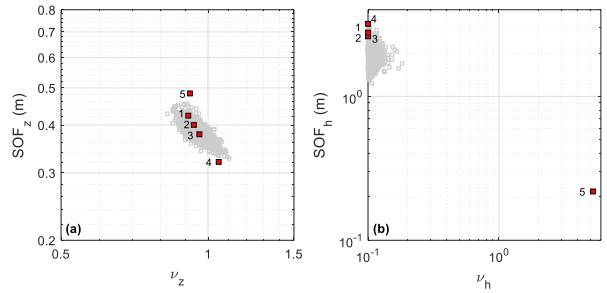


Figure 7. Estimated site-specific SOF and smoothness for different order of trend (k = annotated number): (a) vertical SOF and smoothness; (b) horizontal SOF and smoothness (from Ching et al. 2020). The grey dots are the SBL results.

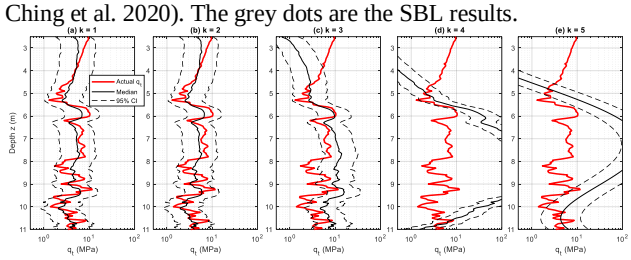


Figure 8. CRF simulation results at the central CPT of Cluster #1 (from Ching et al. 2020).

2.2.3. Bayesian compressive sensing

Besides the aforementioned works, there are works conducted by other scholars that can address the estimation of the site-specific cross-correlation and spatial correlation. In particular, the series of works related to Bayesian compressive sensing (BCS) are herein reviewed. Compressive sensing (Donoho 2006; Candès and Plan 2010; Davenport 2013) is an efficient method of reconstructing a signal if the signal can be represented by sparse BFs, and Bayesian compressive sensing (BCS) (Ji et al. 2008) is its Bayesian version. Zhao and Wang (2018), Xu et al. (2021), Guan & Wang (2021), and Li et al. (2023) implemented BCS to model the site-specific cross-correlation and spatial correlation simultaneously: the site-specific cross-correlation is modelled by a cross-correlation matrix, whereas the spatial correlation is modelled by a sparse set of BFs and their weights, not by the stationary random field model. Given the site-specific data, both site-specific cross-correlation and spatial correlation can be estimated, then cross-correlated CRFs can be subsequently simulated. These BCS studies address the "M" (multivariate), "U" (uncertain), "S" (sparse), and "3X" (spatial correlation) aspects. Although these BCS methods seem to have the potential to handle incomplete data, the "I" aspect is not yet addressed in these BCS works.

With BCS, the spatial variation of soil parameters is directly represented by BFs, hence there is no need to decompose the spatial variation into trend and spatial variability. The spatial variation is completely governed by the chosen BFs. There are cons and pros. The pros are that there is no need to detrend the profile and no need to estimate the auto-correlation parameters. The cons are that if the characteristics of the chosen BFs do not fit well with the actual spatial variation, important features in the actual spatial variation may not be captured. For instance, a spatial variation with non-smooth sample path cannot be represented by smooth BFs. It is fair to say that in BCS, the BFs are responsible for modelling both the trend and auto-correlation structure (such as SOF and smoothness). For BCS to effectively model all these characteristics, the effect of the BF type on the trend and auto-correlation structure should be investigated. Whether or not realistic geotechnical data with a wide range of trend and auto-correlation structure can be represented by a small number of BFs is an open research question that should be pursued further given the obvious attractiveness of BCS.

3. Site-recognition challenge

As discussed earlier, site-specific data are sparse. The data sparsity makes the estimation of site-specific cross-correlation and auto-correlation challenging. This issue is worsened by site-uniqueness, which means that the data from other sites cannot be directly adopted without exercising judgment in current practice to assist the estimation of cross-correlation and auto-correlation for the target site. The site-recognition challenge is to address "site-uniqueness", directly or indirectly, so that databases can be combined with sparse site-specific data in a manner sensitive to site differences to assist the estimation of site-specific cross-correlation and auto-

correlation. It departs from current practice in the application of data-driven methods rather than judgment that is limited to data familiar to engineers only. There are methods in the literature that address the site-recognition challenge:

- Similarity-based methods: find generic sites or records from a database similar to the target site and augment the site-specific data with these similar data.
- Transfer learning: transfer the knowledge learned from sites in a database to the target site.

3.1. Similarity-based methods

3.1.1. Similarity in cross-correlation

The definition of "similarity" can be either in cross-correlation or in auto-correlation. For the similarity in cross-correlation, Ching and Phoon (2020b) proposed a method that can extract "records" from a database that are similar to the target site in terms of the cross-correlation behaviors. Here, a record refers to a row of data in Table 1. More specifically, Ching and Phoon (2020b) proposed an index that measures the similarity between a record in a soil database and the target site, i.e., it is a "record-to-site" similarity index. Figure 9 shows the records from an Onsøy site (Lacasse and Lunne 1982), Norway (red dots) and the records from a database named CLAY/10/7490 (Ching and Phoon 2014). The database records are shown as grey and dark dots, where the dark dots are the records whose similarity indices (S) with respect to the Onsøy site are significant (S > 1). The database records with high similarity can be augmented to the Onsøy-site data for form a larger dataset to estimate the quasi-site-specific cross-correlation.

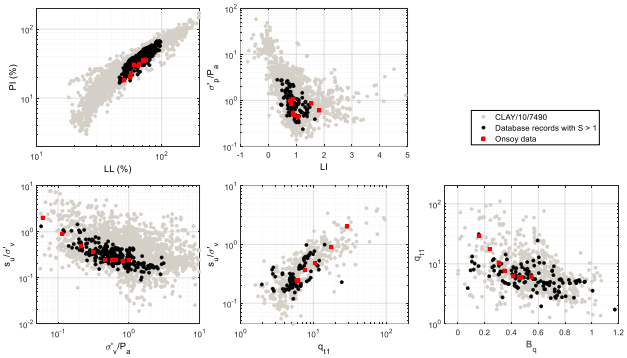


Figure 9. Cross-correlation plots for the Onsøy site and database records [$q_{t1} = (q_t-\sigma_v)/\sigma'_v$] (from Ching and Phoon 2020b).

In contrast, Sharma et al. (2022) proposed a method that can extract "sites" from a database that are similar to the target site in terms of the cross-correlation behaviors. The main difference here is that it is a "site-to-site" similarity index. The database sites highly similar to the target site are first identified. Then, their data are augmented to the site-specific data for form a larger dataset to estimate the quasi-site-specific cross-correlation. More recently, Cai et al. (2024) also proposed a site-to-site similarity index between a database site and the target site, and it is even faster than Sharma et al. (2022). All methods reviewed herein (Ching and Phoon 2020b; Sharma et al. 2022; Cai et al. 2024) can handle multivariate, uncertain, sparse, and incomplete data.

3.1.2. Similarity in auto-correlation

To the best knowledge of the authors, there were no previous studies to address how to quantify the similarity in auto-correlation between two sites until recently. Hu et al. (2024) adopted the Bayesian compressive sensing (BCS) as the basis of quantifying the similarity between the auto-correlation structures of two 2D cross-sections with sparse CPTs. The CPT data on each cross-section are expressed by a discrete-cosine transform (DCT) spectrum using BCS. The similarity between two sites is then quantified as the similarity between the two DCT spectra. In view of the importance of similarity in auto-correlation, more investigations into are needed in this research direction. To the knowledge of the authors, there are no methods that can address similarity in *both* cross-correlation and auto-correlation.

3.2. Transfer learning

Transfer learning is a technique where knowledge learned from a task is re-used to boost performance on a related task. In the context of the site-recognition challenge, the knowledge learned from site(s) in a database may be transferred to the target site to enhance the estimation of site-specific cross-correlation and auto-correlation.

3.2.1. Transfer-learning for cross-correlation

In the past, the transfer learning of cross-correlation has been conducted in a generic (non-site-specific) manner in the form of a "transformation model" (Phoon and Kulhawy 1999). Figure 10 shows the (generic) cross-correlation for the normalized CPT cone tip resistance $(q_t-\sigma_v)/\sigma'_v$ vs. normalized undrained shear strength s_u/σ'_v. The data points in the figure are from the CLAY/10/7490 database. The knowledge learned from the data points is the generic cross-correlation (transformation model), shown as the dark line in the figure. The dashed lines are the 95% CI, which quantifies the "transformation uncertainty" (Phoon and Kulhawy 1999).

In the case where site-specific cross-correlation is not available due to insufficient site-specific data, generic cross-correlation learned from a database (such as that in Figure 10) can be transferred to the target site. There are two drawbacks for generic cross-correlation:

- Generic cross-correlation usually has significant transformation uncertainty (i.e., the 95% CI is wide). This means that the prediction made by transferring the cross-correlation knowledge learned from a generic database to the target site is imprecise.
- As illustrated in Figure 4 earlier, generic cross-correlation has different meanings from site-specific cross-correlation. In fact, site-specific cross-correlation is usually less uncertain. This is illustrated by the data from a site in CLAY/10/7490 in Figure 11a. Figure 11a plots the $(q_t-\sigma_v)/\sigma'_v$ vs. s_u/σ'_v data with site labels (data from the same site have the same label). If we focus on the site labelled as yellow circles in Figure 11a, its site-specific cross-correlation is shown as the red line and 95%

CI shown as the red dashed lines. It is clear that the site-specific uncertainty for this site is much less than the generic uncertainty in Figure 10. This suggests that the prediction made by transferring the cross-correlation knowledge learned from "individual sites" in a database to the target site may be more precise.

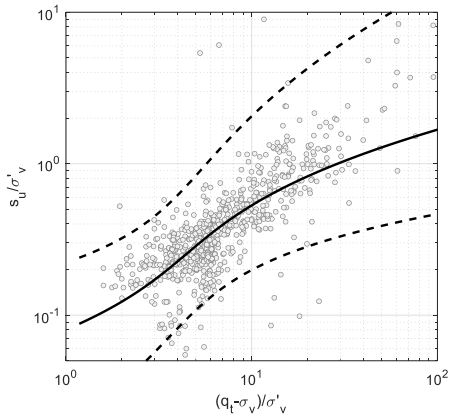


Figure 10. Generic cross-correlation between $(q_t-\sigma_v)/\sigma'_v$ vs. normalized undrained shear strength s_u/σ'_v (data from CLAY/10/7490 database).

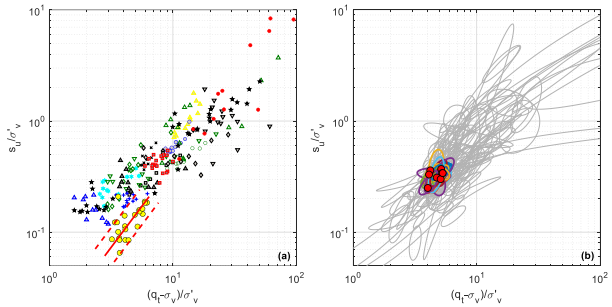


Figure 11. (a) $(q_t-\sigma_v)/\sigma'_v$ vs. s_u/σ'_v data with site labels (data from the same site have the same label); (b) learning and inference outcomes of the HBM.

Transferring the cross-correlation knowledge learned from sites in a database to the target site can be achieved via the hierarchical Bayesian model (HBM) (Gelman and Hill 2006; Zhang et al. 2016; Lu et al. 2018; Bozorgzadeh et al. 2019; Bozorgzadeh and Bathurst 2022). Ching et al. (2021) proposed the HBM shown in Figure 12 to model the cross-correlations of individual sites in a database. The site-uniqueness of the individual sites is explicitly modelled by the HBM. To be more specific, the (transformed) soil parameters for the j-th record at the i-th site in a database are denoted by the vector $\underline{X}_{ij}$, and $\underline{X}_{ij}$ is assumed to follow a multivariate normal distribution with site-specific cross-correlation mean vector = $\underline{\mu}_i$ and site-specific cross-correlation covariance matrix = $\mathbf{C}_i$, denoted by $N(\underline{\mu}_i,\mathbf{C}_i)$. The parameters $(\underline{\mu}_i,\mathbf{C}_i)$ quantify the site-specific cross-correlation of the i-th site. Because of site-uniqueness, $\underline{\mu}_i \neq \underline{\mu}_k$ and $\mathbf{C}_i \neq \mathbf{C}_k$ if $i \neq k$. Nonetheless, $\{\underline{\mu}_i: i = 1, \ldots, n_s\}$ (n_s is the number of database sites) are assumed to follow the same multivariate normal distribution $N(\underline{\mu}_0,\mathbf{C}_0)$, where $\underline{\mu}_0$ is the (hyper) mean vector and $\mathbf{C}_0$ is the (hyper) covariance matrix. Similarly, $\{\mathbf{C}_i: i = 1, \ldots, n_s\}$ are assumed to follow the same inverse-Wishart (IW) distribution (James 1964), denoted by $IW(\mathbf{\Sigma}_0,\nu_0)$, where $\mathbf{\Sigma}_0$ is the (hyper) scale matrix and ν_0 is the (hyper) degree of freedom. The hyper-parameters $(\underline{\mu}_0,\mathbf{C}_0,\mathbf{\Sigma}_0,\nu_0)$ govern the site-unique cross-correlations of the individual sites in the database. The target site has site-specific cross-correlation mean = $\underline{\mu}_s$ and site-specific cross-correlation covariance matrix = $\mathbf{C}_s$. It is assumed that $(\underline{\mu}_s,\mathbf{C}_s)$ are governed by the same hyper-parameters (see Figure 12).

The procedure of the HBM proposed by Ching et al. (2021) consists of two stages: learning stage and inference stage. In the learning stage, the hyper-parameters $(\underline{\mu}_0,\mathbf{C}_0,\mathbf{\Sigma}_0,\nu_0)$ are calibrated to learn the site-unique cross-correlations of the individual sites in the database. The calibrated hyper-parameters can produce a "prior model" for $(\underline{\mu}_s,\mathbf{C}_s)$ of the target site. The behaviors of this prior model can be illustrated by simulating the site-specific mean vector and covariance matrix $(\underline{\mu}_h,\mathbf{C}_h)$ of a "hypothetical site" based on the calibrated hyper-parameters: $\underline{\mu}_h \sim N(\underline{\mu}_0,\mathbf{C}_0)$ and $\mathbf{C}_h \sim IW(\mathbf{\Sigma}_0,\nu_0)$. Each $(\underline{\mu}_h,\mathbf{C}_h)$ simulation is represented as a (skewed) grey ellipse in Figure 11b. The grey ellipses in Figure 11b can be compared to the site data in Figure 11a.

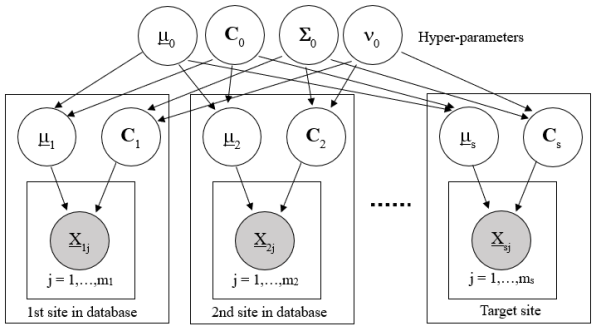


Figure 12. Model structure of the HBM (from Ching et al. 2021).

In the inference stage, the prior model for $(\underline{\mu}_s,\mathbf{C}_s)$ is updated into the posterior model through Bayesian analysis by further conditioning on the site-specific data of the target site, which are usually sparse (illustrated as the red circles in Figure 11b). It is clear from Figure 11b that most (grey) ellipses in Figure 11b are incompatible to the site-specific data, and those compatible to the site-specific data are shown in colors. The colored (compatible) ellipses illustrate the posterior model for $(\underline{\mu}_s,\mathbf{C}_s)$ after the Bayesian analysis. This posterior model is quasi-site-specific because not only the target-site data are used to construct the model but also the database is used to develop its prior. The uncertainty of this quasi-site-specific model can be visualized as the vertical size of the region occupied by the colored ellipses. It is remarkable that the uncertainty of the posterior (quasi-site-specific) model is much less than the uncertainty of the generic model in Figure 10.

The HBM proposed by Ching et al. (2021) can work with the MUSIC-3X method proposed by Ching et al. (2022) to form the HBM-MUSIC-3X method. The role of the HBM is to construct the prior model for $(\underline{\mu}_s,\mathbf{C}_s)$ of the target site (i.e., the learning stage). The prior model is updated by the site-specific data into the posterior (quasi-site-specific) model. The MUSIC-3X method has a certain role during this Bayesian updating because it

provides the cross-correlation and auto-correlation models for the site-specific data. Finally, the MUSIC-3X further conducts CRF simulations based on the posterior model. The CRF simulation results of the HBM-MUSIC-3X method at the borehole B-3 for the Baytown site are illustrated in Figure 13. The borehole B-3 data are left out of the analysis to emulate a scenario with sparse site-specific data. The dark lines are the CRF results by the HBM-MUSIC-3X method, whereas the magenta lines are the CRF results by the MUSIC-3X method. Recall that by leaving out the B-3 data, the number of (σ'_p, s_u) data points reduces from 7 to 4. With only 4 (σ'_p, s_u) data points and without transfer learning (i.e., the MUSIC-3X method), the site-specific cross-correlations related to σ'_p and s_u are highly uncertain, so the 95% CIs for the OCR and s_u CRF simulation results are wide (the dashed magenta lines in the OCR and s_u profile plots). Nonetheless, with transfer learning from the database (i.e., the HBM-MUSIC-3X method), the 95% CIs for the OCR and s_u CRF simulation results become much narrower. This shows that the HBM transfer learning can significantly reduce the uncertainty in cross-correlation due to sparse site-specific data.

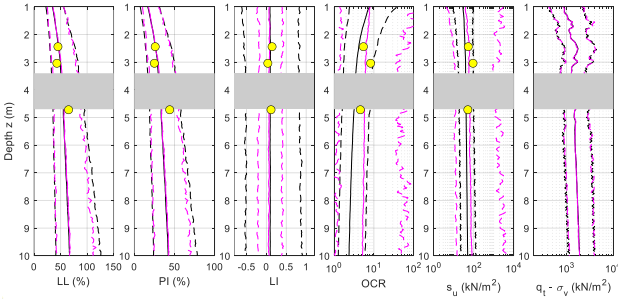


Figure 13. CRF simulation results at the B-3 borehole location of the Baytown site (B-3 data are left out of the analysis; the yellow dots show the observed data at B-3). Dark lines are for HBM-MUSIC-3X, and magenta lines are for MUSIC-3X (from Ching et al. 2022).

It is also possible to adopt the similarity-based method (Section 3.1) to further boost the performance of the HBM. Cai et al. (2024) found that if a small sub-database can be formed from a big database such that this small sub-database only contains sites with cross-correlations similar to the target site, the HBM trained by this smaller sub-database can be more effective than the HBM trained by the big database (more effective in the sense that the uncertainty of the quasi-site-specific prediction is further reduced). This sub-database is called quasi-regional database (Phoon and Ching 2022).

3.2.2. Transfer-learning for auto-correlation

It is also possible to transfer the knowledge learned from the spatial variations of other sites to the target site. As discussed earlier, spatial variation consists of trend and spatial variability, and spatial variability is usually characterized by auto-correlation parameters. As a result, there are two types of knowledge that can be transferred: transfer learning for trend and transfer learning for auto-correlation parameters. To the best knowledge of the authors, transfer learning for trend has not yet been pursued in the literature. Also, site-specific trends of soil parameters may be extremely complex and strongly dependent on local geology. It is an open research question whether it is possible to transfer the trend knowledge from other sites to the target site. Before this question is addressed, probably the only viable way to understand the spatial trend of the target site is through site-specific measurements (e.g., lots of CPTs).

Nonetheless, it is possible to transfer the auto-correlation knowledge from other sites to the target site because the auto-correlation parameters of various sites may vary in relatively narrow ranges. Cami et al. (2020) summarized the ranges of the site-specific horizontal SOF (SOF_h) and vertical SOF (SOF_z) reported in the literature for various soils, shown in Table 2. There are cases in the literature with simultaneous knowledge of site-specific SOF_h and SOF_z. Chapter 3 of ISSMGE-TC304 (2021) summarized the ranges for the site-specific SOFs of these cases as Figure 14.

Table 2. Ranges of site-specific SOF_h and SOF_z for various soil types (from Cami et al. 2020)

Soil type	SOF_h (m)			SOF_z (m)		
	# studies	Range	Mean	# studies	Range	Mean
Alluvial	9	1.07-49	14.2	13	0.07-1.1	0.36
Ankara clay	-	-	-	4	1-6.2	3.63
Chicago clay	-	-	-	2	0.8-1.3	0.91
Clay	9	0.1-164	31.9	16	0.05-3.6	1.29
Clay, sand, silt mix	13	1.2-1000	201.5	28	0.06-21	1.58
Hangzhou clay	2	40.4-45.4	42.9	4	0.5-0.8	0.63
Marine clay	8	8.4-66	30.9	9	0.11-6.1	1.55
Marine sand	1	15	15	5	0.07-7.2	1.43
Offshore soil	1	24.6-66.5	45.6	2	0.5-1.6	1.04
OC clay	1	0.14	0.14	2	0.06-0.3	0.15
Sand	9	1.7-80	24.5	14	0.1-4	1.17
Sensitive clay	-	-	-	2	1.1-2.0	1.55
Silt	3	12.7-45.5	33.2	5	0.1-7.2	2.08
Silty clay	7	9.65-45.4	29.8	14	0.1-6.5	1.40
Soft clay	3	22.2-80	47.6	8	0.14-6.2	1.70
Undrained engineered soil	-	-	-	22	0.3-2.7	1.42
Water content	9	2.8-22.2	12.9	8	0.05-6.2	1.70

More recently, Ching et al. (2023) conducted a comprehensive analysis of a large CPT database. From the database, they extracted hundreds of homogeneous soil units from 42 sites to identify SOF_z and vertical smoothness (ν_z) of the detrended CPT data based on the WM model. Table 3 shows the statistics of the site-specific SOF_z for soil units of various soil behavior types (SBTs) (Robertson 2016). It is evident that SOF_z varies in a relatively narrow range. The site-specific ν_z also varies in a relatively narrow range of 0.1-2.0 for all SBTs. Ching et al. (2023) did not investigate the horizontal auto-correlation parameters.

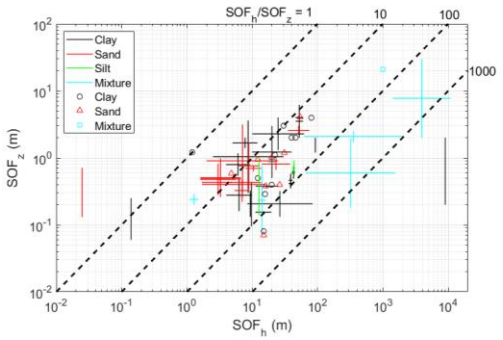


Figure 14. Ranges of SOF_h and SOF_z for various soil conditions (from ISSMGE-TC304 2021).

Table 3. Statistics for site-specific vertical auto-correlation parameters identified from CPT data (from Ching et al. 2023)

SBT	# sites	Site-specific SOF_z (m)		Site-specific ν_z
		Range	Mean	Range
2 (organic)	1	0.30	0.30	0.1-2.0
3 (clay)	7	0.10-0.46	0.26	
4 (silt mixtures)	8	0.13-0.42	0.27	
5 (sand mixtures)	13	0.084-0.58	0.32	
6 (sand)	13	0.31-0.89	0.49	
2, 3, 4 (clay-like)	16	0.10-0.46	0.27	
5, 6 (sand-like)	26	0.084-0.89	0.41	
All	42	0.084-0.89	0.35	

It is noteworthy that the SOF_z ranges in Table 3 (reported in Ching et al. 2023) are significantly narrower than those in Table 2 (reported in Cami et al. 2020) probably due to the following reasons:

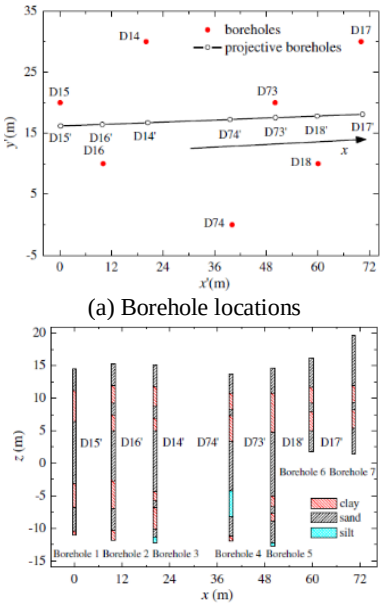
- Ching et al. (2023) analyzed their CPT data using a uniform framework (detrend with a quadratic trend; adopt the WM model; adopt the maximum likelihood estimation method). In contrast, Cami et al. (2020) simply compiled the SOF_z results from the literature. The trend functions, auto-correlation models, and estimation methods adopted in the literature are not uniform. This non-uniformity may result in the wider ranges in Table 2.
- Ching et al. (2023) focused on CPT data, which usually have sufficiently small sampling intervals to consistently identify SOF_z and ν_z. In contrast, the cases compiled by Cami et al. (2020) are not entirely from CPT data: some cases are borehole and vane shear data that have relatively large sampling intervals. For a large sampling interval, not only SOF_z cannot be consistently identified but some fluctuations in the trend may also be incorrectly treated as residuals. If this happens, the identified SOF_z may tend to be large because trend has a large scale of fluctuation.

For a target site with sufficient CPTs, it is recommended that its auto-correlation parameters can be estimated by analyzing the site-specific CPT data (transfer learning of auto-correlation is not necessary). For a target site without sufficient CPTs, it is recommended that the transfer learning of auto-correlation can be conducted: SOF_z and ν_z of the target site may be chosen based on the ranges in Table 3, whereas the ratio SOF_h/SOF_z can be chosen based on the ranges in Figure 14. The most updated reference for soil statistics is Phoon et al. (2024a).

4. Stratification challenge

The purpose of stratification is to delineate soil layers based on limited site-specific data. In the literature, the soil-layer delineation has been addressed by two types of methods:

- Domain-based methods: For these methods, the soil types at unexplored locations (domain) are simulated based on limited site-specific data. The coupled Markov chain (CMC) methods (e.g., Qi et al. 2016; Li et al. 2019; Varkey et al. 2023) and Markov random field (MRF) methods (e.g., Li et al. 2016; Zhao et al. 2021; Wei and Wang 2022) are two examples. Figure 15 shows the locations and soil-type data of the boreholes at a site in Perth city, Australia. There are only 6 boreholes within an area of 40 m × 70 m, and the task is to delineate the soil layers at unexplored locations. This site was analyzed by Qi et al. (2016) using the CMC method (Elfeki and Dekking 2001). Based on the observed soil-type data at the boreholes, the CMC method can simulate the soil types at unexplored locations using the Markov chain theory. Figure 16 shows one realization of the simulated soil types. More recently, machine learning methods are also adopted to learn the patterns in a "training image" of geological formation and subsequently to simulate soil types at unexplored locations based on the learned image patterns. This can be achieved by multiple-point geostatistics (MPG) method (e.g., Caers and Zhang 2004; Hu and Chugunova 2008; Shi and Wang 2021a) or by other methods such as convolutional neural network (CNN) (e.g., Shi and Wang 2021b). Figure 17a shows a training image from a training site, and Figure 17b shows the borehole data at a target site. It is assumed that the geological formation patterns of the training and target sites are similar. The MPG or CNN method can learn the patterns in the training image (e.g., Figure 17a) and then simulate soil types at unexplored locations (e.g., Figure 17b). The simulation result (the most probable stratum) based on the CNN method (Shi and Wang 2021b) is shown in Figure 17c.



(a) Borehole locations

(b) Soil-type data at boreholes

Figure 15. Soil-type data of the boreholes at a site in Perth city, Australia (from Qi et al. 2016).

- Boundary methods: These methods assume that the depth of the boundary between two soil layers is a continuous function (e.g., Figure 18a), and the purpose is to simulate the depths of the boundary at unexplored locations. Boundary methods are ideal for a boundary whose depth is a continuous function. For instance, Zhang and Dasaka (2010) modelled the depth of the soil-bedrock boundary as a 2D random field. Boundary methods can also handle problems

with multiple boundaries (e.g., Cao and Wang 2013; Xiao et al. 2017), e.g., there are two boundaries in Figure 18a. However, boundary methods can only deal with regular continuous soil-layer boundaries such as those in Figure 18a. It is not clear how to deal with diminishing boundaries and lenses in Figure 18b. Possibly due to this reason, boundary methods are not as popular as domain methods. Most recent advancements for stratification are for domain methods.

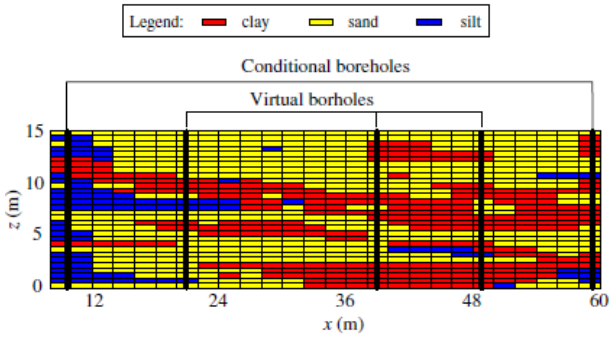
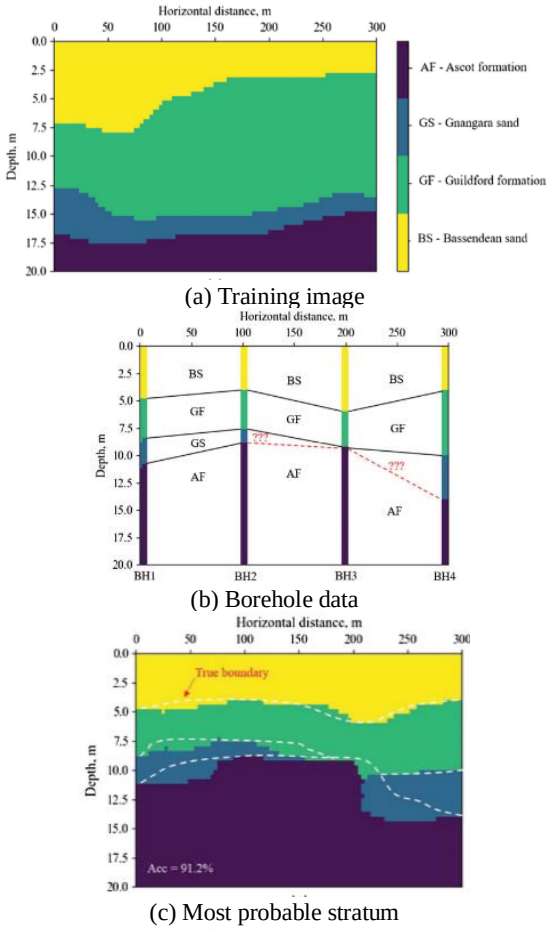


Figure 16. Realization of soil types at unexplored locations using the CMC method (from Qi et al. 2016).



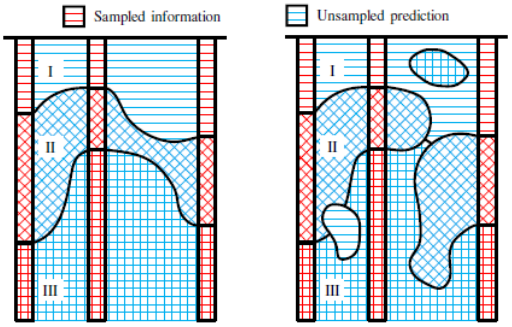
(a) Training image

(b) Borehole data

(c) Most probable stratum

Figure 17. Illustrations for the CNN method using the training image (source: Shi and Wang 2021b).

Most soil-layer delineation methods (e.g., CMC, MRF, and training-image methods) can only adopt soil-type data at boreholes (sand, silt, clay, etc.) as input. However, a routine geotechnical site investigation program usually consists of multiple information (the "M" aspect in MUSIC-3X). These soil-layer delineation methods cannot consider other types of site-specific data such as CPT data, Atterberg limits, water content, SPT N, VST (vane shear test), PMT (pressuremeter test), DMT (dilatometer test), and SASW (surface wave measurements) as inputs. The multivariate data obtained from these tests may be correlated to the soil types (e.g., SPT N values for clays are typically less than those for sands). It is desirable to develop new soil-layer delineation methods that can combine all available multivariate site investigation data from various types of tests to conduct soil-type simulation.



(a) Regular boundaries (b) Diminishing boundary & lenses

Figure 18. Illustration of soil-layer boundaries (from Xiao et al. 2017)

5. Concluding remarks

This paper reviews some recent advancements that address the challenges faced by the area of data-driven site characterization (DDSC), including the ugly-data challenge, site-recognition challenge, and stratification challenge. These challenges have been addressed to a certain amount of success, but there are still unresolved issues yet to be addressed. Some unresolved issues haven been discussed above, but there are some important issues that are not yet discussed:

- The need for databases: DDSC requires learning from real data, so it is necessary to compile large databases for cross-correlation, auto-correlation, and stratification. ISSMGE TC304 has initiated the effort of compiling databases (named 304dB, http://140.112.12.21/issmge/tc304.htm?=6) for cross-correlations and auto-correlations. For auto-correlation databases, besides 304dB providing some CPT data for some sites, the New Zealand Geotechnical Database (https://www.nzgd.org.nz/) contains a large number of CPTs conducted in New Zealand. For stratification, there are not a lot of databases around. In the opinion of the authors, the current compilation of databases is far from enough. For cross-correlation databases, there is a need to cover more common soil/rock parameters. For auto-correlation databases, there is a need to cover more regions. For stratification databases, there is a need to initiate the compilation.
- Computational cost: Realistic DDSC problems are 3D. For 3D problems, the computational cost can become a major challenge. Assumptions (such as Assumptions #1 and #2 in Section 2.2.1), advanced algorithms, and advanced computing methods may be needed to accelerate the computation. Currently,

these assumptions are "plausibly assumed" without supporting evidences. There is a need to verify these assumptions.

- Software: All research efforts taken in DDSC may not propagate to geotechnical engineering practice if they are not converted into useful and reliable software. There is a need to initiate/accelerate this conversion.
- In machine learning, benchmarking is used to compare tools and identify the best-performing ML solutions in the industry. This competition is expected to expedite development of real-world solutions. Phoon et al. (2022b) established the first set of benchmark examples for Project DeepGeo. The purpose of Project DeepGeo is to produce a 3D stratigraphic map of the subsurface volume below a full-scale project site and to estimate the governing engineering properties and soil type at each spatial point based on actual site investigation data and other relevant Big Indirect Data (BID) (Phoon and Ching 2021). However, the first set of benchmark examples only provides site-specific data for training; it does not provide data from "similar" sites in BID (Phoon et al. 2022b). It does not include the responses of geotechnical structures (monitoring data) as well. That is, it is not a benchmark example for machine learning guided observational method (MLOM) (Phoon and Shuku 2024).
- Other challenges: Phoon et al. (2024b) proposed a new taxonomy of site data under "4S" to expand the agenda for future research beyond site characterization. The "4S" are site generalizations, spatial features, sampling characteristics, and smart data. For first "S", the concept of a "site" is fundamental in geotechnical engineering, but the full extent of its complexity is still unfolding. Current research is already pointing to the concept of a "*data discovered site*", rather than a conventional definition based on a project site boundary. For the second "S", the most analysed spatial feature in geotechnical engineering is spatial variability. But spatial features can encompass geotechnical properties, geological features, ground improvement structures, and environmental processes (hydro, thermal, transport, etc.). They are basically 3D functions of space (or 4D when time is included) that influence the behaviour of a geotechnical structure constructed on or in a spatial domain. For the third "S", MUSIC-3X is an initial attempt to describe the attributes of geotechnical data found in a typical site investigation report. However, it does not cover other data attributes such as categorical data. The diverse sampling characteristics associated with different data sources pose challenges to data fusion methods. For the fourth "S", smart data is defined as actionable data at the point of collection. In contrast, conventional data are compiled and analysed in batches. Clearly, timely decisions cannot be made in the batch processing mode. The strategy to maximize the value of data in the presence of Internet of Things (IoT) is not likely the same as conventional monitoring instruments. The contours of a possible Value of Smart Information (VoSI) framework are not defined at this point, but a good start is to imagine how to extract more value from conventional monitoring instruments that can talk to each other by building intelligence into every stage of the data processing chain. VoSI is expected to be an integral part of machine learning guided observational method (MLOM) (Phoon and Shuku 2024).

Machine learning in geotechnics should be approached with an appropriate balance of three elements: (1) data centricity, (2) fit for (and transform) practice, and (3) geotechnical context. This agenda underpins a new interdisciplinary field termed "data-centric geotechnics" (Phoon and Ching 2021; Phoon et al. 2022c). An algorithm is not an end in itself. It is a means to actualize data-centric geotechnics, in which data-driven site characterization is one important application area. The explosive rise of digital technologies presents many opportunities to transform geotechnical practice at the Type 3 level (disruptive) (Phoon and Zhang 2023). More research in collaboration with the industry and government agencies is urgently needed.

References

Bozorgzadeh, N. and R.J. Bathurst. 2022. "Hierarchical Bayesian Approaches to Statistical Modelling of Geotechnical Data." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16(3): 452-469.

Bozorgzadeh, N., J.P. Harrison, and M.D. Escobar. 2019. "Hierarchical Bayesian Modelling of Geotechnical Data: Application to Rock Strength." Géotechnique 69(12): 1056-1070.

Caers, J. and T.F. Zhang. 2004. "Multiple-point Geostatistics: A Quantitative Vehicle for Integrating Geologic Analogs into Multiple Reservoir Models." In: Grammar, G.M., Harris, P.M., and Eberli, G.P. (eds) Integration of Outcrop and Modern Analogs in Reservoir Modeling, American Association of Petroleum Geologists 80: 383- 394.

Cai, Y.M., J. Ching, and K.K. Phoon. 2024. "Tailored Clustering Method to Identify Quasi-Regional Sites." Engineering Geology 333: 107490.

Cami, B., S. Javankhoshdel, K.K. Phoon, and J. Ching. 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.

Candès, E.J. and Y. Plan. 2010. "A Probabilistic and RIPless Theory of Compressed Sensing." IEEE Transactions on Information Theory 57, 7235-7254.

Cao, Z. and Y. Wang. 2013. "Bayesian Approach for Probabilistic Site Characterization Using Cone Penetration Tests." Journal of Geotechnical and Geoenvironmental Engineering 139(2): 267-276.

Ching, J. and K.K. Phoon. 2014. "Transformations and Correlations among Some Parameters of Clays – The Global Database." Canadian Geotechnical Journal 51(6): 663-685.

Ching, J. and K.K. Phoon. 2017. "Characterizing Uncertain Site-specific Trend Function by Sparse Bayesian Learning." Journal of Engineering Mechanics 143(7): 04017028.

Ching, J. and K.K. Phoon. 2018. "Impact of Auto-correlation Function Model on the Probability of Failure." Journal of Engineering Mechanics 145(1): 04018123.

Ching, J. and K.K. Phoon. 2020a. "Constructing a Site-specific Multivariate Probability Distribution Using Sparse, Incomplete, and Spatially Variable (MUSIC-X) Data." Journal of Engineering Mechanics 146(7): 04020061.

Ching, J. and K.K. Phoon. 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.

Ching, J., I. Yoshida, and K.K. Phoon. 2023. “Comparison of Trend Models for Geotechnical Spatial Variability: Sparse Bayesian Learning vs. Gaussian Process Regression.” Gondwana Research 123: 174-183.

Ching, J., K.K. Phoon, A.W. Stuedlein, and M.B. Jaksa. 2019. “Identification of Sample Path Smoothness in Soil Spatial Variability.” Structural Safety 81: 101870.

Ching, J., K.K. Phoon, and S.P. Sung. 2017. “Worst Case Scale of Fluctuation in Basal Heave Analysis Involving Spatially Variable Clays.” Structural Safety 68: 28-42.

Ching, J., K.K. Phoon, Z.Y. Yang, and A.W. Stuedlein. 2022. “Quasi-site-specific Multivariate Probability Distribution Model for Sparse, Incomplete, and Three-dimensional Spatially Varying Soil Data.” Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16(1): 53-76.

Ching, J., M. Uzielli, K.K. Phoon, and X.J. Xu. 2023. “Characterization of Autocovariance Parameters of Detrended Cone Tip Resistance from a Global CPT Database.” Journal of Geotechnical and Geoenvironmental Engineering, 149(10): 04023090.

Ching, J., S. Wu, and K.K. Phoon. 2021. “Constructing Quasi-site-specific Multivariate Probability Distribution Using Hierarchical Bayesian Model.” Journal of Engineering Mechanics 147(10): 04021069.

Ching, J., S.H. Wu, and K.K Phoon. 2016. “Statistical Characterization of Random Field Parameters Using Frequentist and Bayesian Approaches.” Canadian Geotechnical Journal 53(2): 285-298.

Ching, J., W.H. Huang, and K.K. Phoon. 2020. “3D Probabilistic Site Characterization by Sparse Bayesian Learning.” Journal of Engineering Mechanics 146(12): 04020134.

Davenport, M. 2013. “The Fundamentals of Compressive Sensing,” SigView, April 12.

DeGroot, D.J. and G.B. Baecher. 1993. “Estimating Autocovariances of In-situ Soil Properties.” Journal of Geotechnical Engineering 119(1): 147-166.

Donoho, D. 2006. “For Most Large Underdetermined Systems of Linear Equations, the Minimal 1-norm Solution Is Also the Sparsest Solution.” Communications on Pure and Applied Mathematics 59(6): 797–829.

Elfeki, A. and M. Dekking. 2001. “A Markov Chain Model for Subsurface Characterization: Theory and Applications.” Mathematical Geology 33(5): 569-89.

Fenton, G.A. 1999. “Random Field Modeling of CPT Data.” Journal of Geotechnical and Geoenvironmental Engineering 125(6): 486-498.

Fenton, G.A. and D.V. Griffiths. 2003. “Bearing-capacity Prediction of Spatially Random c-φ Soils.” Canadian Geotechnical Journal 40(1): 54-65.

Fenton, G.A., D.V. Griffiths, and M.B. Williams. 2005. “Reliability of Traditional Retaining Wall Design.” Géotechnique 55(1): 55-62.

Firouzianbandpey, S., D.V. Griffiths, L.B. Ibsen, and L.V. Anderson. 2014. “Spatial Correlation Length of Normalized Cone Data in Sand: Case Study in the North of Denmark.” Canadian Geotechnical Journal 51(8): 844-857.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Geman, S. and D. Geman. 1984. “Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images.” IEEE Transactions on Pattern Analysis and Machine Intelligence 6: 721-741.

Gilks, W.R., D.J. Spiegelhalter, and S. Richardson. 1996. Markov Chain Monte Carlo in Practice. Chapman and Hill, London.

Guan Z. and Y. Wang. 2021. “Non-parametric Construction of Site-specific Non-Gaussian Multivariate Joint Probability Distribution from Sparse Measurements.” Structural Safety 91: 102077.

Guttorp, P. and T. Gneiting. 2006. “Studies in the History of Probability and Statistics XLIX on the Matérn Correlation Family.” Biometrika 93(4): 989-995.

http://140.112.12.21/issmge/2021/SOA_Review_on_geotechnical_property_variablity_and_model_uncertainty.pdf

https://doi.org/10.1016/j.enggeo.2024.107445

Hu, L.Y. and T. Chugunova. 2008. “Multiple-point Geostatistics for Modeling Subsurface Heterogeneity: A Comprehensive Review.” Water Resources Research 44: W11413.

Hu, Y., Y. Wang, K.K. Phoon, and M. Beer. 2024. “Similarity Quantification of Soil Spatial Variability between Two Cross-sections Using Auto-correlation Functions.” Engineering Geology 331: 107445.

ISSMGE-TC304 (2021). State-of-the-art Review of Inherent Variability and Uncertainty in Geotechnical Properties and Models. International Society of Soil Mechanics and Geotechnical Engineering (ISSMGE) - Technical Committee TC304 ‘Engineering Practice of Risk Assessment and Management’, March 2nd., 2021.

Jaksa, M.B., J.S. Goldsworthy, G.A. Fenton, W.S. Kaggwa, D.V. Griffiths, Y.L. Kuo, and H.G. Poulos. 2005. “Towards Reliable and Effective Site Investigations.” Geotechnique 55(2): 109-121.

Jaksa, M.B., W.S. Kaggwa, and P.I. Brooker. 1999. “Experimental Evaluation of the Scale of Fluctuation of a Stiff Clay.” Proceedings of the 8th International Conference on Application of Statistics and Probability, A.A. Balkema, Rotterdam, 415-422.

Ji, S., Y. Xue, and L. Carin. 2008. “Bayesian Compressive Sensing.” IEEE Transactions on Signal Processing 56: 2346-2356.

Kulhawy, F.H. and P.W. Mayne. 1990. Manual on Estimating Soil Properties for Foundation Design. Report EL-6800, Electric Power Research Institute, Cornell University, Palo Alto.

Lacasse, S. and T. Lunne. 1982. “Penetration Tests in Two Norwegian Clays.” Proc. 2nd Eur. Symp. on Penetration Testing, Amsterdam, 661-670.

Li, J., Y.M. Cai, X.Y. Li, and L.M. 2019. “Simulating Realistic Geological Stratigraphy Using Direction-dependent Coupled Markov Chain Model.” Computers and Geotechnics 115: 103147.

Li, P., Y. Wang, and Z. Guan. 2023. “Non-parametric Generation of Multivariate Cross-correlated Random Fields Directly from Sparse Measurements Using Bayesian Compressive Sensing and Markov Chain Monte Carlo.” Stochastic Environmental Research and Risk Assessment 37: 4607-4628.

Li, Z., X.R. Wang, H. Wang, and R.Y. Liang. 2016. “Quantifying Stratigraphic Uncertainties by Stochastic Simulation Techniques Based on Markov Random Field.” Engineering Geology 201: 106-122.

Liu, W.F., Y.F. Leung, and M.K. Lo. 2016. “Integrated Framework for Characterization of Spatial Variability of Geological Profiles.” Canadian Geotechnical Journal 54(1): 47-58.

Liu, W.F., Y.F. Leung, and M.K. Lo. 2017. “Integrated Framework for Characterization of Spatial Variability of Geological Profiles.” Canadian Geotechnical Journal 54(1): 47-58.

Lloret-Cabot, M., G.A. Fenton, and M.A. Hicks. 2014. “On the Estimation of Scale of Fluctuation in Geostatistics.” Georisk:

Assessment and Management of Risk for Engineered Systems and Geohazards 8(2): 129-140.

Lu, S., J. Zhang, S. Zhou, and A. Xu. 2018. “Reliability Prediction of the Axial Ultimate Bearing Capacity of Piles: A Hierarchical Bayesian Method.” Advances in Mechanical Engineering 10(11): 1-11.

Phoon, K.K. and F.H. Kulhawy. 1999. “Evaluation of Geotechnical Property Variability.” Canadian Geotechnical Journal 36(4): 625-639.

Phoon, K.K. and J. Ching 2021. “Project DeepGeo – Data-driven 3D Subsurface Mapping.” Journal of GeoEngineering 16(2), 61-74.

Phoon, K.K. and J. Ching. 2022. “Additional Observations on the Site Recognition Challenge.” Journal of GeoEngineering 17(4): 231-247.

Phoon, K.K. and W.G. Zhang. 2023. “Future of Machine Learning in Geotechnics.” Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 17(1): 7-22.

Phoon, K.K. and T. Shuku. 2024. “Future of Machine Learning in Geotechnics (FOMLIG), 5–6 Dec 2023, Okayama, Japan.” Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 18(1): 288-303.

Phoon, K.K., J. Ching, and Y. Tao. 2024a. “Soil and Rock Parametric Uncertainties. Chapter 2 in Uncertainty, Modelling, and Decision Making in Geotechnics, CRC Press, Boca Raton, 39-116.

Phoon, K.K., J. Ching, and C. Tang. 2024b. “Role of Site Characterization Information in Data-centric Geotechnics.” Chapter 1 in Database Approach for Data-centric Geotechnics: Site Characterization.

Phoon, K.K., J. Ching, and T. Shuku. 2022a. “Challenges in Data-driven Site Characterization.” Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16(1): 114-126.

Phoon, K.K., T. Shuku, J. Ching, and I. Yoshida. 2022b. “Benchmark Examples for Data-driven Site Characterization.” Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 16(4): 599-621.

Phoon, K. K., J. Ching, and Z. Cao. 2022c. “Unpacking Data-centric Geotechnics.” Underground Space 7(6): 967–989.

Qi, X.H., D.Q. Li, K.K. Phoon, Z. Cao, and X.S. Tang. 2016. “Simulation of Geologic Uncertainty Using Coupled Markov Chain.” Engineering Geology 207: 129-140.

Rasmussen, C.E. and C.K.I. Williams. 2006. Gaussian Processes for Machine Learning. MIT Press, London.

Robertson, P.K. 2016. “Cone Penetration Test (CPT)-based Soil Behaviour Type (SBT) Classification System – An Update.” Canadian Geotechnical Journal 53: 1910-1927.

Sharma, A., J. Ching, and K.K. Phoon. 2022. “A Hierarchical Bayesian Similarity Measure for Geotechnical Site Retrieval.” Journal of Engineering Mechanics 148(10): 04022062.

Shi, C. and Y. Wang. 2021a. “Non-parametric and Data-driven Interpolation of Subsurface Soil Stratigraphy from Limited Data Using Multiple Point Statistics.” Canadian Geotechnical Journal 58(2): 261-280.

Shi, C. and Y. Wang. 2021b. “Development of Subsurface Geological Cross-section from Limited Site-specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost.” Journal of Geotechnical and Geoenvironmental Engineering 147(9): 04021082.

Soubra, A.H., D.S.Y.A. Massih, and M. Kalfa. 2008. “Bearing Capacity of Foundations Resting on a Spatially Random Soil.” Geotechnical Special Publication 178: 66-73.

Stuedlein, A.W., S.L. Kramer, P. Arduino, and R.D. Holtz. 2012. “Geotechnical Characterization and Random Field Modeling of Desiccated Clay.” Journal of Geotechnical and Geoenvironmental Engineering 138(11): 1301-1313.

Stuedlein, A.W., T.N. Gianella, and G.J. Canivan. 2016. “Densification of Granular Soils Using Conventional and Drained Timber Displacement Piles.” Journal of Geotechnical and Geoenvironmental Engineering 142(12): 04016075.

Tipping, M.E. 2001. “Sparse Bayesian Learning and the Relevance Vector Machine.” Journal of Machine Learning Research 1: 211-244.

Uzielli, M., G. Vannucchi, and K.K. Phoon. 2005. “Random Field Characterisation of Stress-normalised Cone Penetration Testing Parameters.” Geotechnique 55(1): 3-20.

Vanmarcke, E.H. 1977. “Probabilistic Modeling of Soil Profiles.” Journal of Geotechnical Engineering GT11: 1227-1246.

Varkey, D., A.P. van den Eijnden, and M.A. Hicks. 2023. “Predicting Subsurface Stratigraphy using an Improved Coupled Markov Chain Method.” Proceedings of the 14th International Conference on Applications of Statistics and Probability in Civil Engineering, Dublin, Ireland, July 9-13.

Vessia, G., C. Cherubini, J. Pieczyńska, and W. Puła. 2009. “Application of Random Finite Element Method to Bearing Capacity Design of Strip Footing.” Journal of Geoengineering 4(3): 103-112.

Wei, X.X. and H. Wang. 2022. “Stochastic Stratigraphic Modeling Using Bayesian Machine Learning.” Engineering Geology 307: 106789.

Xiao, T., D.Q. Li, Z.J. Cao, and L.M. Zhang. 2018. “CPT-based Probabilistic Characterization of Three-dimensional Spatial Variability Using MLE.” Journal of Geotechnical and Geoenvironmental Engineering 144(5): 04018023.

Xiao, T., L.M. Zhang, X.Y. Li, and D.Q. Li. 2017. “Probabilistic Stratification Modeling in Geotechnical Site Characterization.” ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 3(4): 04017019.

Xu, J., Y. Wang, and L. Zhang. 2021. “Interpolation of Extremely Sparse Geo-data by Data Fusion and Collaborative Bayesian Compressive Sampling.” Computers and Geotechnics 134: 104098.

Yoshida, I., Y. Tomizawa, and Y. Otake. 2021. “Estimation of Trend and Random Components of Conditional Random Field Using Gaussian Process Regression.” Computers and Geotechnics 136: 104179.

Zhang, J., C.H. Juang, J.R. Martin, and H.W. Huang. 2016. “Inter-region Variability of Robertson and Wride Method for Liquefaction Hazard Analysis.” Engineering Geology 203: 191-203.

Zhang, L.M. and S.M. Dasaka. 2010. “Uncertainties in Geologic Profiles versus Variability in Pile Founding Depth.” Journal of Geotechnical and Geoenvironmental Engineering 136(11): 1475-1488.

Zhao T. and Y. Wang. 2018. “Simulation of Cross-correlated Random Field Samples from Sparse Measurements Using Bayesian Compressive Sensing.” Mechanical Systems and Signal Processing 112: 384-400.

Zhao, C., W.P. Gong, T.Z. Li, C.H. Juang, H.M. Tang, and H. Wang. 2021. “Probabilistic Characterization of Subsurface Stratigraphic Configuration with Modified Random Field Approach.” Engineering Geology 288: 106138.