Constructing Site-Specific Multivariate Probability Distribution Model Using Bayesian Machine Learning

Jianye Ching, M.ASCE[1]; and Kok-Kwang Phoon, F.ASCE[2]

Abstract: This study proposes a novel data-driven Bayesian machine learning method for constructing site-specific multivariate probability distribution models in geotechnical engineering. There is a trade-off for constructing a site-specific model: a model developed from generic data may not be fully applicable to a local site, but a model purely developed from limited site-specific data may be very imprecise due to significant statistical uncertainty. The proposed method is based on the hybridization between site-specific and generic data in the way that it is governed by site-specific data when site-specific data are abundant and by generic data when site-specific data are sparse. This method broadly follows how an engineer currently estimates design soil parameters from limited site-specific information. The proposed method admits incomplete multivariate data, so it can handle missing data that are commonly encountered in site investigation. It is a Bayesian method, so uncertainties are rigorously quantified. Actual case studies are used to demonstrate the usefulness of the proposed method. Analysis results show that the proposed method can effectively capture the correlation behaviors in site-specific data and, moreover, can make meaningful predictions even when site-specific data are very sparse. **DOI: 10.1061/(ASCE)EM.1943-7889.0001537.** *© 2018 American Society of Civil Engineers.*

Author keywords: Geotechnical engineering; Site characterization; Multivariate probability distribution model; Transformation model; Uncertainty; Bayesian machine learning.

Introduction

It is common that geotechnical design is conducted under limited site investigation data. First, borings and soundings are widely spaced (e.g., one borehole per 300 m^2), so the spatial distribution of the soil property is unknown at unexplored locations. This source of uncertainty is called spatial variability in the literature (Vanmarcke 1977; Phoon and Kulhawy 1999a). Second, design soil parameters, such as the undrained shear strength of a clay or the friction angle of a sand, typically require relatively sophisticated and expensive sampling and laboratory testing techniques. Test indexes such as the standard penetration test (SPT) blow count or cone penetration test (CPT) values can be adopted to correlate to design soil parameters. However, the correlation equation (or transformation model) usually contains a significant amount of uncertainty. This source of uncertainty is called transformation uncertainty (Phoon and Kulhawy 1999b). Third, the volume of the tested soils is typically small compared with the volume affecting the limit state. It is impossible to determine the parameter affecting the limit state exactly. This source of uncertainty is called statistical uncertainty. Finally, there are measurement errors for all tests. This source of uncertainty is called measurement error. The current paper focuses on the second and third sources of uncertainty: transformation uncertainty associated with transformation models and the statistical uncertainty due to limited data.

Useful compilations of transformation models are available in the literature (e.g., Djoenaidi 1985; Kulhawy and Mayne 1990; Mayne et al. 2001). For instance, it is common to estimate the friction angle (ϕ') of sand based on its SPT blow count (N) through a transformation model derived from calibration data points obtained in the literature. Here, the N value is site specific, and the friction angle of interest is also site specific. However, the N-ϕ' transformation model is not site specific and is typically developed using an N-ϕ' calibration database compiled from a large number of sites. It is common to adopt such a generic transformation model to estimate site-specific design parameters because there are insufficient data at a single site to establish a fully relevant local model. In spite of their successful record, generic transformation models are applied with some degree of trepidation because error bounds are *not* routinely determined in the current practice. An experienced engineer can mitigate this limitation by applying engineer's judgment (reality check), but a probabilistic transformation model can produce a 95% confidence interval (CI) for the estimate explicitly to augment this judgment. To achieve this, a probabilistic transformation model, be it generic or local, will require proper treatment of transformation and statistical uncertainties. The occasional criticism that a generic model is not applicable skirts the practical point that one is compelled to use such a model in the absence of anything better. In the opinion of the authors, the better approach is to replace the current single estimate by a 95% confidence interval that allows an engineer to make an informed decision on applicability of a particular model. An absurdly large interval may alert the engineer that a particular model is almost useless, but this is not always the case. The value of a model should be quantified by the width of the confidence interval, rather than debated in the absence of a database.

When applied to a specific site, the transformation uncertainty of a generic model can be excessively large (e.g., Ching and Phoon 2014b; Ching et al. 2017a, 2018a; Ching 2018) because it is intended to accommodate a wide range of soil types and site conditions. However, if we narrow down a database to a single site, the remaining calibration data points can be too sparse to construct the transformation model with any acceptable degree of statistical significance. More specifically, one exchanges large transformation uncertainty associated with a generic transformation model with

[1] Professor, Dept. of Civil Engineering, National Taiwan Univ., Taipei 106, Taiwan (corresponding author). Email: jyching@gmail.com

[2] Professor, Dept. of Civil and Environmental Engineering, National Univ. of Singapore, Singapore 119077, Singapore.

Note. This manuscript was submitted on December 6, 2017; approved on June 13, 2018; published online on November 10, 2018. Discussion period open until April 10, 2019; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Engineering Mechanics***, © ASCE, ISSN 0733-9399.

large statistical uncertainty associated with a purely site-specific or local transformation model based on insufficient data. There is a trade-off here: a generic model may be less applicable, but a site-specific model may be very imprecise. A discussion on the relative merits of a generic versus site-specific transformation model is incomplete without accounting for the different sources of uncertainties and their respective weights explicitly.

With the preceding practical background in mind, the authors submit that a realistic method for constructing a transformation model for geotechnical data should fulfill the following considerations:

1. The resulting transformation model should be probabilistic, in the sense that not only the point estimate for the design soil parameter can be obtained but also its transformation and statistical uncertainties can be quantified (e.g., expressed as a confidence interval). This is essential for reliability-based design (RBD) and for risk analysis. The determination of the characteristic value in Eurocode 7 [EN 1997-1 (CEN 2004)] may also require the quantification of transformation and statistical uncertainties. More fundamentally, it is arguably a more realistic approach given that transformation and/or statistical uncertainties are usually significant. A single estimate, be it an average or a conservative value, would not provide the engineer with a sufficient sense of imprecision for a basic parametric study or even the relevance of a particular model to a specific site.
2. The resulting transformation model should allow multivariate inputs and can predict multivariate outputs. The advantage of allowing multiple inputs has been elaborated in Ching et al. (2010), Ching and Phoon (2012), and Ching et al. (2014). The key idea is that the design soil parameter is usually correlated to more than one test results, e.g., ϕ' correlated to relative density (D_r), SPT blow count (N), and median grain size (D_{50}) simultaneously. By combining multiple input information, the transformation uncertainty can be further reduced so that a more economic design outcome can be obtained. This has the potential of turning site investigation from a *cost* into an *investment* (Ching et al. 2014).
3. The method should be able to accommodate incomplete multivariate site-specific data. For instance, if a transformation model that predicts ϕ' based on (D_r, N, D_{50}) is to be constructed, in principle we need multivariate data with simultaneous knowledge of (ϕ', D_r, N, D_{50}) at the same depth and reasonably close borehole and test locations. However, it is very rare that such complete multivariate data points are available during a common site investigation program. It is common to measure incomplete multivariate data points at different depths and locations, for instance, some data points have (ϕ', D_r) information, some have (D_r, D_{50}) information, or some only have ϕ' information. If the data points are visualized as a spreadsheet table of size ($m \times 4$), where m is the number of measured depths, incomplete multivariate data means there are missing entries in the spreadsheet table.
4. When there are abundant data at a specific site, the method will mainly rely on the local data to construct the transformation model. When site-specific data are sparse and there is limited local experience (not an uncommon situation), the resulting transformation model should rely more on generic data. It is sensible to rely on data from other sites sharing comparable geology. This is in line with current geotechnical practice where a desk study is integral to site investigation.
5. It is also desirable that the applicability of the method is independent of the nature of the problem. It should be applicable to transformation models for clays, sands, or rocks: namely, it should be a framework that is purely driven by data points, i.e., a machine learning method. This is in alignment to the International Society for Soil Mechanics and Geotechnical Engineering's (ISSMGE's) latest initiative to explore machine learning methods in geotechnical engineering. More importantly, an in-depth physical understanding of the geomaterials at an actual site is rarely available [only available at very few research sites (Tan et al. 2003a, b, 2006a, b)].

In the literature, Yan et al. (2009), Wang and Cao (2013), Feng and Jimenez (2014), Cao and Wang (2014), Ng et al. (2015), Feng and Jimenez (2015), Wang and Aladejare (2015), and Ng et al. (2017) proposed Bayesian methods to construct a probabilistic site-specific model to predict a single output (design soil or rock design parameter). These previous works can address Consideration 1. Wang and Akeju (2016) and Wang and Aladejare (2016) proposed Bayesian methods to construct a probabilistic site-specific model to predict multivariate outputs. These previous works can address Considerations 1 and 2. Ching and Phoon (2014a), Liu et al. (2016), and Ching et al. (2017b, 2018b) adopted a Bayesian method (called the conventional Bayesian analysis from here on) to construct a generic multivariate probability density function (PDF) and derive a probabilistic model to predict multivariate outputs based on multivariate inputs. This conventional Bayesian analysis can address Considerations 1 and 2. It also partially addresses Consideration 3 using an ad hoc method of assembling incomplete generic data (Ching and Phoon 2014a; Ching et al. 2017b, 2018b). However, it does not address Consideration 4 because the multivariate PDF is constructed by generic data only. Because the transformation uncertainty for the generic data is significant, the predicted confidence interval for the design soil parameter is typically excessively large (Ching and Phoon 2014a; Ching et al. 2017b, 2018b). None of the aforementioned previous works can fully address all four considerations.

The purpose of this paper is to propose a hybridization method for constructing a site-specific multivariate PDF from a generic multivariate PDF (available from the conventional Bayesian updating analysis) and multivariate site-specific data. It will be shown that this hybrid model fulfills all considerations outlined previously. In particular, it is able to handle extremely sparse and incomplete site-specific data and it produces more reasonable confidence intervals than those from the conventional Bayesian analysis. The rest of this paper has the following structure. The proposed hybridization method will be presented under the scenario where a multivariate PDF for clay properties is to be constructed based on sparse site-specific data. First, a global clay database called CLAY/10/7490 (Ching and Phoon 2014b) as well as the *generic* multivariate PDF constructed based on CLAY/10/7490 (Ching and Phoon 2014a) will be briefly introduced. This generic multivariate PDF summarizes the global database. Second, a Bayesian framework of constructing the site-specific multivariate PDF purely based on site-specific data is presented. This site-specific multivariate PDF summarizes the site-specific database. Both multivariate PDFs operate in the common multivariate normal space (to be defined subsequently) where exact sampling is possible if appropriate conjugate priors are adopted. This underlying multivariate normality and the conjugate priors are the key assumptions of the proposed method. Third, a simple method of *hybridizing* the generic and site-specific multivariate PDFs is proposed so that the resulting multivariate PDF is governed by the generic PDF if there are very sparse site-specific data, whereas it is governed by the site-specific PDF if there are abundant site-specific data. The generic PDF constructed from the CLAY/10/7490 database is adopted in this paper as an example of generic information. In principle, the engineer is free to select a smaller subset of relevant sites that are considered comparable to the current site of interest to establish a more applicable generic PDF. Finally, two examples will be presented to demonstrate the effectiveness and usefulness of the proposed method. It will be clear that under the conventional Bayesian approach, the confidence interval for the site-specific design soil parameters is excessively large, whereas the hybrid approach produces more reasonable outcomes as shown in the subsequent examples.

Generic Multivariate Probability Density Function

The proposed method requires a generic multivariate PDF that is constructed from a generic database. The word *generic* is in the sense that the database covers a wider range of conditions than those encountered at a single site. It can be either a global database or a regional database. In the current paper, a previously developed global clay database named CLAY/10/7490 (Ching and Phoon 2014b) was adopted, although it can be replaced by another database containing more relevant sites (e.g., Ching and Phoon 2014b; D'Ignazio et al. 2016; Liu et al. 2016; Ching et al. 2017a, 2018a). The CLAY/10/7490 database consists of 7,490 data points for 10 clay parameters from 251 studies in the literature that cover 30 countries and regions worldwide. The proposed method is general: the generic database is not limited to clay databases. It can be a sand database or rock database.

The following is a brief description of the CLAY/10/7490 database that contains 10 dimensionless clay parameters (three index parameters, two in situ stress parameters, two strength parameters, and three CPT parameters):

$$
\begin{aligned}
Y_1 &= \ln(\mathrm{LL}) \\
Y_2 &= \ln(\mathrm{PI}) \\
Y_3 &= \mathrm{LI} \\
Y_4 &= \ln(\sigma_v'/P_a) \\
Y_5 &= \ln(\sigma_p'/P_a) \\
Y_6 &= \ln(s_u/\sigma_v') \\
Y_7 &= \ln(S_t) \\
Y_8 &= B_q \\
Y_9 &= \ln[(q_t - \sigma_v)/\sigma_v'] = \ln(q_{t1}) \\
Y_{10} &= \ln[(q_t - u_2)/\sigma_v'] = \ln(q_{tu})
\end{aligned}
\tag{1}
$$

where LL = liquid limit; PI = plasticity index; LI = liquidity index; σ_v' = vertical effective stress; σ_p' = preconsolidation stress; s_u = undrained shear strength; S_t = sensitivity; q_t = (corrected) cone tip resistance; u_2 = pore pressure behind cone; B_q = pore pressure ratio = $(u_2 - u_0)/(q_t - \sigma_v)$, with u_0 being hydrostatic pore pressure; P_a = atmospheric pressure = 101.3 kPa; $q_{t1} = (q_t - \sigma_v)/\sigma_v'$; and $q_{tu} = (q_t - u_2)/\sigma_v'$. The s_u values are all converted to the mobilized s_u values, denoted by s_u(mob), which is the in situ undrained shear strength mobilized in embankment and slope failures (Mesri and Huvaj 2007).

Based on CLAY/10/7490, the generic multivariate PDF for $(Y_1, Y_2, \ldots, Y_{10})$ was constructed by Ching and Phoon (2014a). The multivariate PDF is a translation PDF (Liu and Der Kiureghian 1986; Li et al. 2012). Basically, Y_i is connected to a standard normal X_i by the following transformations:

$$X_i = \Phi^{-1}[F_i(Y_i)] \tag{2}$$

$$Y_i = F_i^{-1}[\Phi(X_i)] \tag{3}$$

where F_i = cumulative distribution function (CDF) of Y_i; and Φ = CDF for a standard normal random variable. A suitable marginal PDF model should be fitted to the Y_i data to derive the CDF F_i. The Johnson system of distributions (Johnson 1949) was adopted by Ching and Phoon (2014a) to model the marginal PDFs of $(Y_1, Y_2, \ldots, Y_{10})$. With the Johnson system of distributions, Eqs. (2) and (3) reduce to the following concise equations for the transformations between X and Y (Ching and Phoon 2014a):

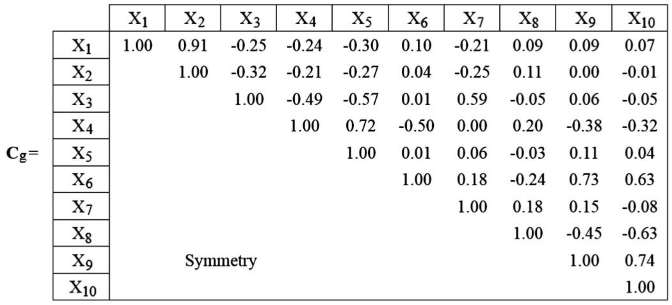
$$
\begin{aligned}
\frac{X_1 + 1.166}{1.636} &= \sinh^{-1}\left(\frac{Y_1 - 3.479}{0.616}\right) \\
\frac{X_2 + 0.265}{1.433} &= \sinh^{-1}\left(\frac{Y_2 - 3.178}{0.918}\right) \\
\frac{X_3 + 1.068}{1.434} &= \sinh^{-1}\left(\frac{Y_3 - 0.358}{0.629}\right) \\
\frac{X_4 - 0.256}{3.150} &= \ln\left(\frac{Y_4 + 7.010}{4.745 - Y_4}\right) \\
\frac{X_5 - 21.548}{4.600} &= \ln\left(\frac{Y_5 + 4.793}{571.992 - Y_5}\right) \\
\frac{X_6 + 0.517}{2.039} &= \sinh^{-1}\left(\frac{Y_6 + 1.461}{1.427}\right) \\
\frac{X_7 + 2.080}{2.393} &= \sinh^{-1}\left(\frac{Y_7 - 0.461}{1.885}\right) \\
\frac{X_8 - 0.161}{2.676} &= \sinh^{-1}\left(\frac{Y_8 - 0.615}{0.513}\right) \\
\frac{X_9 + 0.572}{1.340} &= \sinh^{-1}\left(\frac{Y_9 - 1.476}{0.659}\right) \\
\frac{X_{10} + 1.102}{2.134} &= \sinh^{-1}\left(\frac{Y_9 - 0.657}{1.154}\right)
\end{aligned}
\tag{4}
$$

where $\sinh(x) = [\exp(x) - \exp(-x)]/2$ is the hyperbolic sine function; and $\sinh^{-1}(x) = \ln[x + (1 + x^2)^{0.5}]$ is the inverse hyperbolic sine function. Using these equations, $(Y_1, Y_2, \ldots, Y_{10})$ can be converted to $(X_1, X_2, \ldots, X_{10})$. The resulting transformed variables $(X_1, X_2, \ldots, X_{10})$ roughly follow the standard normal distribution.

Note that $(X_1, X_2, \ldots, X_{10})$ does not necessarily collectively follow a multivariate normal PDF even if each X_i is marginally standard normal. Under the translation PDF, $(X_1, X_2, \ldots, X_{10})$ are *assumed* to follow the multivariate normal PDF

$$f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g) = N(\mathbf{x}|\mathbf{0}, \mathbf{C}_g) = |\mathbf{C}_g|^{-1/2}(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\mathbf{x}^T\mathbf{C}_g^{-1}\mathbf{x}\right] \tag{5}$$

where $N(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$ denotes a multivariate normal PDF for $\mathbf{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$; $n = 10$ is the dimension; $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is a normal variable vector (T means the vector or matrix transpose); $\mathbf{0} = (0, 0, \ldots, 0)^T$ is the zero vector; $\boldsymbol{\mu}_g = \mathbf{0}$ for standard normal variables; $\mathbf{C}_g$ is the covariance matrix for $(X_1, X_2, \ldots, X_{10})$ that characterizes the generic correlation among $(X_1, X_2, \ldots, X_{10})$; and the subscript g is to highlight that $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$ are for generic data. Fig. 1 shows the $\mathbf{C}_g$ matrix interpreted from CLAY/10/7490 (Ching and Phoon 2014a). In this study, $\boldsymbol{\mu}_g = \mathbf{0}$ and $\mathbf{C}_g$ in Fig. 1 are treated as a fixed vector and matrix without uncertainty because there is a significant amount of data in CLAY/10/7490, so the statistical uncertainty is relatively small.



$\mathbf{C_g}$=

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.00	0.91	-0.25	-0.24	-0.30	0.10	-0.21	0.09	0.09	0.07
X_2		1.00	-0.32	-0.21	-0.27	0.04	-0.25	0.11	0.00	-0.01
X_3			1.00	-0.49	-0.57	0.01	0.59	-0.05	0.06	-0.05
X_4				1.00	0.72	-0.50	0.00	0.20	-0.38	-0.32
X_5					1.00	0.01	0.06	-0.03	0.11	0.04
X_6						1.00	0.18	-0.24	0.73	0.63
X_7							1.00	0.18	0.15	-0.08
X_8								1.00	-0.45	-0.63
X_9		Symmetry							1.00	0.74
X_{10}										1.00

Fig. 1. $\mathbf{C}_g$ matrix for CLAY/10/7490. (Data from Ching and Phoon 2014a.)

The underlying multivariate normality for $(X_1, X_2, \ldots, X_{10})$ is the key assumption adopted by the current paper. Based on the authors' experience, soil parameters do not strictly follow a multivariate normal PDF, even when they are transformed individually to normal random variables. Nonetheless, Ching and Phoon (2012, 2014a) and Ching et al. (2017b, 2018a) performed fairly extensive validation studies to demonstrate that physically meaningful results can be obtained even in the absence of rigorous justification for this multivariate normality assumption. The details for this CLAY/10/7490 global database and the resulting multivariate PDF can be found in Ching and Phoon (2014a, b). The ensuing analysis in this paper will be presented in the $(X_1, X_2, \ldots, X_{10})$ space. In this standard normal space, the multivariate normal PDF $f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ summarizes the generic database CLAY/10/7490 in terms of capturing the dependencies between the 10 parameters.

Site-Specific Multivariate Probability Density Function

This section proposes a novel and computationally efficient method for constructing a site-specific multivariate PDF that can accommodate very sparse and incomplete site-specific data while quantifying the associated large statistical uncertainty correctly. Information from the generic multivariate PDF is not used in this section.

Site-specific data are typically sparse. Table 1 shows the site investigation results for a silty clay layer in a Taipei, Taiwan, site (Ou and Liao 1987). For this Taipei site, three boreholes and three CPT soundings were conducted. The data in Table 1 are based on one borehole and one nearby CPT sounding. The depth intervals for the data range from 0.5 to 2.6 m. The site investigation data in the table can be used to derive (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_9). Data for (Y_7, Y_8, Y_{10}) are completely missing. These data will be used to train the site-specific multivariate PDF. As a result, these data will be referred to as the training data. The training data are sparse in the following sense:

1. The number of boreholes and soundings is limited.
2. Test results (e.g., Atterberg's limits, σ_p', s_u) are only available at limited depths. One exception is the CPT sounding (q_c), which is continuous with depth.
3. Data at a given depth may be incomplete, i.e., there are missing data (empty entries).

These sparsity characteristics are commonly encountered in site investigation reports.

It is challenging to construct the site-specific multivariate PDF based on the sparse training data in Table 1 for the following reasons:

1. The amount of training data is limited, so there is significant statistical uncertainty for the PDF parameters. It is not trivial to quantify the statistical uncertainty.
2. Not many estimation methods can cope with missing training data.

Define (site-specific X_i) = $\Phi^{-1}[F_i(\text{site-specific } Y_i)]$, similar to Eq. (2). This CDF transform was previously used for the conversion between generic (X_i, Y_i). Now, the same CDF transform is used for the conversion between site-specific (X_i, Y_i). Basically, Eq. (4) is used as the transforms between (X, Y) for both generic and site-specific data. In this paper, it was further assumed that site-specific X_i is normal and that, moreover, site-specific $(X_1, X_2, \ldots, X_{10})$ is multivariate normal

$$f(\mathbf{x}|\boldsymbol{\mu}_s, \mathbf{C}_s) = N(\mathbf{x}|\boldsymbol{\mu}_s, \mathbf{C}_s)$$
$$= |\mathbf{C}_s|^{-1/2}(2\pi)^{-n/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_s)^T\mathbf{C}_s^{-1}(\mathbf{x}-\boldsymbol{\mu}_s)\right] \tag{6}$$

where $\boldsymbol{\mu}_s$ is the mean vector for site-specific $(X_1, X_2, \ldots, X_{10})$; $\mathbf{C}_s$ is the covariance matrix for site-specific $(X_1, X_2, \ldots, X_{10})$ that characterizes the site-specific correlation among $(X_1, X_2, \ldots, X_{10})$; and the subscript s is to highlight that $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ are for site-specific data. The multivariate normality for both generic and site-specific $(X_1, X_2, \ldots, X_{10})$ is the key assumption adopted by the current paper. With the multivariate normality, exact sampling is possible, which is important for processing big data sets efficiently. Site-specific X_i is not standard normal because it typically spans a narrower range than generic X_i. In particular, its mean value is in general not zero ($\boldsymbol{\mu}_s$ is not a zero vector). Also, $\mathbf{C}_s$ is not the same as $\mathbf{C}_g$. For example, the diagonal elements of $\mathbf{C}_s$ are not unity because the standard deviation of X_i is not 1.

Table 1. Site investigation results for a silty clay layer at a Taipei, Taiwan, site

				Test results (training data, $\mathbf{X}^o$)						
Depth (m)	s_u (kN/m^2)		s_u(mob) (kN/m^2)	LL (Y_1)	PI (Y_2)	LI (Y_3)	σ_v'/P_a (Y_4)	σ_p'/P_a (Y_5)	s_u(mob)/σ_v' (Y_6)	q_{t1} (Y_9)
12.8	UU	55.2	46.9	30.1	9.1	1.20	1.26	1.71	0.37	5.17
14.8	VST	50.7	52.9	32.8	12.8	1.43	1.43	N/A	0.36	4.22
16.1	UU	61.9	51.7	36.4	14.5	1.24	1.54	N/A	0.33	4.12
17.8	UU	54.2	42.8	41.9	18.9	0.90	1.68	1.79	0.25	4.03
18.3	VST	59.5	59.3	N/A	N/A	N/A	1.72	N/A	0.34	5.27
20.2	UU	73.1	60.5	38.1	17.3	0.70	1.88	N/A	0.32	4.53
22.7	VST	63.3	64.4	37.0	16.0	0.58	2.08	N/A	0.31	4.76
24.0	UU	82.2	67.5	38.0	16.2	0.75	2.19	2.19	0.30	5.12
26.6	UU	98.1	82.1	34.8	13.8	0.80	2.41	N/A	0.34	5.32

Note: UU = unconsolidated undrained test; and VST = vane shear test.

Bayesian Framework

The site-specific parameters $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ are unknown and are to be inferred by the site-specific training data alone. As mentioned previously, the statistical uncertainty in the inferred $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ can be significant if the training data are sparse, which is common for geotechnical site investigation. This is in contrast to the generic parameters $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$, which have relatively small statistical uncertainty because there is a significant amount of data in CLAY/10/7490. The statistical uncertainty for $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ will be rigorously quantified by the novel Bayesian analysis proposed in this section.

In Bayesian analysis, the prior PDFs $f(\boldsymbol{\mu}_s)$ and $f(\mathbf{C}_s)$ characterize the uncertainty in $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ prior to any site investigation. At the site-specific level, we are updating $f(\boldsymbol{\mu}_s)$ and $f(\mathbf{C}_s)$ into their posterior PDFs $f(\boldsymbol{\mu}_g|\mathbf{X}^o)$ and $f(\mathbf{C}_g|\mathbf{X}^o)$ to capture site-specific statistical uncertainty. One can adopt uniform noninformative (flat) prior PDFs for both $f(\boldsymbol{\mu}_s)$ and $f(\mathbf{C}_s)$. Uniform priors with typical ranges are particularly useful when there are very limited data (e.g., Cao et al. 2016). However, these uniform priors are not conjugate to the multivariate normal model, so the exact sampling will not be possible. To maintain the prior conjugacy, it is desirable to adopt the conjugate priors and tune their parameters so that they become noninformative.

Conjugate prior PDFs for $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ exist because $(X_1, X_2, \ldots, X_{10})$ is assumed to be multivariate normal. Subsequently, we will show that exact sampling is possible because of the use of conjugate priors. The conjugate prior PDF for $\boldsymbol{\mu}_s$ is multivariate normal (Gelman et al. 2013)

$$f(\boldsymbol{\mu}_s) = N(\boldsymbol{\mu}_s|\boldsymbol{\mu}_0, \mathbf{C}_0) = |\mathbf{C}_0|^{-1/2} \cdot (2\pi)^{-n/2} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_0)^T \mathbf{C}_0^{-1} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_0)\right] \quad (7)$$

where $\boldsymbol{\mu}_0$ and $\mathbf{C}_0$ are the prior mean vector and prior covariance matrix for $\boldsymbol{\mu}_s$ ($\boldsymbol{\mu}_0$ and $\mathbf{C}_0$ are prescribed and fixed). If $\boldsymbol{\mu}_0$ is taken to be a zero vector and $\mathbf{C}_0$ is taken to be a diagonal matrix with very large diagonal elements, the prior PDF $f(\boldsymbol{\mu}_s)$ tends to be noninformative. The conjugate prior PDF for $\mathbf{C}_s$ is inverse Wishart (Gelman et al. 2013)

$$f(\mathbf{C}_s) = \mathrm{IW}(\mathbf{C}_s|\boldsymbol{\Sigma}, \nu) = \frac{|\boldsymbol{\Sigma}|^{\nu/2}}{2^{n\nu/2} \cdot \Gamma_n(\nu/2)} \cdot |\mathbf{C}_s|^{-\frac{\nu+n+1}{2}} \cdot \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma} \times \mathbf{C}_s^{-1})\right] \quad (8)$$

where $\mathrm{IW}(\mathbf{C}_s|\boldsymbol{\Sigma}, \nu)$ denotes the inverse-Wishart distribution (James 1964; Mardia et al. 1979) with scale matrix $\boldsymbol{\Sigma}$ and degree of freedom ν; $\Gamma_n(\cdot)$ is the multivariate gamma function (James 1964) with dimension = n; and $\mathrm{tr}(\cdot)$ is the matrix trace, i.e., the summation of the diagonal elements of a square matrix. It is challenging to make the prior PDF $f(\mathbf{C}_s)$ truly noninformative simply by adopting certain $(\boldsymbol{\Sigma}, \nu)$. A noninformative $f(\mathbf{C}_s)$ means that all Pearson correlation coefficients extracted from $\mathbf{C}_s$ are independent and uniformly distributed over $[-1, 1]$ and all variances (or standard deviations) extracted from $\mathbf{C}_s$ are independent and have relatively flat distributions. Also, the correlation coefficients are weakly correlated with the standard deviations. A popular choice is to let $\boldsymbol{\Sigma}$ = identity matrix and $\nu = n + 1$. This choice induces Pearson correlation coefficients uniformly distributed over $[-1, 1]$. However, the marginal distribution for the standard deviation is not flat [low density near zero; see Gelman (2006)], and there is a dependence between correlation coefficient and standard deviation (Tokuda et al. 2011; Alvarez et al. 2014). The nonflat prior for the standard deviation is undesirable because the proposed Bayesian method is meant to be applied to a wide variety of geotechnical parameters and the coefficients of variation (COVs) for different geotechnical parameters can vary significantly. It is not desirable to have a nonflat prior for standard deviation that can potentially prefer certain COV values. The prior dependency between correlation coefficient and standard deviation is also undesirable because there is no strong evidence in the literature to support such dependence. To make $f(\mathbf{C}_s)$ more noninformative while maintaining the prior conjugacy, Huang and Wand (2013) proposed the following hierarchical $f(\mathbf{C}_s)$:

$$f(\mathbf{C}_s|a_1, a_2, \ldots, a_n) = \mathrm{IW}(\mathbf{C}_s|\boldsymbol{\Sigma}, \nu)$$

$$\boldsymbol{\Sigma} = 2(\nu - n + 1) \cdot \begin{bmatrix} 1/a_1 & 0 & \cdots & 0 \\ 0 & 1/a_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1/a_n \end{bmatrix}$$

$$f(a_i) = \mathrm{IG}(a_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot a_i^{-\alpha-1} \cdot \exp\left(-\frac{\beta}{a_i}\right) \quad (9)$$

where $(a_1, a_2, \ldots, a_n)$ are hyperparameters (also random) that parameterize the scale matrix $\boldsymbol{\Sigma}$; $\mathrm{IG}(a|\alpha, \beta)$ denotes the inverse-gamma distribution with shape parameter α and scale parameter β; and $\Gamma(\cdot)$ is the gamma function. Huang and Wand (2013) showed that if ν is taken to be $n + 1$, α is taken to be 0.5, and β is taken to be a small number, not only are the Pearson correlation coefficients uniformly distributed over $[-1, 1]$, but the marginal distribution for the resulting standard deviation is also flat. Moreover, dependence between correlation coefficient and standard deviation become significantly weaker. Selecting the previous noninformative priors for the PDFs of $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ is important because prior information does not exist in a realistic geotechnical setting and this study takes a position that $\boldsymbol{\mu}_s$ and $\mathbf{C}_s$ can be quite different from their respective generic versions ($\boldsymbol{\mu}_g$ and $\mathbf{C}_g$).

Let $\mathbf{X}$ be an $m \times 10$ site-specific data matrix converted from the $m \times 10$ $\mathbf{Y}$ data matrix in Table 1 through Eq. (4). Let $\mathbf{x}_j^T$ be the jth row of $\mathbf{X}$: $\mathbf{x}_j^T$ is a 1×10 vector containing the $(X_1, \ldots, X_{10})$ values for the jth depth. For the time being, we assume that all entries in $\mathbf{X}$ are observed (no missing entries in Table 1). If the site-specific data at different depths are independent (e.g., depth intervals are larger than the vertical scale of fluctuation), the multivariate PDF for $\mathbf{X}$ can be written as

$$f(\mathbf{X}|\boldsymbol{\mu}_s, \mathbf{C}_s) = \left[\prod_{j=1}^m N(\mathbf{x}_j|\boldsymbol{\mu}_s, \mathbf{C}_s)\right] = |\mathbf{C}_s|^{-m/2}(2\pi)^{-(m\cdot n)/2} \exp\left[-\frac{1}{2}\sum_{j=1}^m (\mathbf{x}_j - \boldsymbol{\mu}_s)^T \mathbf{C}_s^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_s)\right] \quad (10)$$

The complete multivariate PDF for all variables can be written as

$$f(\mathbf{X},\boldsymbol{\mu}_s,\mathbf{C}_s,\mathbf{a}) = f(\mathbf{X}|\boldsymbol{\mu}_s,\mathbf{C}_s)\cdot f(\boldsymbol{\mu}_s)\cdot f(\mathbf{C}_s|\mathbf{a})\cdot f(\mathbf{a}) = \left[\prod_{j=1}^{m} N(\mathbf{x}_j|\boldsymbol{\mu}_s,\mathbf{C}_s)\right]\cdot N(\boldsymbol{\mu}_s|\boldsymbol{\mu}_0,\mathbf{C}_0)\cdot \mathrm{IW}(\mathbf{C}_s|\boldsymbol{\Sigma},\nu)\cdot\left[\prod_{i=1}^{n}\mathrm{IG}(a_i|\alpha,\beta)\right]$$

$$\propto |\mathbf{C}_s|^{-(m+2n+2)/2}\cdot|\boldsymbol{\Sigma}|^{(n+1)/2}\cdot\left(\prod_{i=1}^{n}a_i^{-\alpha-1}\right)\cdot e^{-(1/2)\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)^{\mathrm{T}}\mathbf{C}_s^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_s)-(1/2)\boldsymbol{\mu}_s^{\mathrm{T}}\mathbf{C}_0^{-1}\boldsymbol{\mu}_s-(1/2)\mathrm{tr}(\boldsymbol{\Sigma}\times\mathbf{C}_s^{-1})-\sum_{i=1}^{n}(\beta/a_i)} \tag{11}$$

where $\mathbf{a}$ denotes $(a_1, a_2, \ldots, a_n)$.

Gibbs Sampler

Gibbs sampler (GS) (Geman and Geman 1984; Gilks et al. 1996) is a special instance for the Markov chain Monte Carlo methods (Metropolis et al. 1953; Hastings 1970; Gilks et al. 1996). The basic idea of GS is to decompose the random parameters into groups and to randomly sample each group by conditioning on the remaining groups. By doing so sequentially among all groups, it can be shown that the samples are asymptotically distributed as the desired distribution. In our case, the random parameters include ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$, $\mathbf{a}$). With the assumed multivariate normality and conjugate priors discussed previously, the following conditional PDFs have conjugate forms ($\boldsymbol{\mu}_s$ remains multivariate normal, $\mathbf{C}_s$ remains inverse Wishart, and a_i remains inverse gamma)

$$f(\boldsymbol{\mu}_s|\mathbf{X},\mathbf{C}_s,\mathbf{a}) \propto e^{-(1/2)\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)^{T}\mathbf{C}_s^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_s)-(1/2)\boldsymbol{\mu}_s^{T}\mathbf{C}_0^{-1}\boldsymbol{\mu}_s} \propto e^{-(1/2)[\boldsymbol{\mu}_s^{T}(\mathbf{C}_0^{-1}+m\mathbf{C}_s^{-1})\boldsymbol{\mu}_s-2\left(\sum_{j=1}^{m}\mathbf{C}_s^{-1}\mathbf{x}_j\right)^{T}\boldsymbol{\mu}_s]}$$

$$= N\left\{\boldsymbol{\mu}_s|(\mathbf{C}_0^{-1}+m\mathbf{C}_s^{-1})^{-1}\left(\sum_{j=1}^{m}\mathbf{C}_s^{-1}\mathbf{x}_j\right),(\mathbf{C}_0^{-1}+m\mathbf{C}_s^{-1})^{-1}\right\} \tag{12}$$

$$f(\mathbf{C}_s|\mathbf{X},\boldsymbol{\mu}_s,\mathbf{a}) \propto |\mathbf{C}_s|^{-(m+2n+2)/2}\cdot e^{-(1/2)\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)^{T}\mathbf{C}_s^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_s)-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}\times\mathbf{C}_s^{-1})} \propto |\mathbf{C}_s|^{-[(m+n+1)+n+1]/2}\cdot e^{-(1/2)\mathrm{tr}\left(\left[\boldsymbol{\Sigma}+\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)(\mathbf{x}_j-\boldsymbol{\mu}_s)^{T}\right]\times\mathbf{C}_s^{-1}\right)}$$

$$= \mathrm{IW}\left\{\mathbf{C}_s|\boldsymbol{\Sigma}+\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)(\mathbf{x}_j-\boldsymbol{\mu}_s)^{T}, n+m+1\right\} \tag{13}$$

$$f(a_i|\mathbf{X},\boldsymbol{\mu}_s,\mathbf{C}_s,\mathbf{a}_{\backslash i}) \propto a_i^{-(n+1)/2-\alpha-1}\cdot e^{-(1/2)\mathrm{tr}(\boldsymbol{\Sigma}\times\mathbf{C}_s^{-1})-(\beta/a_i)} \propto a_i^{-[\alpha+(n+1)/2]-1}\cdot e^{-(\beta+2\mathbf{C}_{s,ii}^{-1})/a_i} = \mathrm{IG}\left(a_i|\alpha+\frac{n+1}{2},\beta+2\mathbf{C}_{s,\mathrm{ii}}^{-1}\right) \tag{14}$$

where $\mathbf{a}_{\backslash i}$ denotes $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$; and $\mathbf{C}_{s,ii}^{-1}$ denotes the (i, i) entry in the $\mathbf{C}_s^{-1}$ matrix. Therefore, GS can be conducted effectively because exact sampling from one group conditional on the remaining groups is possible. Unlike common Markov chain Monte Carlo methods, GS never rejects samples and does not require a proposal PDF. With GS, it is also possible to accommodate missing entries in $\mathbf{X}$. Denote all missing (unknown) entries in $\mathbf{x}_j$ by $\mathbf{x}_j^u$ and all nonmissing (observed) entries by $\mathbf{x}_j^o$. The covariance matrix $\mathbf{C}_s$ is also partitioned accordingly. The PDF of $\mathbf{x}_j^u$ conditioned on $\mathbf{x}_j^o$ is still multivariate normal

$$f(\mathbf{x}_j^u|\mathbf{x}_j^o,\boldsymbol{\mu}_s,\mathbf{C}_s,\mathbf{a}) = N\{\mathbf{x}_j^u|\boldsymbol{\mu}_s^u+\mathbf{C}_s^{uo}(\mathbf{C}_s^o)^{-1}(\mathbf{x}_j^o-\boldsymbol{\mu}_s^o),\mathbf{C}_s^u - \mathbf{C}_s^{uo}(\mathbf{C}_s^o)^{-1}\mathbf{C}_s^{ou}\} \tag{15}$$

In the current paper, it was assumed that the noninformative priors parameters were used: $\boldsymbol{\mu}_0$ = zero vector, $\mathbf{C}_0$ = diagonal matrix with diagonal elements equal to 10^4, $\nu = n+1$, $\alpha = 0.5$, and $\beta = 10^{-4}$. The GS iterates over the following steps:

1. Initialize ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$) samples at arbitrary values, where $\mathbf{X}^u = \mathbf{x}_1^u, \mathbf{x}_2^u, \ldots, \mathbf{x}_m^u$ denotes all missing entries in $\mathbf{X}$. Note that $\mathbf{X}$ does not contain missing entries after the initialization.
2. Draw $\boldsymbol{\mu}_s$ sample from $f(\boldsymbol{\mu}_s|\mathbf{X},\mathbf{C}_s,\mathbf{a})$, a multivariate normal PDF with the following mean vector and covariance matrix (MATLAB command mvnrnd):

$$\boldsymbol{\mu}_s \sim N\left\{\boldsymbol{\mu}_s|(\mathbf{C}_0^{-1}+m\mathbf{C}_s^{-1})^{-1}\left(\sum_{j=1}^{m}\mathbf{C}_s^{-1}\mathbf{x}_j\right),(\mathbf{C}_0^{-1}+m\mathbf{C}_s^{-1})^{-1}\right\} \tag{16}$$

3. Draw $\mathbf{C}_s$ sample from $f(\mathbf{C}_s|\mathbf{X},\boldsymbol{\mu}_s,\mathbf{a})$, an inverse-Wishart PDF with the following scale matrix and degrees of freedom (MATLAB command iwishrnd):

$$\mathbf{C}_s \sim \mathrm{IW}\left\{\mathbf{C}_s|\boldsymbol{\Sigma}+\sum_{j=1}^{m}(\mathbf{x}_j-\boldsymbol{\mu}_s)(\mathbf{x}_j-\boldsymbol{\mu}_s)^{T}, n+m+1\right\} \tag{17}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(4/a_1, 4/a_2, \ldots, 4/a_n)$.

4. For $i = 1, \ldots, n$, draw a_i sample from $f(a_i|\mathbf{X},\boldsymbol{\mu}_s,\mathbf{C}_s,\mathbf{a}_{\backslash i})$, an inverse-gamma PDF with the following shape and scale parameters [MATLAB command $1/\mathrm{gamrnd}(\mathrm{shape}, 1/\mathrm{scale})$]:

$$a_i \sim \mathrm{IG}\left(a_i|\frac{n+2}{2}, 10^{-4}+2\mathbf{C}_{s,ii}^{-1}\right) \tag{18}$$

5. For $j = 1, \ldots, m$, draw $\mathbf{x}_j^u$ sample from $f(\mathbf{x}_j^u|\mathbf{x}_j^o,\boldsymbol{\mu}_s,\mathbf{C}_s,\mathbf{a})$, a multivariate normal PDF with the following mean vector and covariance matrix (MATLAB command mvnrnd):

$$\mathbf{x}_j^u \sim N\{\mathbf{x}_j^u|\boldsymbol{\mu}_s^u + \mathbf{C}_s^{uo}(\mathbf{C}_s^o)^{-1}(\mathbf{x}_j^o - \boldsymbol{\mu}_s^o), \mathbf{C}_s^u - \mathbf{C}_s^{uo}(\mathbf{C}_s^o)^{-1}\mathbf{C}_s^{ou}\} \tag{19}$$

6. Cycle Steps 2–5 T times to obtain T samples for ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$).

GS starts with an initial sample of ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$), then it sequentially draws samples from the conditional PDFs based on the latest parameter values. The ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$, $\mathbf{a}$, $\mathbf{X}^u$) samples after the burn-in period (determined by visual inspection) are collected. These samples are distributed as the posterior PDFs $f(\boldsymbol{\mu}_s, \mathbf{C}_s, \mathbf{a}, \mathbf{X}^u|\mathbf{X}^o)$, where $\mathbf{X}^o = (\mathbf{x}_1^o, \mathbf{x}_2^o, \ldots, \mathbf{x}_m^o)$ denotes the collection of all training data. The GS produces the exact solution for the Bayesian analysis in the sense that it draws samples from the full posterior PDFs $f(\boldsymbol{\mu}_s, \mathbf{C}_s, \mathbf{a}, \mathbf{X}^u|\mathbf{X}^o)$. This exact sampling is possible because of the assumed multivariate normality and the use of conjugate priors. The scatter of the ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) samples quantifies the site-specific statistical uncertainty. Also, the applicability of the proposed GS method is independent of the nature of the data. It can be used to construct the site-specific PDF model for clays, sands, or rocks: namely, it is a machine learning framework that is purely driven by data. Bayesian machine learning methods such as Bayesian network (Heckerman et al. 1995), Bayesian neural network (MacKay 1995), Gaussian processes (Rasmussen and Williams 2006), relevance vector machine (Tipping 2001), Bayesian deep learning (Wang and Yeung 2016), Bayesian model class selection (Beck and Yuen 2004; Yuen 2010), and Bayesian simulation (MacKay 1998; Gilks et al. 1996; Doucet et al. 2001) have made significant advancement in recent years. The GS method proposed in the current study belongs to Bayesian simulation methods.

Site-Specific Prediction

It is of practical interest to simulate the properties of a clay layer at a new depth ($\mathbf{x}_{\text{new}}$) that does not appear in the training data in Table 1. It is clear that such a simulation is only meaningful if points located within the same geologic layer are selected. The argument here is that the properties at these points belong to the same population as the training data. By assuming $\mathbf{x}_{\text{new}}$ to be from the same population as $\mathbf{X}$, the multivariate PDF for $\mathbf{x}_{\text{new}}$ is also multivariate normal with mean = $\boldsymbol{\mu}_s$ and covariance matrix = $\mathbf{C}_s$. However, ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) are uncertain and their conditional samples have been obtained by GS. Based on the total probability theorem, the conditional multivariate PDF $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ is a mixture of multivariate normal PDFs

$$\begin{aligned} f(\mathbf{x}_{\text{new}}|\mathbf{X}^o) &= \int f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_s, \mathbf{C}_s) \cdot f(\boldsymbol{\mu}_s, \mathbf{C}_s|\mathbf{X}^o) \cdot d\boldsymbol{\mu}_s d\mathbf{C}_s \\ &\approx \frac{1}{T - t_b}\left[\sum_{t=t_b+1}^{T} N(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t})\right] \end{aligned} \tag{20}$$

where $(\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t})$ are the GS samples at time step t; and t_b = end of the burning period. Samples for $\mathbf{x}_{\text{new}}$ can be readily sampled using the following steps:

1. Sample the t index randomly among the indexes ($t_b + 1$, $t_b + 2, \ldots, T$).
2. Sample $\mathbf{x}_{\text{new}} \sim N(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t})$, where t = sampled t in Step 1.

These samples have incorporated the site-specific training data. The statistical uncertainty due to limited training data is also characterized by the samples.

It is also possible that certain properties in $\mathbf{x}_{\text{new}}$ have been observed, denoted by $\mathbf{x}_{\text{new}}^o$. For instance, the effective overburden stress (σ_v') can usually be accurately estimated even at unexplored depths. Denote the remaining unobserved properties by $\mathbf{x}_{\text{new}}^u$. Let ($\boldsymbol{\mu}_{s,t}$, $\mathbf{C}_{s,t}$) be also partitioned accordingly. The conditional PDF $f(\mathbf{x}_{\text{new}}^u|\mathbf{X}^o, \mathbf{x}_{\text{new}}^o)$ can be expressed as

$$\begin{aligned} f(\mathbf{x}_{\text{new}}^u|\mathbf{X}^o, \mathbf{x}_{\text{new}}^o) &\propto \sum_{t=t_b+1}^{T} N(\mathbf{x}_{\text{new}}^u, \mathbf{x}_{\text{new}}^o|\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t}) \\ &\propto \sum_{t=t_b+1}^{T} w_t \times N(\mathbf{x}_{\text{new}}^u|\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t}, \mathbf{x}_{\text{new}}^o) \end{aligned} \tag{21}$$

where w_t = weight for each multivariate normal PDF

$$\begin{aligned} w_t &= N(\mathbf{x}_{\text{new}}^o|\boldsymbol{\mu}_{s,t}^o, \mathbf{C}_{s,t}^o) \\ &= |\mathbf{C}_{s,t}^o|^{-1/2} \times (2\pi)^{-n/2} \times \exp\left[-\frac{1}{2}\boldsymbol{\mu}_{s,t}^{o\,T}(\mathbf{C}_{s,t}^o)^{-1}\boldsymbol{\mu}_{s,t}^o\right] \end{aligned} \tag{22}$$

The previous steps can be modified to draw $\mathbf{x}_{\text{new}}^u$ samples that are distributed as $f(\mathbf{x}_{\text{new}}^u|\mathbf{X}^o, \mathbf{x}_{\text{new}}^o)$:

1. Sample the t index among ($t_b + 1, t_b + 2, \ldots, T$) according to the normalized weights $w_t/(w_{tb+1} + w_{tb+2} + \cdots + w_T)$, where w_t is defined in Eq. (22).
2. Sample $\mathbf{x}_{\text{new}}^u \sim N\{\mathbf{x}_{\text{new}}^u|\boldsymbol{\mu}_{s,t}^u + \mathbf{C}_{s,t}^{uo} + (\mathbf{C}_{s,t}^o)^{-1}(\mathbf{x}_{\text{new}}^o - \boldsymbol{\mu}_{s,t}^o), \mathbf{C}_{s,t}^u - \mathbf{C}_{s,t}^{uo}(\mathbf{C}_{s,t}^o)^{-1}\mathbf{C}_{s,t}^{ou}\}$.

These $\mathbf{x}_{\text{new}}^u$ samples can be further converted to the physical soil parameters $\mathbf{y}_{\text{new}}^u$ through Eq. (4). The $\mathbf{y}_{\text{new}}^u$ samples can be used to predict the statistics of the design parameters of the clay at the new depth.

Hybridizing Site-Specific and Generic Multivariate PDFs

When site-specific training data are very sparse, the statistical uncertainty in ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) can be significant, so the soil properties to be predicted can be quite uncertain. In this case, it is sensible to rely more on generic data. This is reasonable: if local experience is absent, a reasonable choice is to rely on generic experience. This is in line with current standard practice where a desk study is integral to site investigation. In this paper, prediction purely based on generic experience is equivalent to drawing $\mathbf{x}_{\text{new}}$ samples from $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g) = N(\mathbf{x}|\mathbf{0}, \mathbf{C}_g)$. In contrast, when site-specific training data are abundant, it is sensible to rely more on site-specific data. In this paper, prediction purely based on site-specific data is equivalent to drawing $\mathbf{x}_{\text{new}}$ samples from $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$. In this section, a method is proposed to hybridize the generic and site-specific multivariate PDFs so that the hybrid PDF approaches the generic PDF $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ when site-specific data are sparse and approaches the site-specific PDF $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ when site-specific data are abundant.

The idea of hybridization proposed in this study is straightforward: the hybrid multivariate PDF is proportional to the direct product between the generic and site-specific multivariate PDFs

$$f(\mathbf{x}_{\text{new}}|hb) \propto f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g) \cdot f(\mathbf{x}_{\text{new}}|\mathbf{X}^o) \tag{23}$$

where $f(\mathbf{x}_{\text{new}}|hb)$ denotes the hybrid multivariate PDF; $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ is the generic multivariate PDF [Eq. (5)]; and $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ is the site-specific multivariate PDF [Eq. (20)]. Fig. 2 illustrates this idea of hybridization and explains why it works. The generic PDF $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ (the solid curves in the figure) does not change with respect to the amount of site-specific training data because it only depends on $\boldsymbol{\mu}_g$ and $\mathbf{C}_g$. However, the site-specific PDF $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ depends on the amount of site-specific training data

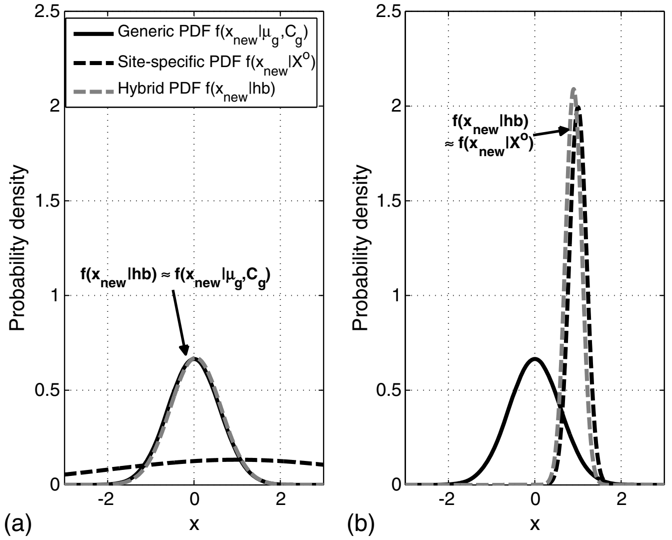


Fig. 2. Hybridization: (a) sparse site-specific data; and (b) abundant site-specific data.

$\mathbf{X}^o$: it is relatively flat when $\mathbf{X}^o$ is sparse [Fig. 2(a)] and is relatively peaked when $\mathbf{X}^o$ is abundant [Fig. 2(b)]. When $\mathbf{X}^o$ is sparse [Fig. 2(a)], $f(\mathbf{x}_{\text{new}}|hb) \propto f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g) \times$ (a relatively flat PDF) $\approx f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$, hence the hybrid PDF approaches $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$. When site-specific training data are abundant [Fig. 2(b)], the opposite happens: the hybrid PDF $\propto$ (a relatively flat PDF) $\times f(\mathbf{x}_{\text{new}}|\mathbf{X}^o) \approx f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$, hence the hybrid PDF approaches $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$.

In its process, hybridization is similar to multiplying the prior PDF and likelihood function together to produce the posterior PDF in Bayesian updating. However, hybridization multiplies $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ and $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ to produce $f(\mathbf{x}_{\text{new}}|hb)$, whereas the conventional Bayesian analysis multiplies the prior $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ and the likelihood $f(\mathbf{X}^o|\mathbf{x}_{\text{new}}, \boldsymbol{\mu}_g, \mathbf{C}_g)$ to produce the posterior $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o, \boldsymbol{\mu}_g, \mathbf{C}_g)$. The conventional Bayesian analysis assumes that both $\mathbf{X}^o$ and $\mathbf{x}_{\text{new}}$ have mean and covariance ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$). This may not be true because $\mathbf{X}^o$ and $\mathbf{x}_{\text{new}}$ are site specific, so their mean and covariance may not be ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$). In contrast, the hybridization adopts $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$, in which $\mathbf{X}^o$ and $\mathbf{x}_{\text{new}}$ have mean and covariance equal to ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) that are disconnected with ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$). Moreover ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) are assumed to follow the noninformative prior PDF before $\mathbf{X}^o$ is known. After $\mathbf{X}^o$ is known, their posterior PDF is updated using GS so that $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ captures the statistical uncertainty in ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) conditioned on $\mathbf{X}^o$. When $\mathbf{X}^o$ is very sparse, $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ becomes very flat [Fig. 2(b)] and not very useful. The hybridization adopts the generic PDF $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ to multiply to $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ so that the hybrid $f(\mathbf{x}_{\text{new}}|hb)$ resembles $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$. By doing so, $f(\mathbf{x}_{\text{new}}|hb)$ is still useful even when $\mathbf{X}^o$ is very sparse.

By combining Eqs. (5) and (20), it can be shown that the hybrid PDF $f(\mathbf{x}_{\text{new}}|hb)$ is still a mixture of multivariate normal PDF

$$\begin{aligned} f(\mathbf{x}_{\text{new}}|hb) &\propto f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g) \cdot f(\mathbf{x}_{\text{new}}|\mathbf{X}^o) \\ &\propto N(\mathbf{x}_{\text{new}}|\mathbf{0}, \mathbf{C}_g) \times \left[\sum_{t=t_b+1}^{T} N(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_{s,t}, \mathbf{C}_{s,t})\right] \\ &\propto \sum_{t=t_b+1}^{T} \mathrm{w}_t \times N(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_{hb,t}, \mathbf{C}_{hb,t}) \end{aligned} \tag{24}$$

where w_t = weight for each multivariate normal PDF

$$w_t = |\mathbf{C}_g + \mathbf{C}_{s,t}|^{-1/2} \times (2\pi)^{-n/2} \times \exp\left[-\frac{1}{2}\boldsymbol{\mu}_{s,t}^T(\mathbf{C}_g + \mathbf{C}_{s,t})^{-1}\boldsymbol{\mu}_{s,t}\right] \tag{25}$$

and

$$\begin{aligned} \boldsymbol{\mu}_{hb,t} &= (\mathbf{C}_g^{-1} + \mathbf{C}_{s,t}^{-1})^{-1}\mathbf{C}_{s,t}^{-1}\boldsymbol{\mu}_{s,t} \\ \mathbf{C}_{hb,t} &= (\mathbf{C}_g^{-1} + \mathbf{C}_{s,t}^{-1})^{-1} \end{aligned} \tag{26}$$

The hybrid PDF $f(\mathbf{x}_{\text{new}}|hb)$ can be readily sampled using the following steps:

1. Sample the t index among $(t_b+1, t_b+2, \ldots, T)$ according to the normalized weights $w_t/(w_{tb+1} + w_{tb+2} + \cdots + w_T)$, where w_t is defined in Eq. (25).
2. Sample $\mathbf{x}_{\text{new}} \sim N(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_{hb,t}, \mathbf{C}_{hb,t})$, where t = sampled t in Step 1.

These samples have incorporated training data $\mathbf{X}^o$. The statistical uncertainty for ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) is also characterized by the samples. The samples also have incorporated the generic multivariate PDF $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ by hybridization.

In the case where certain properties in $\mathbf{x}_{\text{new}}$ have been observed, denoted by $\mathbf{x}^o_{\text{new}}$, the conditional hybrid PDF $f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}})$ should be used

$$\begin{aligned} f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}}) &\propto f(\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g, \mathbf{x}^o_{\text{new}}) \cdot f(\mathbf{x}^u_{\text{new}}|\mathbf{X}^o, \mathbf{x}^o_{\text{new}}) \\ &\propto \sum_{t=t_b+1}^{T} w_t \times N(\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}_{hb,t}, \mathbf{C}_{hb,t}, \mathbf{x}^o_{\text{new}}) \end{aligned} \tag{27}$$

where $f(\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g, \mathbf{x}^o_{\text{new}})$ is the conditional generic PDF; $f(\mathbf{x}^u_{\text{new}}|\mathbf{X}^o, \mathbf{x}^o_{\text{new}})$ is the conditional site-specific PDF; $f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}})$ is the conditional hybrid PDF; and w_t is the weight for each conditional hybrid PDF

$$\begin{aligned} w_t = {} & |\mathbf{C}_g + \mathbf{C}_{s,t}|^{-1/2} \times |\mathbf{C}^o_{s,t}|^{-1/2} \times (2\pi)^{-n} \\ & \times \exp\left[-\frac{1}{2}\boldsymbol{\mu}_{s,t}^T(\mathbf{C}_g + \mathbf{C}_{s,t})^{-1}\boldsymbol{\mu}_{s,t} - \frac{1}{2}\boldsymbol{\mu}^{o}_{s,t}{}^T(\mathbf{C}^o_{s,t})^{-1}\boldsymbol{\mu}^o_{s,t}\right] \end{aligned} \tag{28}$$

The preceding steps can be slightly modified to draw $\mathbf{x}^u_{\text{new}}$ samples from $f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}})$:

1. Sample the t index among $(t_b+1, t_b+2, \ldots, T)$ according to the normalized weights $w_t/(w_{tb+1} + w_{tb+2} + \cdots + w_T)$, where w_t is defined in Eq. (28).
2. Sample $\mathbf{x}^u_{\text{new}} \sim N\{\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}^u_{hb,t} + \mathbf{C}^{uo}_{hb,t}(\mathbf{C}^o_{hb,t})^{-1}(\mathbf{x}^o_{\text{new}} - \boldsymbol{\mu}^o_{hb,t}), \mathbf{C}^u_{hb,t} - \mathbf{C}^{uo}_{hb,t}(\mathbf{C}^o_{hb,t})^{-1}\mathbf{C}^{ou}_{hb,t}\}$.

These $\mathbf{x}^u_{\text{new}}$ samples can be further converted to the physical soil parameters $\mathbf{y}^u_{\text{new}}$ through the inverse transform of Eq. (4). The $\mathbf{y}^u_{\text{new}}$ samples can be used to predict the statistics of the design parameters of the clay at the new depth.

Case Study

Sweden Case

Consider a clay site at Lilla Mellösa, Sweden, extracted from D'Ignazio et al. (2016). Table 2 shows the site investigation results of this clay site. The site investigation data in the table can be used to derive $(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$. Data for $(Y_1, Y_2, \ldots, Y_6)$ are complete, whereas data for (Y_7, Y_8, Y_9, Y_{10}) are missing. The $(X_1, X_2, \ldots, X_6)$ data can be obtained from $(Y_1, Y_2, \ldots, Y_6)$ through Eq. (4). The data were divided into two groups: 13 training

Table 2. Site investigation results for a clay site at Lilla Mellösa, Sweden

Depth	s_u value (VST)	s_u(mob)	Test indexes					
(m)	(kN/m^2)	(kN/m^2)	LL (Y_1)	PI (Y_2)	LI (Y_3)	σ'_v/P_a (Y_4)	σ'_p/P_a (Y_5)	s_u(mob)$/\sigma'_v$ (Y_6)
				Training cases, $\mathbf{X}^o$				
2.1	8.7	5.64	129.7	82.2	1.01	0.15	0.21	0.38
3.6	8.6	5.59	124.2	80.5	0.88	0.22	0.25	0.26
4.2	9.4	6.17	119.3	78.3	0.90	0.24	0.28	0.25
5.0	10.3	7.00	110.0	71.8	0.98	0.28	0.32	0.25
5.7	10.8	7.47	105.1	69.0	0.94	0.32	0.35	0.23
6.4	11.2	7.74	100.7	69.0	0.95	0.35	0.40	0.22
7.9	13.2	9.86	84.8	57.5	0.97	0.43	0.49	0.23
8.5	14.2	10.65	82.1	55.9	1.01	0.46	0.54	0.23
9.0	17.0	13.22	76.0	51.0	0.88	0.49	0.64	0.26
9.1	15.3	11.70	78.8	53.7	0.99	0.50	0.58	0.23
9.9	17.4	13.49	73.8	51.5	0.97	0.55	0.64	0.24
10.7	18.4	14.68	71.1	47.7	1.00	0.60	0.71	0.24
12.4	18.6	14.50	73.3	50.4	1.20	0.74	0.86	0.19
				Testing cases, $\mathbf{x}_{\text{new}}$				
2.8	8.4	5.45	129.7	82.7	0.91	0.18	0.21	0.30
7.1	12.1	8.68	93.0	63.0	1.04	0.39	0.44	0.22
11.5	18.6	14.45	73.3	50.9	1.02	0.67	0.78	0.21

Source: Data from D'Ignazio et al. (2016).

cases and three testing cases. The training cases were used to train the site-specific model, whereas the properties for the testing cases were to be predicted by the trained model. To illustrate the behaviors of the proposed method, the following scenarios for the training cases were considered:

1. No data: the training cases in Table 2 are not used. This no data scenario may not be realistic. However, it was adopted here to illustrate the behavior of the site-specific model purely based on the noninformative prior.
2. Two data points: two training cases at depths 5 and 9.1 m are used.
3. Five data points: five training cases at depths 2.1, 5, 7.9, 9.1, and 12.4 m are used.
4. Abundant data: all 13 training cases are used.

For all four scenarios, the effect of hybridization will be illustrated. For all scenarios, it was assumed that the depth intervals for the training data were larger than the vertical scale of fluctuation so that the data at different depths can be assumed independent.

To demonstrate the behaviors of the trained site-specific model, consider a new depth in the same clay layer at the Lilla Mellösa site with (transformed) property $\mathbf{x}_{\text{new}}$. The GS samples for $\mathbf{x}_{\text{new}}$ can be readily obtained and converted to $\mathbf{y}_{\text{new}}$ samples (e.g., LL, PI, LI, σ'_v/P_a, σ'_p/P_a, and s_u/σ'_v) through the inverse transform of Eq. (4). It was found that GS samples with $T = 20{,}000$ can produce stable statistics (e.g., median, confidence interval, and correlation). The total sampling step size was taken to be $T = 20{,}000$, and the end of the burn-in period was determined to be $t_b = 1{,}000$. Fig. 3 shows the marginal cumulative density function (CDF) for the resulting LL samples. Observations for other variables, i.e., the marginal CDFs for PI, LI, σ'_v/P_a, σ'_p/P_a, s_u/σ'_v, were similar, so only the marginal CDF for LL is illustrated. In Fig. 3, the marginal CDFs based on $f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g)$, $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$, and $f(\mathbf{x}_{\text{new}}|hb)$ are referred to as generic CDF, site-specific CDF, and hybrid CDF, respectively. The empirical CDF was based on the training data. In essence, $f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ reflected the data distribution of CLAY/10/7490, $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ was trained by the training data, and $f(\mathbf{x}_{\text{new}}|hb)$ was the hybridization between $f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ and $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$. In Fig. 3(a), there are no training data, so the site-specific CDF is governed by the noninformative prior. The hybrid CDF converges to the generic CDF. In Fig. 3(d), there are abundant training data, and the site-specific CDF contains much information and follows the empirical CDF closely. The hybrid CDF converges to the site-specific CDF. Figs. 3(b and c) are intermediate cases where the hybrid CDF is the trade-off between the generic and site-specific CDFs.

Fig. 3 only shows the marginal distributions for $\mathbf{y}_{\text{new}}$ samples. Fig. 4 shows the correlation behaviors among $\mathbf{y}_{\text{new}}$ sample pairs. Conclusions obtained from various $\mathbf{y}_{\text{new}}$ sample pairs are similar, so only the LI-σ'_p (X_3 versus X_5) sample pair is illustrated. In Fig. 4(a), there are no training data, and the LI-σ'_p samples from $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ spread widely due to the extremely large uncertainty in the noninformative prior, and the samples from $f(\mathbf{x}_{\text{new}}|hb)$ resemble those from $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$. In Fig. 4(d), there are abundant training data, and the LI-σ'_p samples from $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ spread narrowly, and the samples from $f(\mathbf{x}_{\text{new}}|hb)$ resemble those from $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$. Figs. 4(b and c) are intermediate cases.

During GS, the site-specific covariance matrix $\mathbf{C}_s$ is sampled. From each $\mathbf{C}_s$ sample, a sample for the correlation coefficient between (X_3, X_5), denoted by $\delta_{s,35}$, can be extracted. Note that $\delta_{s,35}$ represents the site-specific correlation between (X_3, X_5). Figs. 5(a–d) show the histograms for $\delta_{s,35}$. When there are no training data [Fig. 5(a)], the histogram is flat between $[-1, 1]$ because $\delta_{s,35}$ is governed by the noninformative prior. When there are abundant training data [Fig. 5(d)], the histogram becomes peaked. Figs. 5(b and c) are intermediate cases. For comparison, the generic correlation between (X_3, X_5) exhibited in CLAY/10/7490, denoted by $\delta_{g,35}$, is -0.57 (Fig. 1) and is shown as vertical lines in Figs. 5(a–d). It is remarkable that $\delta_{s,35}$ can take a value that is very different from $\delta_{g,35}$. In general, there seems to be no clear connection between a site-specific correlation and a generic correlation: δ_s can be greater or smaller than δ_g or even take a different sign, exemplified by Fig. 5(d) where $\delta_{s,35}$ is very likely to be positive but $\delta_{g,35}$ is negative. This suggests that the conventional Bayesian analysis [i.e., multiplying the prior $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ and the likelihood $f(\mathbf{X}^o|\mathbf{x}_{\text{new}}, \boldsymbol{\mu}_g, \mathbf{C}_g)$ to produce the posterior $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o, \boldsymbol{\mu}_g, \mathbf{C}_g)$] is not appropriate because it assumes both site-specific data $\mathbf{X}^o$ and $\mathbf{x}_{\text{new}}$ have the same mean and covariance ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$).

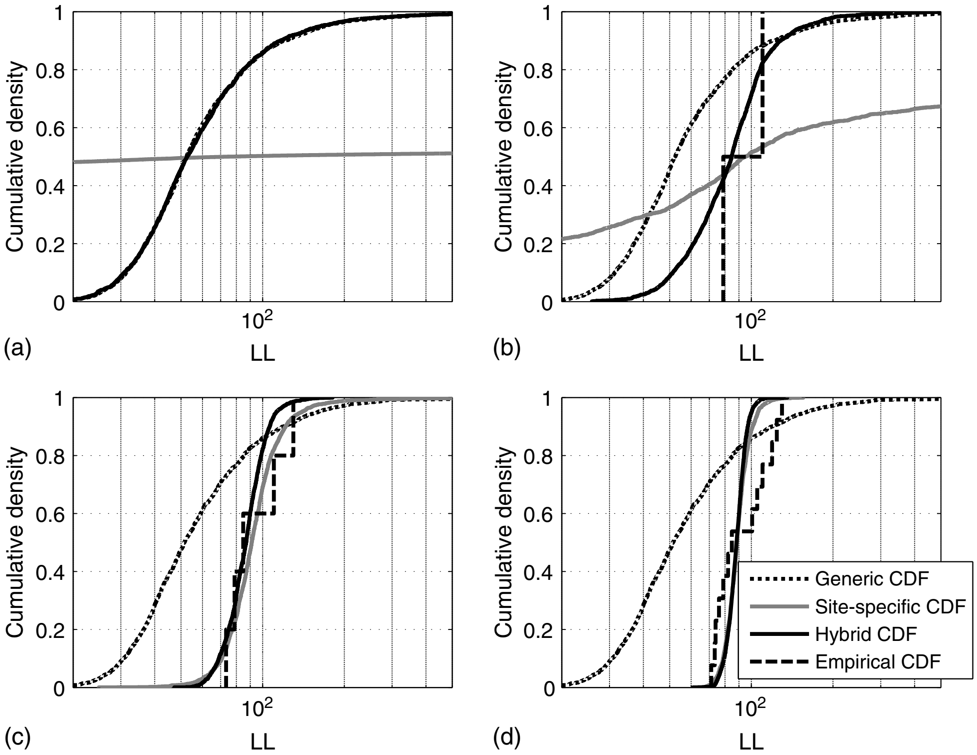


Fig. 3. Marginal CDFs for the LL samples (Lilla Mellösa site): (a) no data; (b) two data points; (c) five data points; and (d) abundant data.

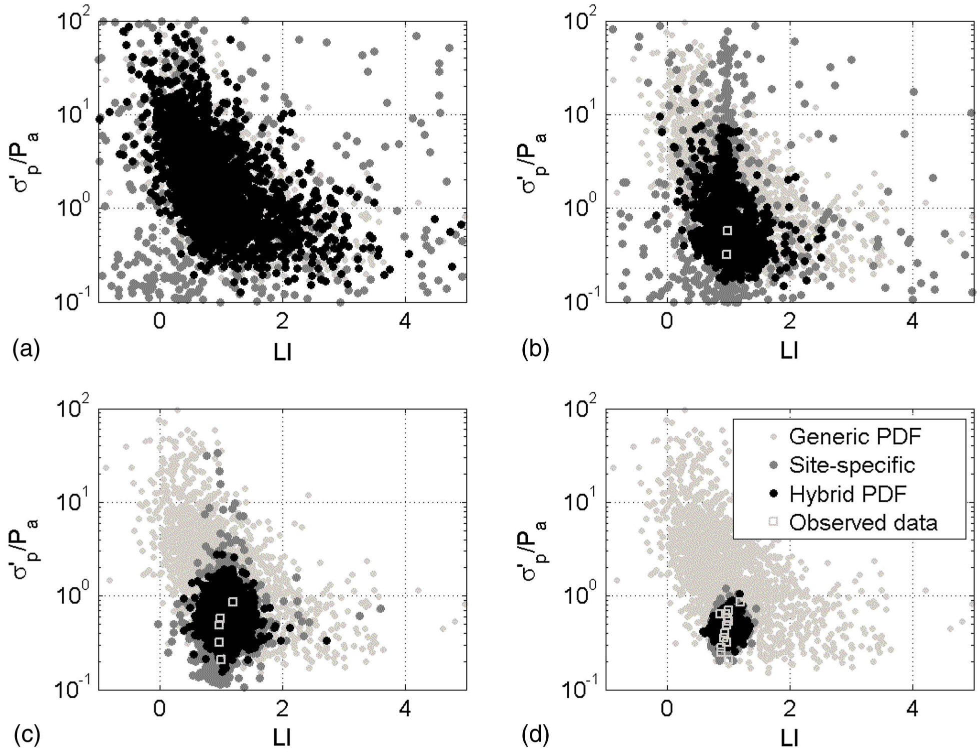


Fig. 4. Correlation behavior for LI-σ'_p samples (Lilla Mellösa site): (a) no data; (b) two data points; (c) five data points; and (d) abundant data.

The preceding demonstration is for the behaviors of the trained model. Now let us consider the prediction for s_u(mob) at a new depth. Consider the testing cases at three new depths in Table 2. Suppose that (LL, PI, LI, σ'_v, σ'_p) at these new depths are known (i.e., they are $\mathbf{x}^o_{\text{new}}$), and the purpose is to predict the s_u(mob) values at these depths. For each new depth, its $(Y_1, Y_2, \ldots, Y_5)$ can be computed from its (LL, PI, LI, σ'_v, σ'_p) and can be subsequently converted to $\mathbf{x}^o_{\text{new}} = (X_1, X_2, \ldots, X_5)$ through Eq. (4).

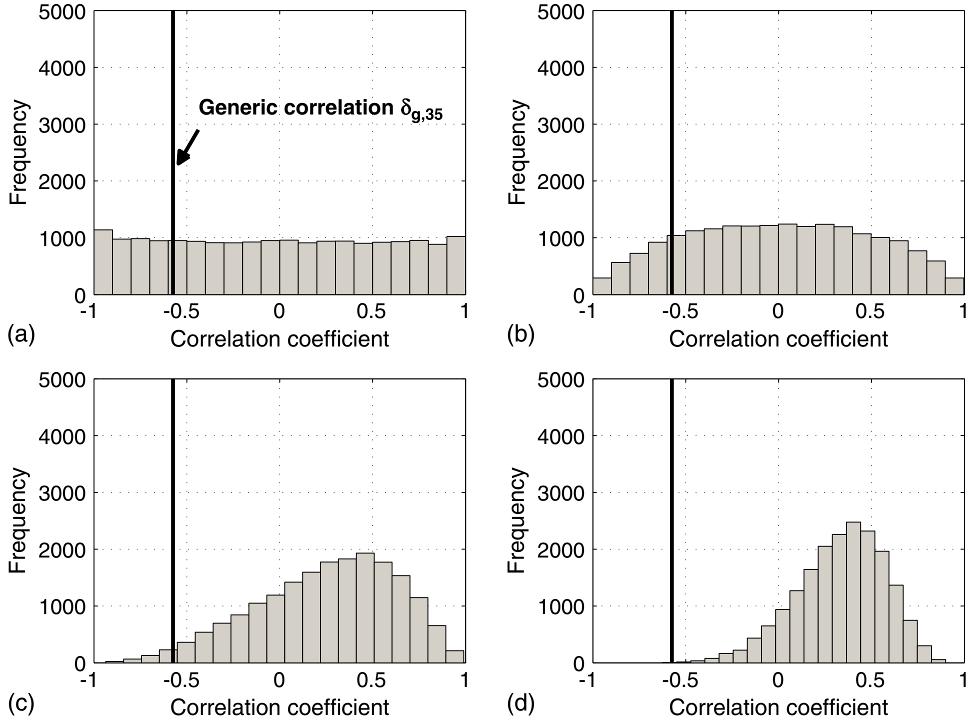


Fig. 5. Site-specific correlation coefficient $\delta_{s,35}$ (Lilla Mellösa site): (a) no data; (b) two data points; (c) five data points; and (d) abundant data.

Conditioning on the $\mathbf{x}^o_{\mathrm{new}}$ for each depth, samples for $\mathbf{x}^u_{\mathrm{new}} = X_6$ can be drawn from $f(\mathbf{x}^u_{\mathrm{new}}|\mathbf{X}^o, \mathbf{x}^o_{\mathrm{new}})$ without hybridization or from $f(\mathbf{x}^u_{\mathrm{new}}|hb, \mathbf{x}^o_{\mathrm{new}})$ with hybridization. The samples of X_6 are then converted to $Y_6 = \ln(s_u/\sigma'_v)$ samples through the inverse transform of Eq. (4), and subsequently s_u(mob) samples can be readily obtained. Finally, the sample median value and sample 95% CI for s_u(mob) can be obtained for each new depth for the purpose of prediction. The actual s_u(mob) values are known at the three new depths, so genuine validation between the prediction results and the actual s_u(mob) can be made.

Figs. 6(a–d) show the medians and 95% CIs based on $f(\mathbf{x}^u_{\mathrm{new}}|\mathbf{X}^o, \mathbf{x}^o_{\mathrm{new}})$, i.e., no hybridization. When there are no data or two data points, s_u(mob) sample values can become absurd (e.g., close to 0 $\mathrm{kN/m^2}$ or more than 1,000 $\mathrm{kN/m^2}$). The resulting 95% CIs are extremely wide, so the prediction results are not plotted in Figs. 6(a and b). Figs. 6(c and d) show that the 95% CIs become narrower when there are more data points. To prevent extremely wide CIs when data are sparse, hybridization can be adopted. By doing so, s_u(mob) samples rely more on the generic PDF when data are sparse, so s_u(mob) sample values are no longer absurd. We submit that adopting a generic experience by hybridization is reasonable if local experience is absent or very sparse. This is in line with current standard practice where a desk study is integral to site investigation. Figs. 6(e–h) show the results based on $f(\mathbf{x}^u_{\mathrm{new}}|hb, \mathbf{x}^o_{\mathrm{new}})$, i.e., hybridization is adopted. Fig. 6(e) shows the prediction results for the no data scenario but now with hybridization. Compared with the extremely wide CIs in Fig. 6(a), the 95% CIs in Fig. 6(e) are now more reasonable. However, the 95% CIs in Fig. 6(e) are wide because the generic PDF is constructed by generic data points so it has to accommodate a wide range of scenarios. Figs. 6(e–h) further show that extra site-specific training data are very effective in reducing the CI size. With only two data points together with hybridization, the CI size is significantly reduced [Fig. 6(f)]. The CI size is further reduced when more training data are available [Figs. 6(g and h)].

By hybridization, the CI size is typically reduced. By comparing between Figs. 6(a and e) or Figs. 6(b and f), it is evident that hybridization reduces the CI size significantly when there are no training data or two training data points. By comparing between Figs. 6(c and g) or Figs. 6(d and h), hybridization reduces the CI size slightly.

Taipei, Taiwan, Case

Now consider the silty clay layer for the Taipei, Taiwan, site (Table 1). There are missing entries in the table: (Y_1, Y_2, Y_3) are missing at one depth; Y_6 is missing at six depths; (Y_7, Y_8, Y_{10}) are completely missing. All available training data in the table were adopted as $\mathbf{X}^o$ to train the site-specific model. It was assumed that the depth intervals for the training data were larger than the vertical scale of fluctuation so that the data at different depths can be assumed independent. At the initial stage of GS, the missing entries in Table 1 (i.e., $\mathbf{X}^u$) were set to be zeros. For GS, the total sampling step size was taken to be $T = 20{,}000$, and the end of the burn-in period was determined to be $t_b = 1{,}000$. During GS, $\mathbf{X}^u$ were randomly sampled using Eq. (19) as one step of an iterative GS loop that cycles from Eq. (16) to Eq. (19). Again, two key results are presented subsequently: (1) behaviors of the trained model, and (2) prediction results for s_u(mob).

Regarding the behaviors of the trained model, Fig. 7 shows the marginal CDFs for the resulting $\mathbf{y}_{\mathrm{new}}$ samples. The site-specific CDF usually closely follows the empirical CDF, except when there are very few data [Fig. 7(e)]. The hybrid CDF is the trade-off between the generic and site-specific CDFs. Fig. 8 shows the correlation behaviors among some $\mathbf{y}_{\mathrm{new}}$ sample pairs. The observations obtained in Fig. 8 are similar to those obtained in Fig. 5. Regarding the prediction for s_u(mob), suppose that LL, PL, w_n (nature water content), σ_v, σ'_v, σ'_p, and q_c (cone resistance) profiles are known by interpolation in the silty clay layer at this Taipei, Taiwan, site [Figs. 9(a–c)], and the purpose is to predict the s_u(mob) profile. Figs. 9(d and e) show the prediction results. The actual s_u(mob)

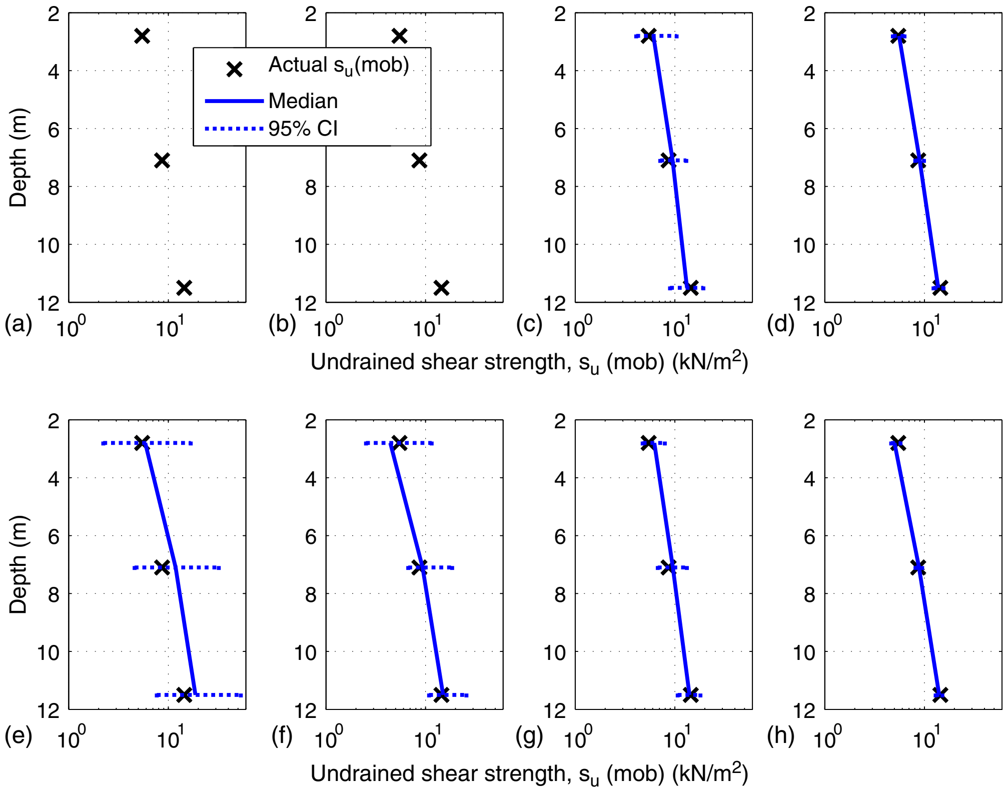


Fig. 6. Prediction results for s(mob) (Lilla Mellösa site) based on $f(\mathbf{x}^u_{\text{new}}|\mathbf{X}^o, \mathbf{x}^o_{\text{new}})$: (a) no data; (b) two data points; (c) five data points; and (d) abundant data. Prediction results based on $f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}})$: (e) no data; (f) two data points; (g) five data points; and (h) abundant data.

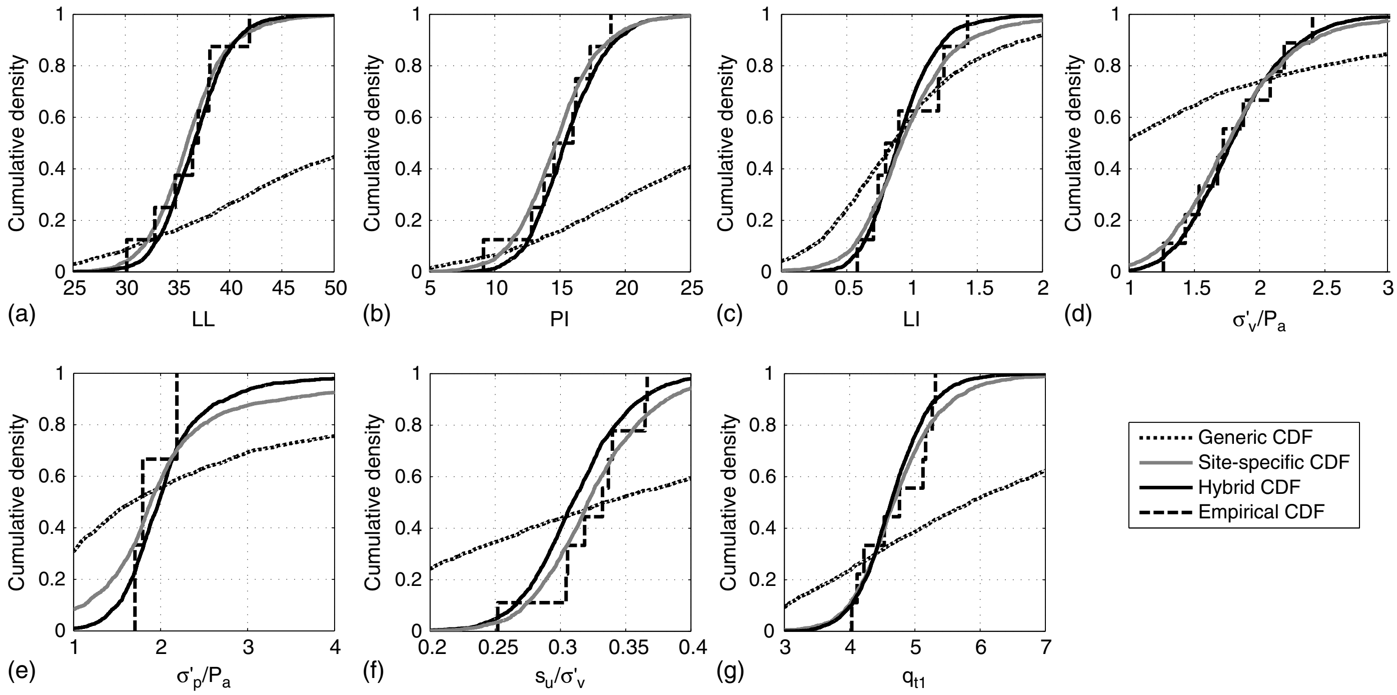


Fig. 7. Marginal CDFs for the $\mathbf{y}_{\text{new}}$ samples (Taipei, Taiwan, site).

data points in these figures are the s_u(mob) values in the training data. They are plotted just for reference, not for genuine validation. The 95% CIs are narrow, and the median profile fits well to the actual s_u(mob) data. By comparing between Figs. 9(d and e), it is evident that hybridization reduces the CI size slightly. The proposed method can only predict the marginal statistics (median and 95% CI) of s_u(mob) at each depth. The proposed method does not address the spatial correlation between different depths.

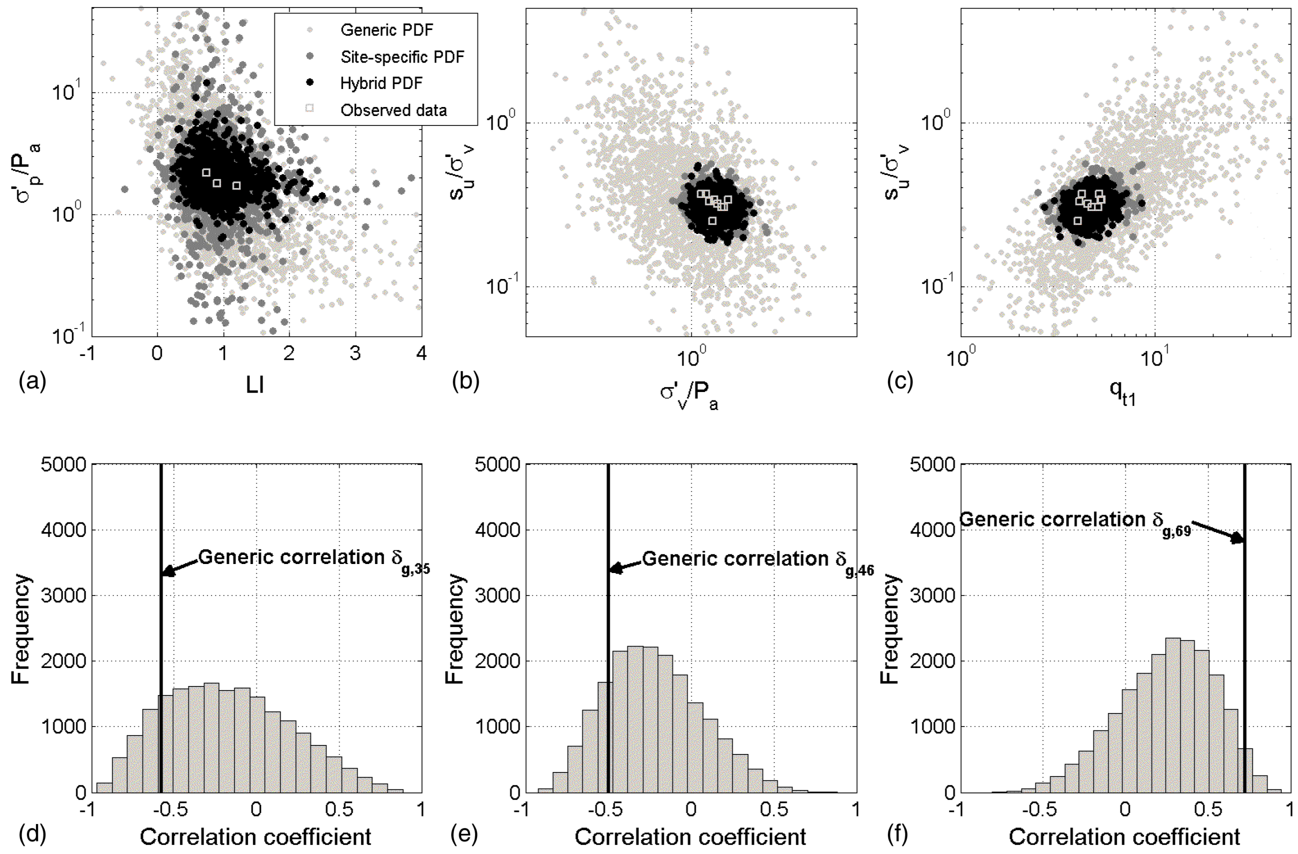


Fig. 8. Correlation behaviors among some $\mathbf{y}_{\text{new}}$ sample pairs (Taipei, Taiwan, site).

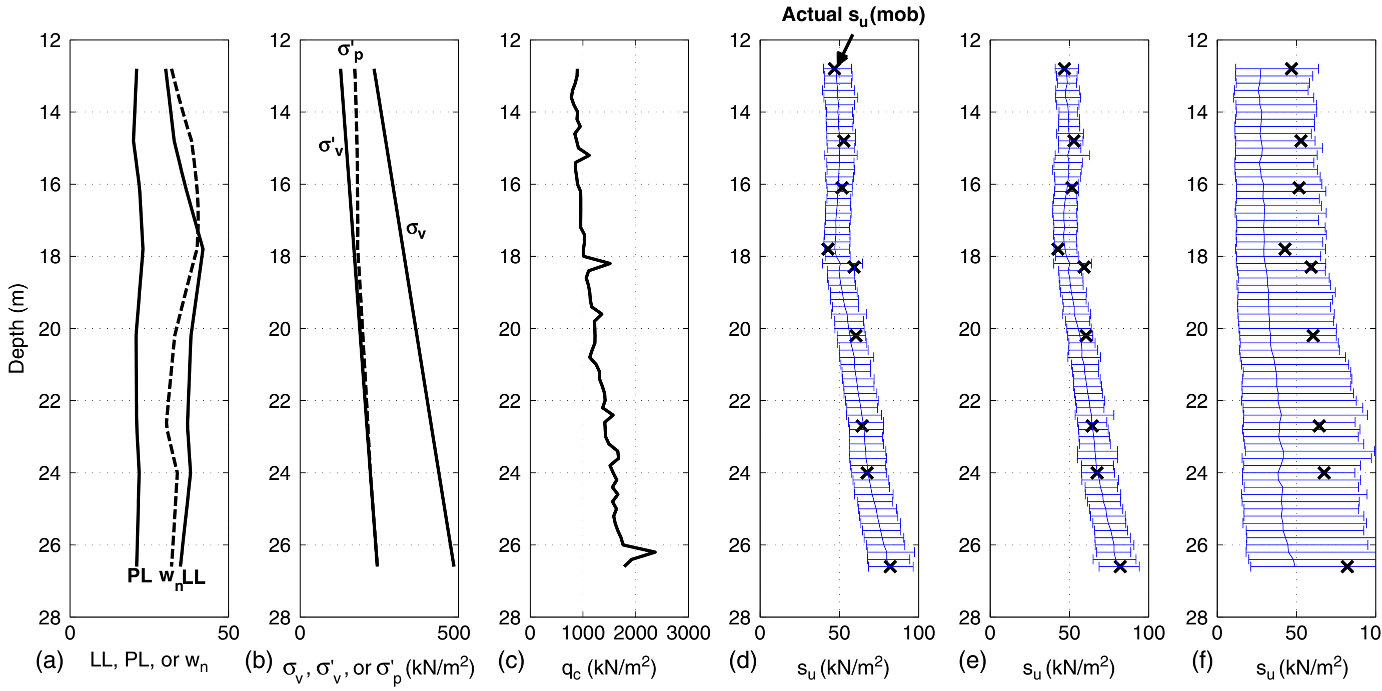


Fig. 9. Clay property profiles: (a) LL, PL, and w_n profiles; (b) σ_v, σ'_v, and σ'_p profiles; (c) q_c profile, as well as prediction results for s_u(mob) based on (d) $f(\mathbf{x}^u_{\text{new}}|\mathbf{X}^o, \mathbf{x}^o_{\text{new}})$; (e) $f(\mathbf{x}^u_{\text{new}}|hb, \mathbf{x}^o_{\text{new}})$; and (f) conventional Bayesian analysis, $f(\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g, \mathbf{x}^o_{\text{new}})$.

Discussion

The analysis results show that the trained model can correctly capture the correlation behaviors in the site-specific training data $\mathbf{X}^o$. When $\mathbf{X}^o$ is very sparse (e.g., no data or two data points in the Lilla Mellösa case), the trained model $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ has extremely large statistical uncertainty, both in marginal PDFs and in correlation coefficients, which is anticipated because there is very little information. When $\mathbf{X}^o$ is abundant, the trained model $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$ has less statistical uncertainty, which is also anticipated. The main value for hybridization is in the former case (very sparse $\mathbf{X}^o$). It is shown in the case studies that hybridization is necessary to produce meaningful results when $\mathbf{X}^o$ is very sparse. In general, hybridization tends to further reduce uncertainty. The uncertainty reduction is significant when $\mathbf{X}^o$ is very sparse and less significant when $\mathbf{X}^o$ is abundant.

Comparison with Conventional Bayesian Analysis

Previously, Ching and Phoon (2014b) and Ching et al. (2017a, 2018a) compiled generic soil and rock databases and developed the generic mean and covariance ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$) based on the databases and subsequently the generic PDF $f(\mathbf{x}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ (Ching and Phoon 2014a; Ching et al. 2017b, 2018b). With the generic PDF, they focused on the simulation of $\mathbf{x}_{\text{new}}$ based on the generic PDF $f(\mathbf{x}_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g)$ as well as the prediction of $\mathbf{x}^u_{\text{new}}$ based on the conditional generic PDF $f(\mathbf{x}^u_{\text{new}}|\boldsymbol{\mu}_g, \mathbf{C}_g, \mathbf{x}^o_{\text{new}})$. This method is called the conventional Bayesian analysis in the current paper. There is no notion of site-specific data or site-specific mean and covariance ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) in the conventional Bayesian analysis. In other words, $\mathbf{x}_{\text{new}}$ is assumed to be from the generic population and have mean and covariance ($\boldsymbol{\mu}_g$, $\mathbf{C}_g$). The prediction results for the Taipei, Taiwan, case based on the conventional Bayesian analysis is plotted in Fig. 9(f). The key distinction between the current study and the conventional Bayesian analysis is that there is now a notion of site-specific data $\mathbf{X}^o$ and site-specific mean and covariance ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$). In other words, $\mathbf{x}_{\text{new}}$ is now assumed to be from the local site and have mean and covariance ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$). Moreover, ($\boldsymbol{\mu}_s$, $\mathbf{C}_s$) are uncertain and to be updated by $\mathbf{X}^o$. The site-specific PDF of $\mathbf{x}_{\text{new}}$ is now $f(\mathbf{x}_{\text{new}}|\mathbf{X}^o)$. A local site typically has less transformation uncertainty than the generic population. For the Taipei case, this can be seen by comparing Fig. 9(d) with Fig. 9(f). The 95% CIs in Fig. 9(d) are narrower than those in Fig. 9(f) because the local transformation uncertainty is less than the generic one. On the other hand, if $\mathbf{X}^o$ is sparse, a local site can have large statistical uncertainty. Hybridization is proposed in the current study so that the hybrid PDF converges to the generic PDF when $\mathbf{X}^o$ is sparse: the results in Fig. 6(e) converge to those based on the generic PDF.

Conclusion

This study proposes a novel method of constructing a site-specific multivariate probability distribution model for soil properties. It has an interesting feature that when site-specific data are abundant, the method is governed by site-specific data, and when site-specific data are sparse, the method is governed by a non-site-specific (or generic) database. The proposed method allows incomplete multivariate inputs. It can rigorously quantify transformation and statistical uncertainties. The applicability of the proposed method is independent of the nature of the data. It can be used to construct the site-specific model for clays, sands, or rocks: namely, it is a machine learning framework that is purely driven by data.

Real case studies were used to demonstrate the usefulness of the proposed method. Analysis results showed that the proposed method can effectively capture the correlation behaviors in site-specific data and, moreover, can make reasonable predictions even when site-specific data are very sparse. There is a key step of hybridizing between generic and site-specific data in the proposed method. It is recommended that when site-specific data are very sparse, this hybridization must be adopted in order to produce reasonable prediction results. The proposed method can only predict the marginal statistics (median and 95% CI) of a soil parameter at each depth. The proposed method does not address the spatial correlation between different depths.

Acknowledgments

The authors would like to thank the members of the TC304 Committee on Engineering Practice of Risk Assessment and Management of the International Society of Soil Mechanics and Geotechnical Engineering for developing Database 304dB (ISSMGE 2018) used in this study and making it available for scientific inquiry.

References

Alvarez, I., J. Niemi, and M. Simpson. 2014. “Bayesian inference for a covariance matrix.” In *Proc., 26th Annual Conf. on Applied Statistics in Agriculture*. Manhattan, KS: Kansas State Univ.

Beck, J. L., and K. V. Yuen. 2004. “Model selection using response measurements: Bayesian probabilistic approach.” *J. Eng. Mech.* 130 (2): 192–203. https://doi.org/10.1061/(ASCE)0733-9399(2004)130:2(192).

Cao, Z., and Y. Wang. 2014. “Bayesian model comparison and characterization of undrained shear strength.” *J. Geotech. Geoenviron. Eng.* 140 (6): 04014018. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001108.

Cao, Z., Y. Wang, and D. Li. 2016. “Quantification of prior knowledge in geotechnical site characterization.” *Eng. Geol.* 203: 107–116. https://doi.org/10.1016/j.enggeo.2015.08.018.

CEN (European Committee for Standardization). 2004. *Eurocode 7: Geotechnical design—Part 1: General rules*. EN 1997-1. Brussels, Belgium: CEN.

Ching, J. 2018. “What does the soil parameter estimated from a transformation model really mean?” *J. GeoEng.* 13 (3): 105–113. http://dx.doi.org/10.6310/jog.201809_13(3).2.

Ching, J., K. H. Li, K. K. Phoon, and M. C. Weng. 2018a. “Generic transformation models for some intact rock properties.” *Can. Geotech. J.* in press. https://doi.org/10.1139/cgj-2017-0537.

Ching, J., G. H. Lin, J. R. Chen, and K. K. Phoon. 2017a. “Transformation models for effective friction angle and relative density calibrated based on a multivariate database of coarse-grained soils.” *Can. Geotech. J.* 54 (4): 481–501. https://doi.org/10.1139/cgj-2016-0318.

Ching, J., G. H. Lin, K. K. Phoon, and J. R. Chen. 2017b. “Correlations among some parameters of coarse-grained soils—The multivariate probability distribution model.” *Can. Geotech. J.* 54 (9): 1203–1220. https://doi.org/10.1139/cgj-2016-0571.

Ching, J., and K. K. Phoon. 2012. “Value of geotechnical site investigation in reliability-based design.” *Adv. Struct. Eng.* 15 (11): 1935–1945. https://doi.org/10.1260/1369-4332.15.11.1935.

Ching, J., and K. K. Phoon. 2014a. “Correlations among some clay parameters—The multivariate distribution.” *Can. Geotech. J.* 51 (6): 686–704. https://doi.org/10.1139/cgj-2013-0353.

Ching, J., and K. K. Phoon. 2014b. “Transformations and correlations among some parameters of clays: The global database.” *Can. Geotech. J.* 51 (6): 663–685. https://doi.org/10.1139/cgj-2013-0262.

Ching, J., K. K. Phoon, K. H. Li, and M. C. Weng. 2018b. “Multivariate probability distribution for some intact rock properties.” *Can. Geotech. J.* in press.

Ching, J., K. K. Phoon, and J. W. Yu. 2014. “Linking site investigation efforts to final design savings with simplified reliability-based design methods.” *J. Geotech. Geoenviron. Eng.* 140 (3): 04013032. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001049.

Ching, J., K.-K. Phoon, and Y.-C. Chen. 2010. “Reducing shear strength uncertainties in clays by multivariate correlations.” *Can. Geotech. J.* 47 (1): 16–33. https://doi.org/10.1139/T09-074.

D’Ignazio, M., K. K. Phoon, S. A. Tan, and T. Lansivaara. 2016. “Correlations for undrained shear strength of Finnish soft clays.” *Can. Geotech. J.* 53 (10): 1628–1645. https://doi.org/10.1139/cgj-2016-0037.

Djoenaidi, W. J. 1985. "A compendium of soil properties and correlations." M.Eng. thesis, School of Civil and Mining Engineering, Univ. of Sydney.

Doucet, A., N. de Freitas, and N. Gordon. 2001. *Sequential Monte Carlo methods in practice*. New York: Springer.

Feng, X., and R. Jimenez. 2014. "Bayesian prediction of elastic modulus of intact rocks using their uniaxial compressive strength." *Eng. Geol.* 173: 32–40. https://doi.org/10.1016/j.enggeo.2014.02.005.

Feng, X., and R. Jimenez. 2015. "Estimation of deformation modulus of rock masses based on Bayesian model selection and Bayesian updating approach." *Eng. Geol.* 199: 19–27. https://doi.org/10.1016/j.enggeo.2015.10.002.

Gelman, A. 2006. "Prior distributions for variance parameters in hierarchical models." *Bayesian Anal.* 1 (3): 515–534. https://doi.org/10.1214/06-BA117A.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian data analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Geman, S., and D. Geman. 1984. "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images." *IEEE Trans. Pattern Anal. Machine Intell.* 6 (6): 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Gilks, W. R., D. J. Spiegelhalter, and S. Richardson. 1996. *Markov chain Monte Carlo in practice*. London: Chapman and Hill.

Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57 (1): 97–109. https://doi.org/10.1093/biomet/57.1.97.

Heckerman, D., D. Geiger, and D. M. Chickering. 1995. "Learning Bayesian networks: The combination of knowledge and statistical data." *Mach. Learn.* 20 (3): 197–243. https://doi.org/10.1023/A:1022623210503.

Huang, A., and M. P. Wand. 2013. "Simple marginally noninformative prior distributions for covariance matrices." *Bayesian. Anal.* 8 (2): 439–452. https://doi.org/10.1214/13-BA815.

ISSMGE (International Society of Soil Mechanics and Geotechnical Engineering). 2018. "Database 304dB." TC304 Engineering Practice of Risk Assessment and Management. Accessed October 6, 2017. http://140.112.12.21/issmge/tc304.htm?=6.

James, A. 1964. "Distributions of matrix variates and latent roots derived from normal samples." *Ann. Math. Stat.* 35 (2): 475–501. https://doi.org/10.1214/aoms/1177703550.

Johnson, N. L. 1949. "Systems of frequency curves generated by methods of translation." *Biometrika* 36 (1/2): 149–176. https://doi.org/10.2307/2332539.

Kulhawy, F. H. and P. W. Mayne. 1990. *Manual on estimating soil properties for foundation design*. Rep. No. EL6800. Palo Alto, CA: Electric Power Research Institute.

Li, D. Q., S. B. Wu, C. B. Zhou, and K. K. Phoon. 2012. "Performance of translation approach for modeling correlated non-normal variables." *Struct. Saf.* 39 (2): 52–61. https://doi.org/10.1016/j.strusafe.2012.08.001.

Liu, P. L., and A. Der Kiureghian. 1986. "Multivariate distribution models with prescribed marginals and covariances." *Probab. Eng. Mech.* 1 (2): 105–112. https://doi.org/10.1016/0266-8920(86)90033-0.

Liu, S., H. Zou, G. Cai, B. V. Bheemasetti, A. J. Puppala, and J. Lin. 2016. "Multivariate correlation among resilient modulus and cone penetration test parameters of cohesive subgrade soils." *Eng. Geol.* 209: 128–142. https://doi.org/10.1016/j.enggeo.2016.05.018.

MacKay, D. J. C. 1995. "Probable networks and plausible predictions: A review of practical Bayesian methods for supervised neural networks." *Network: Comput. Neural Syst.* 6 (3): 469–505. https://doi.org/10.1088/0954-898X_6_3_011.

MacKay, D. J. C. 1998. "Introduction to Monte Carlo methods." In *Learning in graphical models*, edited by M. Jordan. Cambridge, MA: MIT Press.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate analysis*. London: Academic Press.

Mayne, P. W., Christopher, B. R., and DeJong, J. 2001. *Manual on subsurface investigations*. Publication No. FHWA NHI-01-031. Washington, DC: Federal Highway Administration.

Mesri, G., and N. Huvaj. 2007. "Shear strength mobilized in undrained failure of soft clay and silt deposits." In *Advances in measurement and modeling of soil behavior (GSP 173)*, edited by D. J. DeGroot, C. Vipulanandan, J. A. Yamamuro, V. N. Kaliakin, P. V. Lade, M. Zeghal, U. El Shamy, N. Lu, and C. R. Song, 1–22. Reston, VA: ASCE.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. "Equation of state calculations by fast computing machines." *J. Chem. Phys.* 21 (6): 1087–1092. https://doi.org/10.1063/1.1699114.

Ng, I. T., K. V. Yuen, and L. Dong. 2017. "Estimation of undrained shear strength in moderately OC clays based on field vane test data." *Acta Geotech.* 12 (1): 145–156. https://doi.org/10.1007/s11440-016-0433-0.

Ng, I. T., K. V. Yuen, and C. H. Lau. 2015. "Predictive model for uniaxial compressive strength for Grade III granitic rocks from Macao." *Eng. Geol.* 199: 28–37. https://doi.org/10.1016/j.enggeo.2015.10.008.

Ou, C. Y., and J. T. Liao. 1987. *Geotechnical engineering research report*. GT96008. Taipei, Taiwan: National Taiwan Univ. of Science and Technology.

Phoon, K. K., and F. H. Kulhawy. 1999a. "Characterization of geotechnical variability." *Can. Geotech. J.* 36 (4): 612–624. https://doi.org/10.1139/t99-038.

Phoon, K. K., and F. H. Kulhawy. 1999b. "Evaluation of geotechnical variability." *Can. Geotech. J.* 36 (4): 625–639. https://doi.org/10.1139/t99-039.

Rasmussen, C. E., and C. K. Williams. 2006. *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Tan, T. S., K. K. Phoon, D. W. Hight, and S. Leroueil. 2003a. Vol. 1 of *Characterisation and engineering properties of natural soils*. Rotterdam, Netherlands: A.A. Balkema.

Tan, T. S., K. K. Phoon, D. W. Hight, and S. Leroueil. 2003b. Vol. 2 of *Characterisation and engineering properties of natural soils*. Rotterdam, Netherlands: A.A. Balkema.

Tan, T. S., K. K. Phoon, D. W. Hight, and S. Leroueil. 2006a. Vol. 3 of *Characterisation and engineering properties of natural soils*. London: Taylor & Francis.

Tan, T. S., K. K. Phoon, D. W. Hight, and S. Leroueil. 2006b. Vol. 4 of *Characterisation and engineering properties of natural soils*. London: Taylor & Francis.

Tipping, M. E. 2001. "Sparse Bayesian learning and the relevance vector machine." *J. Mach. Learn. Res.* 1: 211–244.

Tokuda, T., B. Goodrich, I. Van Mechelen, and A. Gelman. 2011. "Visualizing distributions of covariance matrices." Accessed August 3, 2017. http://www.stat.columbia.edu/~gelman/research/unpublished/Visualization.pdf.

Vanmarcke, E. H. 1977. "Probabilistic modeling of soil profiles." *J. Geotech. Eng.* 103 (11): 1227–1246.

Wang, H., and D. Y. Yeung. 2016. "Towards Bayesian deep learning: A framework and some existing methods." *IEEE Trans. Knowl. Data Eng.* 28 (12): 3395–3408. https://doi.org/10.1109/TKDE.2016.2606428.

Wang, Y., and O. V. Akeju. 2016. "Quantifying the cross-correlation between effective cohesion and friction angle of soil from limited site-specific data." *Soils Found.* 56 (6): 1055–1070. https://doi.org/10.1016/j.sandf.2016.11.009.

Wang, Y., and A. E. Aladejare. 2015. "Selection of site-specific regression model for characterization of uniaxial compressive strength of rock." *Int. J. Rock Mech. Min. Sci.* 75: 73–81. https://doi.org/10.1016/j.ijrmms.2015.01.008.

Wang, Y., and A. E. Aladejare. 2016. "Bayesian characterization of correlation between uniaxial compressive strength and Young's modulus of rock." *Int. J. Rock Mech. Min. Sci.* 85: 10–19. https://doi.org/10.1016/j.ijrmms.2016.02.010.

Wang, Y., and Z. J. Cao. 2013. "Probabilistic characterization of Young's modulus of soil using equivalent samples." *Eng. Geol.* 159: 106–118. https://doi.org/10.1016/j.enggeo.2013.03.017.

Yan, W. M., K. V. Yuen, and G. L. Yoon. 2009. "Bayesian probabilistic approach for the correlations of compression index for marine clays." *J. Geotech. Geoenviron. Eng.* 135 (12): 1932–1940. https://doi.org/10.1061/(ASCE)GT.1943-5606.0000157.

Yuen, K. V. 2010. "Recent developments of Bayesian model class selection and applications in civil engineering." *Struct. Saf.* 32 (5): 338–346. https://doi.org/10.1016/j.strusafe.2010.03.011.