Role of generic soil database in site-specific soil property estimation

Jianye Ching[1] and K.K. Phoon[2]

[1] Depart. of Civil Engineering, National Taiwan Univ., #1 Roosevelt Road Sect. 4, Taipei 10617, Taiwan
[2] Depart. of Civil & Environ. Engineering, National Univ. of Singapore, Blk E1A, #07-03, 1 Engineering Drive 2, Singapore 117576

ABSTRACT

Geotechnical estimation is site-specific due to the significant inherent variability in the soil properties. Soil data from one site may not be applicable to another site. Ideally, site-specific soil property estimation should rely on site-specific data rather than generic data obtained from other sites. However, site-specific data are often sparse and insufficient to support site-specific estimation. In practice, non-site-specific (generic) soil data is widely used to estimate soil properties in one site by appealing to similarity in soil behavior primarily based on the judgment of the engineer. This paper focuses on the possible role of generic databases in site-specific soil property estimation. It is proposed that the site-specific data can be combined with generic data in a suitable way to support site-specific estimation using Bayesian machine learning. The research outcomes are demonstrated through the problem of estimating the soil properties at a target site, in which a generic database previously compiled by the authors is adopted to support site-specific soil property estimation.

Keywords: soil properties; generic database; site-specific estimation; site characterization; data analysis

1 INTRODUCTION

In geotechnical engineering, there are two facts that are in contrast to each other. On one hand, when it comes to a local site, site-specific data are sparse. For a typical site investigation program, a few boreholes and cone penetration test (CPT) soundings may be conducted. For the boreholes, only a limited number of depths are measured (sparsity in the vertical direction), whereas for both boreholes and CPT soundings, only a limited number of horizontal locations are measured (sparsity in the horizontal direction). The volume of the investigated soil mass is very small compared with the total soil volume mobilized by the actual structure. Also, only a limited number of soil samples is available to verify the correlation among different soil properties, e.g., Atterberg limits, SPT N, preconsolidation stress, undrained shear strength, etc. are simultaneously measured at close proximity, and this multivariate information is rarely complete (complete means that all soil properties are simultaneously measured in a particular location and at a particular depth). It is fair to say that we are in a data-poor scenario for a local site, both in space and in correlation. Phoon (2018) described this situation as MUSIC: site-specific data are Multivariate, Uncertain and Unique, Sparse, and InComplete. Geotechnical engineers routinely need to make decisions under MUSIC site-specific data. Phoon et al. (2019) venture to suggest that MUSIC can be re-interpreted to cover extremes: Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted. The screening for extremes or outliers is clearly important, but not covered in this paper. Ching et al. (2019) proposed a simple chi-square approach to identify outliers in a rock property database, but this approach does not work for MUSIC data.

On the other hand, it is widely known that generic (non-site-specific) data are abundant. Phoon et al. (2019) coined the phrase “Big Indirect Data” (BID) to emphasize that common perception of data sparsity in geotechnical engineering is only accurate within a site-specific context. Indirect data arising from sites outside of the project boundary can range from irrelevant to relevant, but one can imagine abundance of the order of tens of thousands of soil records at a regional/national scale. It is pedantic to ignore BID – an experienced engineer will consider data from comparable sites but he/she is unlikely to find time to trawl tens of thousands of potentially useful soil records systematically. Hence, comparable sites are mostly restricted to those within his/her experience base likely to be restricted to a few municipalities/regions. The challenge is how to complement current practice steeped in empiricism with data-driven methods to extract maximum value from BID. While engineering judgment remains pivotal in decision making, it is ineffective in dealing with MUSIC and BID and their

complex inter-relationships.

The compilation of generic multivariate soil databases is gaining interest in recent years. Table 1 summarizes some databases, labeled as (geo-material type)/(number of parameters of interest)/(number of data points). For example, the CLAY/10/7490 database (Ching and Phoon 2014a) consists of 7490 data points for 10 clay parameters. These generic databases can produce empirical transformation models among different soil properties, such as those shown in Figure 1. These models are based on generic databases covering multiple sites such as those presented in Table 1, because data from a single site are insufficient to establish a transformation model. It is worth noting that a transformation model constructed based on a generic database is a generic model. The generic correlation trend may not be the same as the local correlation trend for a particular site. Figure 2 shows an example of the site-specific effect in the correlation trend. Local trends are evidently different from the generic trend. This is also widely known, but there is no method of characterizing this site effect in a routine project quantitatively, because site-specific data are sparse. Hence, the need to appeal to qualitative understanding of geology and engineering judgment for property estimation.

Although it is preferable that site-specific soil property estimation is made based on site-specific transformation models, site-specific data are MUSIC. Transformation models based on site-specific data may not have sufficient robustness. Adopting a generic transformation model is one possible solution, but the generic trend may not be the same as the site-specific trend. The generic transformation uncertainty is also large because the generic database covers diverse soil types. One practical outcome of a large transformation uncertainty is that a reasonable lower bound estimation, say based on the lower bound of the 95% confidence interval, will be very conservative.

The purpose of the current paper is to explore the possibility of conducting site-specific soil property estimation based on site-specific data with the aid from a generic database. First, a Bayesian method capable of analyzing MUSIC site-specific data is presented. The method can construct the site-specific model even when the site-specific data are very sparse. The data sparsity is rigorously treated and is reflected as the statistical uncertainty. Then, two strategies of making use of generic database are presented. One strategy is called "Bayesian data mining". This strategy finds generic cases that are "similar" to the site of interest and constructs a quasi-site-specific transformation model based on the combination of the site-specific data and the similar generic cases. The other strategy is called "hybridization". This strategy hybridizes the generic model and site-specific model in a rational way. Both methods exhibit reasonable behaviors: when the site-specific are abundant, the result converges to the site-specific model, and when the site-specific data are very sparse, the result converges to the generic model. Finally, a recent development that further considers spatial correlation is briefly mentioned. Case histories are used to demonstrate the use of these strategies.

Table 1. Summary of some soil/rock databases

Database	Reference	Parameters of interest	# Data points	# Sites/studies
CLAY/5/345	Ching and Phoon (2012)	LI, s_u, s_u^{re}, σ'_p, σ'_v	345	37 sites
CLAY/6/535	Ching et al. (2014)	s_u/σ'_v, OCR, q_{t1}, q_{tu}, $(u_2-u_0)/\sigma'_v$, B_q	535	40 sites
CLAY/7/6310	Ching and Phoon (2013)	s_u from 7 different test procedures	6310	164 studies
CLAY/10/7490	Ching and Phoon (2014a)	LL, PI, LI, σ'_v/P_a, S_t, B_q, σ'_p/P_a, s_u/σ'_v, q_{t1}, q_{tu}	7490	251 studies
F-CLAY/7/216	D'Ignazio et al. (2016)	s_u^{FV}, σ'_v, σ'_p, w_n, LL, PL, S_t	216	24 sites
FG/KSAT-1358	Feng and Vardanega (2019a, b)	e, k_{sat}, LL, PI	1358	33 studies
J-Clay/5/124	Liu et al. (2016)	M_r, q_c, f_s, w_n, γ_d	124	16
SAND/7/2794	Ching et al. (2017)	D_{50}, C_u, D_r, σ'_v/P_a, ϕ', q_{t1}, $(N_1)_{60}$	2794	176 studies
ROCK/9/4069	Ching et al. (2018)	n, γ, R_L, S_h, σ_{bt}, I_{s50}, V_p, σ_c, E	4069	184 studies

Note: LL = liquid limit; PL = plastic limit; PI = plasticity index; LI = liquidity index; w_n = natural water content; M_r = resilient modulus; q_c = cone tip resistance; f_s = sleeve friction; γ_d = dry density; D_{50} = median grain size; C_u = coefficient of uniformity; D_r = relative density; e = void ratio; k_{sat} = saturated hydraulic conductivity; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; s_u = undrained shear strength; s_u^{FV} = undrained shear strength from field vane; s_u^{re} = remoulded s_u; ϕ' = effective friction angle; S_t = sensitivity; OCR = overconsolidation ratio; $q_{t1} = (q_t/P_a)\times C_N$ (C_N is the correction factor for overburden stress); $q_{tu} = (q_t-u_2)/\sigma'_v$ = effective cone tip resistance; u_0 = hydrostatic pore pressure; B_q = pore pressure ratio = $(u_2-u_0)/(q_t-\sigma_v)$; P_a = atmospheric pressure = 101.3 kPa;; $(N_1)_{60} = N_{60}\times C_N$ (N_{60} is the N value corrected for the energy ratio); n = porosity; γ = unit weight; R = Schmidt hammer hardness (R_L = L-type Schmidt hammer hardness); S_h = Shore scleroscope hardness; σ_{bt} = Brazilian tensile strength; I_s = point load strength index (I_{s50} = I_s for diameter 50 mm); V_p = P-wave velocity; σ_c = uniaxial compressive strength; E = Young's modulus.

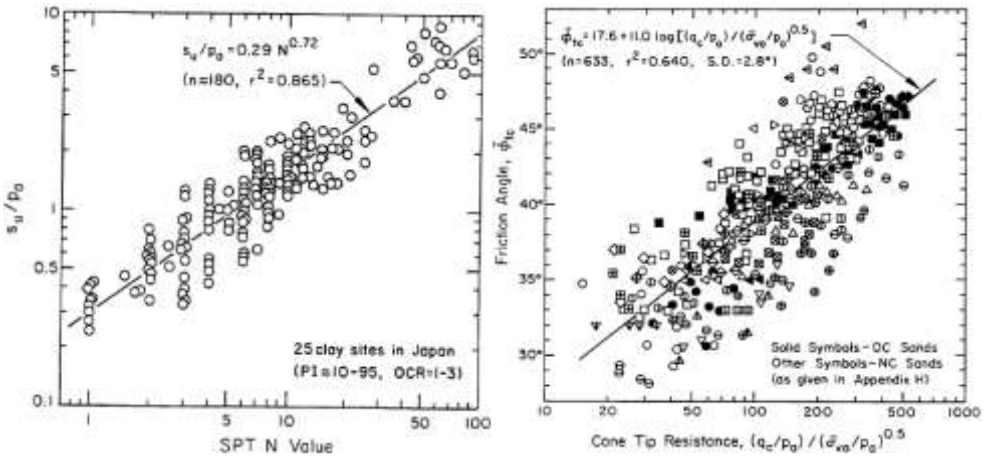


Fig. 1. Examples of transformation models in EPRI EL-6800 (Kulhawy and Mayne 1990)

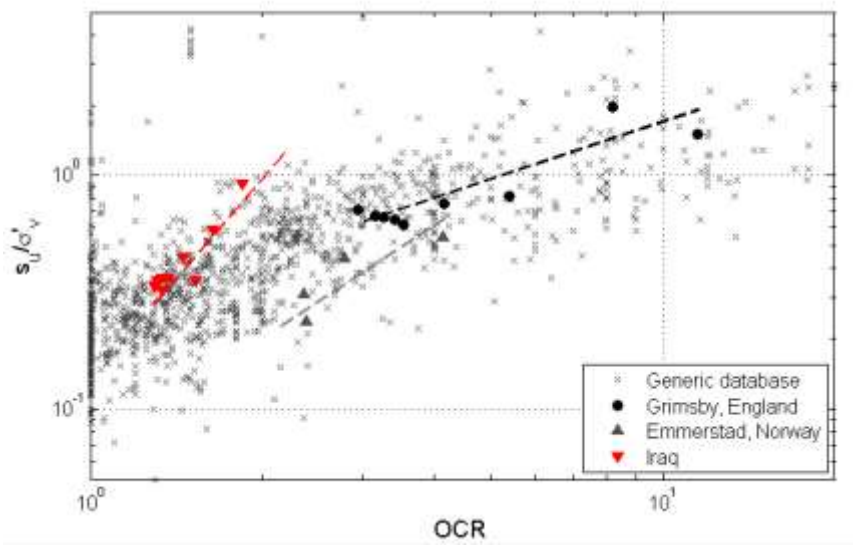


Fig. 2. Example of site-specific effects in the correlation trend (Ching and Phoon 2019a)

2 GENERIC DATABASE

Consider the generic database CLAY/10/7490 (Ching and Phoon 2014a). This database consists of 7490 records for 10 dimensionless clay parameters from 251 studies in the literature that cover 30 countries/regions worldwide. It is not specific to any local site. The 10 clay parameters are denoted by (Y_1, Y_2, …, Y_{10}):

$$\begin{aligned} Y_1 &= \ln(LL) & Y_2 &= \ln(PI) \\ Y_3 &= LI & Y_4 &= \ln(\sigma'_v/P_a) \\ Y_5 &= \ln(\sigma'_p/P_a) & Y_6 &= \ln(s_u/\sigma'_v) \\ Y_7 &= \ln(S_t) & Y_8 &= B_q \\ Y_9 &= \ln(q_{t1}) & Y_{10} &= \ln(q_{tu}) \end{aligned} \quad (1)$$

where LL = liquid limit; PI = plasticity index; LI = liquidity index; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; P_a = atmospheric pressure = 101.3 kPa; 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 = $(u_2-u_0)/(q_t-\sigma_v)$; u_0 = hydrostatic pore pressure; q_{t1} = $(q_t-\sigma_v)/\sigma'_v$; q_{tu} = $(q_t-u_2)/\sigma'_v$. The s_u values are all converted to the "mobilized" s_u values, which is the in-situ undrained shear strength mobilized in embankment and slope failures (Mesri and Huvaj 2007). The records can be visualized as a spreadsheet table of size (N_{db} × m), where N_{db} = 7490 is the total number of records in the target database and m = 10 is the dimension of each record, there are lots of missing entries in the spreadsheet table. Each record (row) is denoted as $\underline{y}_{db}$, a vector containing 10 values. A missing value in a record $\underline{y}_{db}$ means that a particular test has not been carried out for this record.

Ching and Phoon (2014b) adopted a transform based on the cumulative distribution function (CDF) of the Johnson distribution (Johnson 1949) to convert each record $\underline{y}_{db}$ in CLAY/10/7490 to a (roughly) multivariate standard normal record $\underline{x}_{db}$ = (X_1, X_2, …, X_{10}). Ching

and Phoon (2014b) further showed that the $\underline{x}_{db}$ records in CLAY/10/7490 roughly follow a multivariate normal probability density function (PDF):

$$f_{db}(\underline{x}) = |\mathbf{C}_{db}|^{-\frac{1}{2}} (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_{db})^T \mathbf{C}_{db}^{-1}(\underline{x}-\underline{\mu}_{db})} \quad (2)$$

where the subscript 'db' denotes 'database'; $\underline{\mu}_{db}$ is the mean vector ($\underline{\mu}_{db} = \underline{0}$); $\mathbf{C}_{db}$ is the covariance matrix, which can be found in Table 8 in Ching and Phoon (2014b).

3 SITE-SPECIFIC DATA

At a site, geotechnical data are typically MUSIC, multivariate, unique (site-specific), uncertain, sparse, and incomplete. Table 2 shows the site investigation results for a clay site in Onsøy, Norway (Lacasse and Lunne 1982). The dataset can be visualized as a spreadsheet table of size (9 × 10), where $N_s = 9$ is the total number of records (measured depths) in Table 2 and m = 10 to match information available in CLAY/10/7490. "Incomplete" means there are missing entries in the spreadsheet table. The term "sparse" refers to a small N_s. Each record (row) in Table 2, denoted by $\underline{y}_s$, are also converted to $\underline{x}_s$ using the same Johnson CDF transform proposed by Ching and Phoon (2014b). It is further assumed that the resulting $\underline{x}_s$ also follows a multivariate normal PDF:

$$f_s(\underline{x}) = |\mathbf{C}_s|^{-\frac{1}{2}} (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_s)^T \mathbf{C}_s^{-1}(\underline{x}-\underline{\mu}_s)} \quad (3)$$

where the subscript 's' denotes 'site'. Note that $\underline{\mu}_s$ (the site-specific mean vector) and $\mathbf{C}_s$ (the site-specific covariance matrix) are unknown. In fact, $\underline{\mu}_s$ and $\mathbf{C}_s$ can be highly uncertain if the site-specific data (e.g., Table 2) are MUSIC.

Suppose that the purpose is to construct the site-specific transformation model between OCR and s_u/σ'_v (classical SHANSEP model). Figure 3 shows the OCR-s_u/σ'_v relationship for the Onsøy data points in Table 2. With eight data points only, it is challenging to construct the site-specific OCR-s_u/σ'_v relationship with high precision. For comparison, the OCR-s_u/σ'_v relationship for the records in CLAY/10/7490 is also shown in the figure. In the following, two strategies of making use the CLAY/10/7490 database to enhance the precision of the OCR-s_u/σ'_v transformation model will be presented next. One method is called "Bayesian data mining" and the other is called "hybridization".

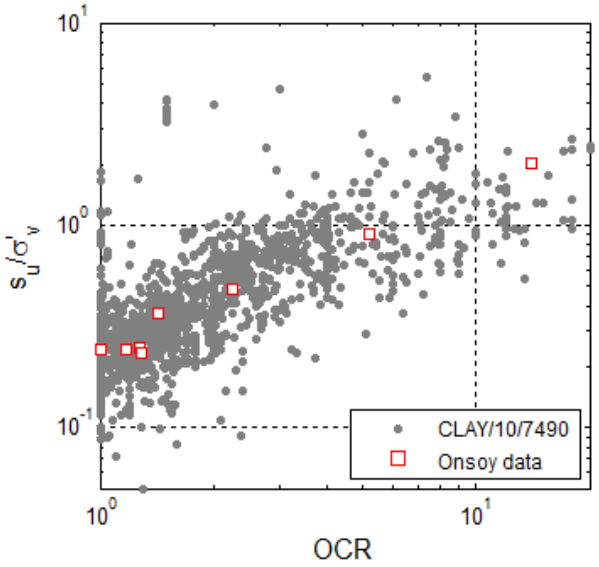


Fig. 3. OCR-s_u/σ'_v relationship.

4 BAYESIAN DATA MINING

The Bayesian data mining approach proposed by Ching and Phoon (2019a) contains three steps. In the first step, a Gibbs sampler method is proposed to construct the site-specific PDF, denoted by $f_s(\underline{x}|\mathbf{D})$, where **D** denotes the site data in Table 2. In its essence, $f_s(\underline{x}|\mathbf{D})$ summarizes the correlation behaviors among the soil properties at the Onsøy site as a multivariate PDF. The sharpness of this PDF depends on the amount of the site-specific data. The PDF is sharp if the site-specific data are abundant and is flat if the data are sparse and incomplete. This behavior is reasonable, because it is not possible to say which realizations are more likely when data are sparse. In other words, in the near absence of information, all realizations are equally likely.

Table 2 Site investigation data for a site in Onsøy, Norway (Source: Lacasse and Lunne 1982).

Index	Depth (m)	Site-specific data **Y**										
		LL (Y_1)	PI (Y_2)	LI (Y_3)	σ'_v/P_a (Y_4)	σ'_p/P_a (Y_5)	s_u/σ'_v (Y_6)	S_t (Y_7)	B_q (Y_8)	q_{t1} (Y_9)	q_{tu} (Y_{10})	OCR
1	1.0	56.2	20.0	1.54	0.06	0.85	2.03	6	0.16	29.11	25.57	13.99
2	1.9	50.2	18.1	1.82	0.12	0.60	0.91	14	0.24	17.69	14.58	5.20
3	3.5	59.9	30.5	0.93	0.22	0.48	0.48	15	0.30	10.52	8.41	2.26
4	5.2	56.8	22.9	1.07	0.32	0.45	0.37	7	0.35	7.70	6.11	1.42
5	7.6	66.3	31.5	0.87	0.47	0.54	0.24	14	0.47	5.89	4.25	1.17
6	9.5	65.1	29.6	0.97	0.58		0.25	12	0.41	6.19	4.74	
7	10.8	74.4	36.1	0.81	0.65	0.84	0.25	9	0.46	5.93	4.31	1.28
8	13.4	71.4	35.8	0.87	0.81	1.05	0.24		0.47	5.95	4.24	1.29

9	16.3	72.7	34.7	0.76	0.99	0.99	0.24		0.55	6.13	3.88	1.00

In the second step, a similarity measure quantifying the similarity between a database record (denoted by $\underline{x}_{db}$) and D is calculated. Here, $\underline{x}_{db}$ is a record in the database, one row in the (N_{db} × m) spreadsheet for CLAY/10/7490. This similarity measure, denoted by $S(\underline{x}_{db})$, is constructed such that a record with a larger $S(\underline{x}_{db})$ is more similar to the Onsøy site. In the third step, the quasi-site-specific transformation model is constructed based on the combination of the site-specific data and database records.

4.1 Step 1: Construction of the site-specific PDF

The main technical challenge for constructing the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ is that **D** can be incomplete, because most parameter estimation techniques require complete **D**. Ching and Phoon (2019a) showed that this challenge can be addressed by adopting the Gibbs sampler (GS) (Geman and Geman 1984; Gilks et al. 1996) in conjunction with the assumed non-informative conjugate prior PDFs. The GS is capable of drawing ($\underline{\mu}_s$, $\mathbf{C}_s$) samples conditioning on incomplete **D**, and based on the ($\underline{\mu}_s$, $\mathbf{C}_s$) samples, the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ can be approximated as the following mixture of multivariate normal PDF:

$$f_s(\underline{x}\,|\,\mathbf{D}) \approx \frac{1}{T-t_b}\left[\sum_{t=t_b+1}^{T} |\mathbf{C}_{s,t}|^{-\frac{1}{2}} (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_{s,t})^T (\mathbf{C}_{s,t})^{-1}(\underline{x}-\underline{\mu}_{s,t})}\right] \quad (4)$$

where ($\underline{\mu}_{s,t}$, $\mathbf{C}_{s,t}$) denote the ($\underline{\mu}_s$, $\mathbf{C}_s$) sample at time step t in GS; T is the total number of time steps in GS; t_b is the end of the burning-in period. Because of the use of non-informative prior PDFs, the resulting $f_s(\underline{x}|\mathbf{D})$ can sensibly reflect the statistical uncertainty associated with MUSIC Onsøy data.

To illustrate the behavior of the GS method, Figure 3 illustrates the shape of $f_s(\underline{x}|\mathbf{D})$ for a simulated example, the histogram of the mean of X_1, and the histogram of the correlation coefficient for N_s = 2, 10, and 100 data points simulated from a bivariate standard normal distribution (X_1, X_2) with mean = $\underline{\mathbf{0}}$ and correlation coefficient = 0.8. The resulting site-specific PDF $f_s(\underline{x}|\mathbf{D})$ is not a bivariate normal distribution. It is flat or non-informative when N_s = 2, because there is almost no site data to "learnt" from. The histogram of the mean covers a wide range and the histogram of the correlation coefficient is not too far from a uniform distribution as to be expected. Nonetheless, when N_s increases, $f_s(\underline{x}|\mathbf{D})$ converges to the underlying PDF with zero mean and correlation coefficient = 0.8.

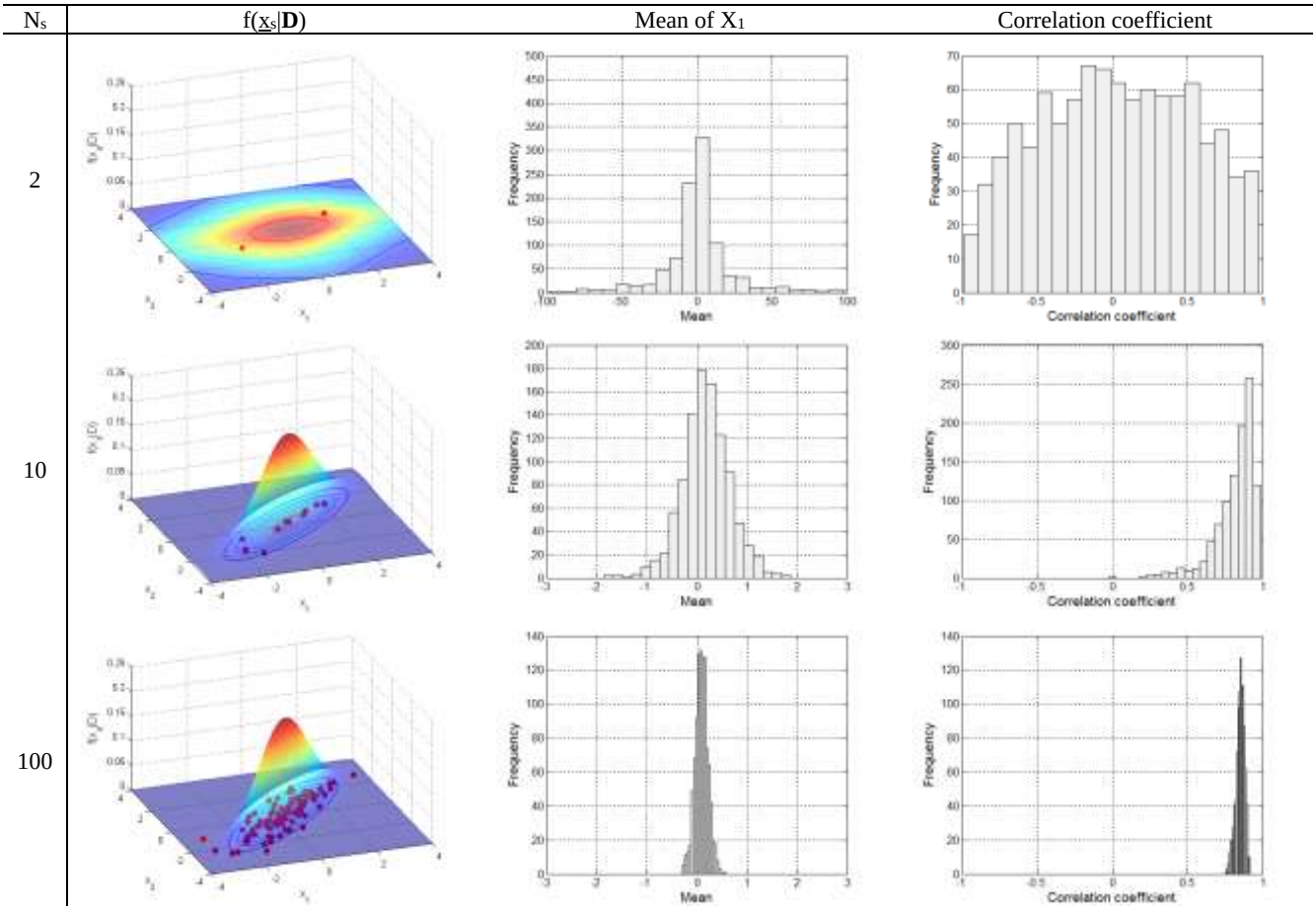


Fig. 3. Site-specific PDF $f(\underline{x}_s|\mathbf{D})$ and the histograms of mean and correlation coefficient "learnt" from 2, 10, and 100 measured data points simulated from a bivariate standard normal distribution (X_1, X_2) with mean = 0 and correlation coefficient = 0.8.

4.2 Step 2: Computation of the similarity measure

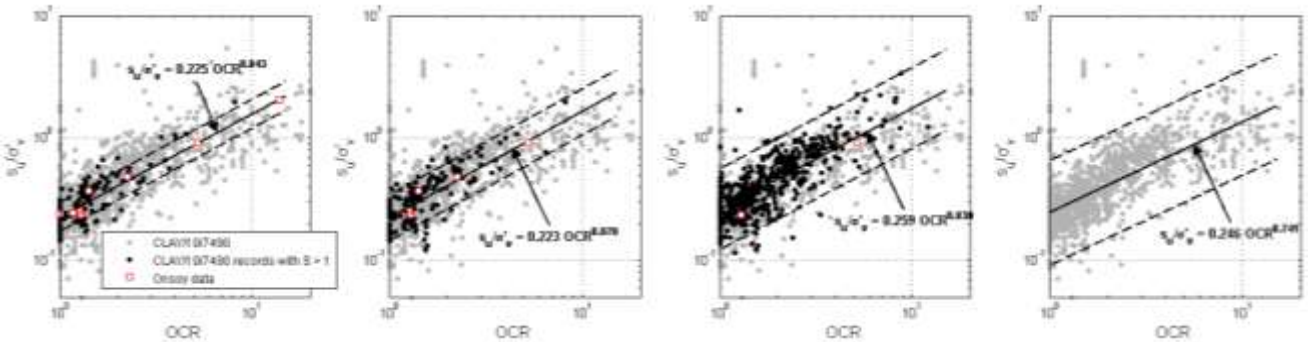
The site-specific PDF $f_s(\underline{x}|\mathbf{D})$ obtained from Step 1 summarizes the data structure at the Onsøy site. In Step 2, a similarity measure $S(\underline{x}_{db})$ is proposed in Ching and Phoon (2019a) to measure the similarity between $f_s(\underline{x}|\mathbf{D})$ and a database record $\underline{x}_{db}$. The main technical challenge here is, again, $\underline{x}_{db}$ can be incomplete, i.e., $\underline{x}_{db}$ is a (10×1) vector with missing entries. Let us denote $\underline{x}_{db}^{o}$ a vector that contains only the observed entries. For instance, if only the 1st, 3rd, and 7th entries in $\underline{x}_{db}$ are observed. The 'o' superscript in $\underline{x}_{db}^{o}$ means only the 1st, 3rd, and 7th entries are selected so that $\underline{x}_{db}^{o}$ is now a (3×1) vector. Ching and Phoon (2019a) proposed the following similarity measure $S(\underline{x}_{db})$:

$$S(\underline{x}_{db}) = \frac{\sum_{t=t_b+1}^{T} \left|\mathbf{C}_{s,t}^{o}\right|^{-\frac{1}{2}} \times e^{-\frac{1}{2}\left(\underline{x}_{db}^{o}-\underline{\mu}_{s,t}^{o}\right)^{T}\left(\mathbf{C}_{s,t}^{o}\right)^{-1}\left(\underline{x}_{db}^{o}-\underline{\mu}_{s,t}^{o}\right)}}{\sum_{t=t_b+1}^{T} \left|\mathbf{C}_{db}^{o}+\mathbf{C}_{s,t}^{o}\right|^{-\frac{1}{2}} \times e^{-\frac{1}{2}\left(\underline{\mu}_{db}^{o}-\underline{\mu}_{s,t}^{o}\right)^{T}\left(\mathbf{C}_{db}^{o}+\mathbf{C}_{s,t}^{o}\right)^{-1}\left(\underline{\mu}_{db}^{o}-\underline{\mu}_{s,t}^{o}\right)}} \quad (5)$$

where $\underline{\mu}_{db}$ and $\mathbf{C}_{db}$ are the mean and covariance matrix that summarize the second-moment statistics of the CLAY/10/7490 database for the illustrative example discussed in this paper (see Eq. 2); the 'o' superscript in ($\underline{\mu}_{s,t}^{o}$, $\underline{\mu}_{db}^{o}$, $\mathbf{C}_{s,t}^{o}$, $\mathbf{C}_{db}^{o}$) has the similar meaning, e.g., $\underline{\mu}^{o}$ is a (3×1) sub-vector that only contains the 1st, 3rd, and 7th entries in $\underline{\mu}$, and $\mathbf{C}^{o}$ is a (3×3) sub-matrix. For each record $\underline{x}_{db}$ in CLAY/10/7490, its similarity measure with respect to the Onsøy site can be computed. For $S(\underline{x}_{db})$ computed based on Eq. (5), Ching and Phoon (2019a) showed that for a randomly chosen record in CLAY/10/7490, its $S(\underline{x}_{db})$ is "on average" 1, regardless of whether there are missing entries or how many the missing entries are. In other words, a record $\underline{x}_{db}$ with $S(\underline{x}_{db}) > 1$ is more similar to the Onsøy site than an average record in CLAY/10/7490 regardless of where or how many the missing entries are. Let us denote the i-th record in CLAY/10/7490 by $\underline{x}_{db}^{i}$. Because the average $S(\underline{x}_{db})$ is the same, $S(\underline{x}_{db}^{i})$ and $S(\underline{x}_{db}^{j})$ can be compared even though $\underline{x}_{db}^{i}$ and $\underline{x}_{db}^{j}$ have different observed components. Table 3 shows the top 10 records in CLAY/10/7490 with the highest $S(\underline{x}_{db})$ values. These records all have $S(\underline{x}_{db})$ values significantly larger than 1, suggesting that they are way more similar to the Onsøy site than an average record in CLAY/10/7490.

Table 3 Top 10 records in CLAY/10/7490 with the highest $S(\underline{x}_{db})$ values.

Rank	$S(\underline{x}_{db})$	LL (%)	PI (%)	LI	σ'_v/P_a	σ'_p/P_a	s_u/σ'_v	S_t	B_q	q_{t1}	q_{tu}	OCR	Location
1	389.0	61.8	28.1	1.10	0.44	0.46	0.38	12.0				1.04	Okishin (Japan)
2	71.2	73.6	36.5		0.46	0.66	0.49		0.40	7.76	5.34	1.43	Bothkennar (UK)
3	53.6	67.0	35.0	0.80	0.73	1.54	0.26		0.35	14.28	10.19	2.10	Anacostia (USA)
4	51.9	64.2	37.4	0.96	0.12	0.49	1.00					4.03	-
5	45.0	72.7	46.8	0.82	0.70	0.70	0.22	6.3				1.00	Shellhaven (UK)
6	44.8	78.2	42.5	0.69	0.74	0.79	0.24	4.0				1.06	Shellhaven (UK)
7	40.0	64.4	40.0	1.00	0.68	0.78	0.23					1.15	Canada
8	37.2	60.0	30.0	0.93	0.17	0.38	0.54					2.28	USA
9	33.2	75.8	60.5	0.77	0.74	1.13	0.21	3.0	0.50	5.37	3.67	1.54	Drammen (Norway)
10	28.0	62.0	32.0	1.09	0.29	0.33	0.32					1.13	USA



(a) All Onsøy records (b) 6 Onsøy records (c) 2 Onsøy records (d) no record

Fig. 4. Quasi-site-specific transformation models by combining Onsøy data with records in CLAY/10/7490.

4.3 Step 3: Construct the quasi-site-specific transformation model

Ching and Phoon (2019a) proposed a method for constructing the quasi-site-specific transformation model based on the combination of the site-specific records (e.g., Onsøy) and database records (e.g., CLAY/10/7490). For the Onsøy data, a "leave-one-out method" is adopted to compute $S(\underline{x}_s)$ for each record in the Onsøy data with respect to the remaining N_s-1 Onsøy records. This is done by first constructing the leave-one-out site-specific PDF using the N_s-1 Onsøy records and then computing the $S(\underline{x}_s)$ for the left-out $\underline{x}_s$ using Eq. 5. By conducting this leave-one-out method, each Onsøy record is associated with an $S(\underline{x}_s)$ value. The two types of records, Onsøy and CLAY/10/7490, are next combined. Each record in this combined dataset has an $S(\underline{x})$ value. A bootstrap method is then adopted to resample the records in the combined dataset based on their weights that are proportional to their $S(\underline{x})$ values. The proportionality relationship has been established using the leave-one-out method described above. The resampled records are used to construct the quasi-site-specific transformation model of interest, the OCR-s_u/σ'_v relationship:

$$\ln\left(s_u/\sigma'_v\right) = a + b\cdot\ln\left(OCR\right) + \varepsilon \qquad (6)$$

where (a,b) are unknown SHANSEP parameters to be determined, and ε is assumed to be a zero-mean normal variable with standard deviation = σ, also unknown and to be determined. This bootstrap method is conducted many times to obtain many bootstrap samples for (a,b,σ). These bootstrap samples can be used to establish the median and 95% confidence interval for the OCR-s_u/σ'_v relationship.

Figure 4 shows the median estimate and 95% confidence interval for the OCR-s_u/σ'_v relationship based on the bootstrap method. The median is the solid line. The 95% confidence interval is given by the pair of dashed lines. The CLAY/10/7490 records with $S(\underline{x}_{db})$ > 1 are shown as dark circles to illustrate a representative sample of "similar" records - analysis is not restricted to these samples. The resulting formula for the median quasi-site-specific transformation model are also annotated in the plots. To demonstrate the effect of sparsity of the site-specific data, four scenarios are considered: (a) all site-specific data in Table 2 are adopted; (b) only six records (rows) in Table 2 (depths = 1.9, 3.5, 5.2, 9.5, 10.8, and 13.4) are available; (c) only two records (rows) in Table 2 (depths = 1.9 and 13.4) are available; (c) no site-specific data is available. From Fig. 4, it is clear that there are less records with $S(\underline{x}_{db})$ > 1 when site-specific are abundant (Fig. 4a), and the number of records with $S(\underline{x}_{db})$ > 1 increases when the site-specific data are sparse. This is because the proposed method correctly captures the statistical uncertainty. Moreover, the transformation uncertainty, quantified by the 95% confidence interval, seems to increase with decreasing amount of site-specific data. When there are more site-specific data points (Fig. 4a), the analysis is significantly affected by the site-specific data. When there is no site-specific data (Fig. 4d), the analysis is completely governed by the CLAY/10/7490 database. Fig. 4c may represent the scenario where site-specific data are very sparse. With only two site-specific OCR-s_u/σ'_v records, traditional regression analysis may not be able to construct a transformation model with any acceptable robustness. Nonetheless, with the proposed method that adopts the union dataset (e.g., Onsøy + CLAY/10/7490), it is now possible to construct a quasi-site-specific transformation model with acceptable robustness.

4 HYBRIDIZATION

The hybridization method is an alternative approach proposed by Ching and Phoon (2019b) of dealing with MUSIC site-specific data. When site-specific data are MUSIC (sparse and incomplete), the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ has significant statistical uncertainty. Figure 3a shows one such example. It may not be possible to construct a useful OCR-s_u/σ'_v relationship purely based on the site-specific data. In this case, it is sensible to rely more on generic database, CLAY/10/7490. 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 contrast, when site-specific training data are abundant, it is sensible to rely more on site-specific data. Estimation purely based on site-specific data is equivalent to adopting $f_s(\underline{x}|\mathbf{D})$ in Eq. (4), and estimation purely based on generic database is equivalent to adopting $f_{db}(\underline{x})$ in Eq. (2). In this section, a method is proposed to hybridize $f_s(\underline{x}|\mathbf{D})$ and $f_{db}(\underline{x})$ so that the hybrid PDF approaches $f_{db}(\underline{x})$ when site-specific data are very sparse and approaches $f_s(\underline{x}|\mathbf{D})$ when site-specific data are abundant.

The idea of hybridization proposed by Ching and Phoon (2019b) is straightforward: the hybrid multivariate PDF, denoted by $f_{hb}(\underline{x}|\mathbf{D})$, is proportional to the direct product between $f_s(\underline{x}|\mathbf{D})$ and $f_{db}(\underline{x})$:

$$f_{hb}\left(\underline{x}\,|\,\mathbf{D}\right) \propto f_{db}\left(\underline{x}\right)\cdot f_s\left(\underline{x}\,|\,\mathbf{D}\right) \qquad (7)$$

Figure 5 illustrates the hybridization idea and explains why it works. The generic PDF $f_{db}(\underline{x})$ (the solid curves in the figure) does not change with respect to the amount of site-specific data **D** because it only depends on $\underline{\mu}_{db}$ and $\mathbf{C}_{db}$. However, the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ depends on the amount of **D**: it is relatively flat when **D** is sparse and incomplete (Fig. 5a) and is relatively peaked when **D** is abundant (Fig. 5b). When **D** is sparse and incomplete (Fig. 5a), $f_{hb}(\underline{x}|\mathbf{D}) \propto f_{db}(\underline{x})\times$(a relatively flat PDF) $\propto f_{db}(\underline{x})$, hence the hybrid PDF approaches $f_{db}(\underline{x})$. When **D** is abundant (Fig. 5b), the opposite happens: the hybrid PDF $\propto$ (a relatively flat

PDF)×$f_s(\underline{x}|\mathbf{D})$ ∝ $f_s(\underline{x}|\mathbf{D})$, hence the hybrid PDF approaches $f_s(\underline{x}|\mathbf{D})$.

By combining Eqs. (2) and (5), Ching and Phoon (2019b) showed that the hybrid PDF $f_{hb}(\underline{x}|\mathbf{D})$ is still a mixture of multivariate normal PDF:

$$\begin{aligned} f_{hb}\left(\underline{x}\,|\,\mathbf{D}\right) &\propto f_{db}\left(\underline{x}\right)\cdot f_s\left(\underline{x}\,|\,\mathbf{D}\right) \\ &\propto \sum_{t=t_b+1}^{T} w_t \times \left|\mathbf{C}_{hb,t}\right|^{-\frac{1}{2}} (2\pi)^{-\frac{m}{2}} e^{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_{hb,t}\right)^T \left(\mathbf{C}_{hb,t}\right)^{-1}\left(\underline{x}-\underline{\mu}_{hb,t}\right)} \end{aligned} \quad (8)$$

where w_t is the weight for each multivariate normal PDF:

$$w_t = \left|\mathbf{C}_g + \mathbf{C}_{s,t}\right|^{-\frac{1}{2}} \times (2\pi)^{-\frac{m}{2}} \times e^{-\frac{1}{2}\underline{\mu}_{s,t}^T \left(\mathbf{C}_g+\mathbf{C}_{s,t}\right)^{-1}\underline{\mu}_{s,t}} \quad (9)$$

and

$$\underline{\mu}_{hb,t} = \left(\mathbf{C}_g^{-1} + \mathbf{C}_{s,t}^{-1}\right)^{-1} \mathbf{C}_{s,t}^{-1}\underline{\mu}_{s,t} \qquad \mathbf{C}_{hb,t} = \left(\mathbf{C}_g^{-1} + \mathbf{C}_{s,t}^{-1}\right)^{-1} \quad (10)$$

To demonstrate the effect of hybridization, Fig. 6 shows how the hybrid PDF varies with respect to the amount of site-specific data. In the figure, the probability density contours for the OCR-s_u/σ'_v bivariate hybrid PDFs are shown. The four scenarios with different data sparsity in Fig. 4 are also considered in Fig. 6. In general, the observations in Fig. 6 are similar to those in Fig. 4: the transformation uncertainty increases with decreasing amount of site-specific data. When there is no site-specific data (Figs. 4d & 6d), the result is completely governed by the CLAY/10/7490 database.

The OCR-s_u/σ'_v bivariate hybrid PDF can be further used to deduce the median estimate and 95% confidence interval for the OCR-s_u/σ'_v relationship, shown as the thick red lines in Fig. 7. Figure 7 can also be compared with Fig. 4 (the results in Fig. 4 are shown as thin dark lines in Fig. 7). It is clear that the median estimates and 95% confidence intervals obtained from the Bayesian data mining method (Fig. 4) and from the hybridization method (Fig. 7) are qualitatively similar (e.g., the confidence interval is wide when site-specific data are sparse) but are quantitatively different. This suggests that the two methods (Bayesian data mining and hybridization) are not equivalent.

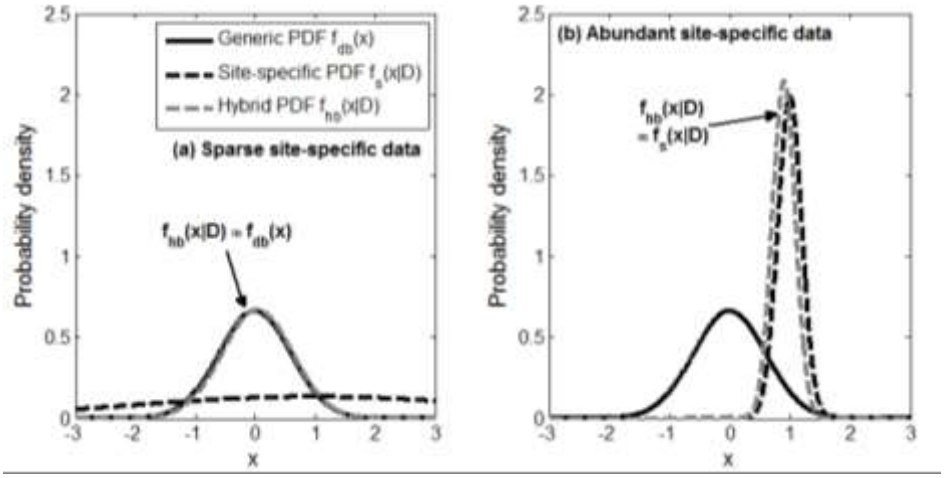
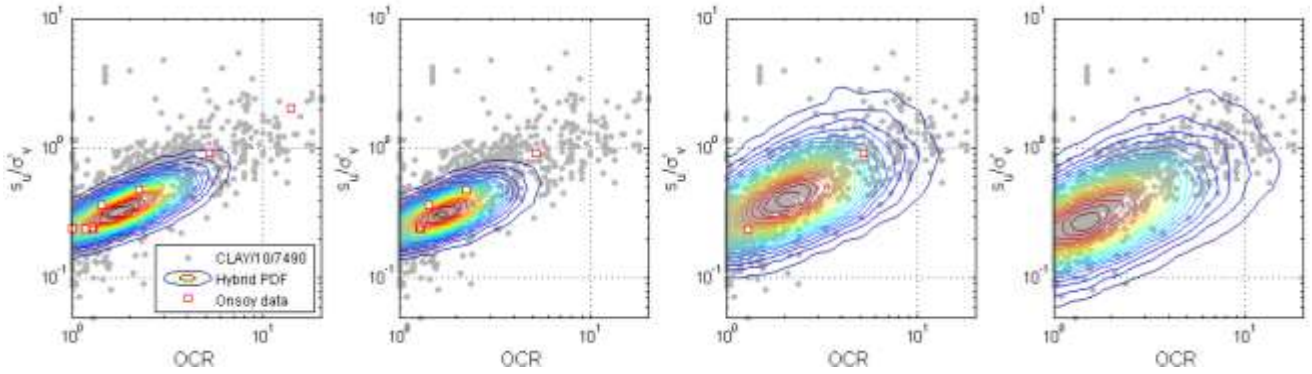
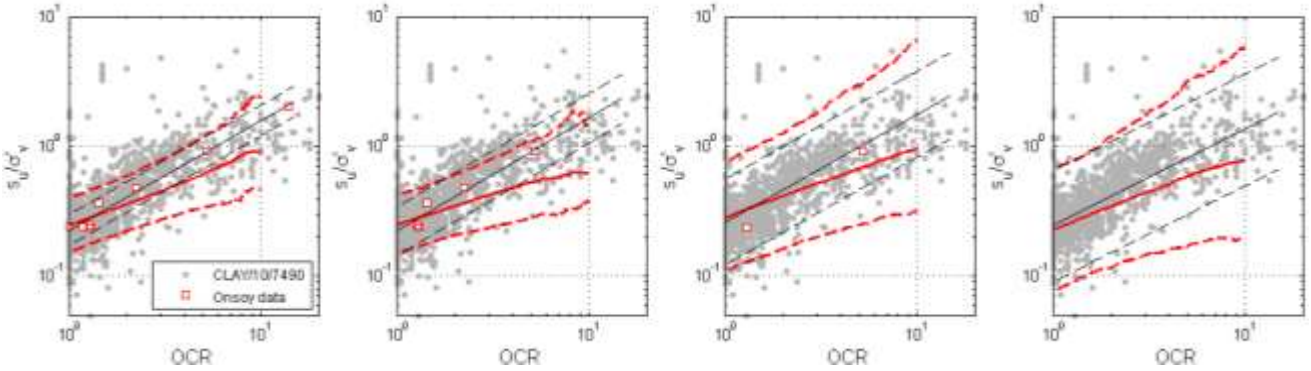


Fig. 5. Illustration of hybridization: (a) sparse site-specific data; (b) abundant site-specific data.



(a) All Onsøy records (b) 6 Onsøy records (c) 2 Onsøy records (d) no record

Fig. 6. Contour plots for the OCR-s_u/σ'_v bivariate hybrid PDF.



(a) All Onsøy records (b) 6 Onsøy records (c) 2 Onsøy records (d) no record

Fig. 7. Median estimate and 95% confidence interval for the OCR-s_u/σ'_v relationship deduced from the bivariate PDF.

5 SPATIAL CORRELATION

The above two methods (Bayesian data mining and hybridization) only considered the correlation among soil properties at the same depth when constructing the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ or the hybrid PDF $f_{hb}(\underline{x}|\mathbf{D})$. The correlation among soil properties at the same depth is called the cross correlation in this paper. However, these methods did not consider correlation among different depths. This is called the spatial correlation. In reality, soil properties are spatially correlated, i.e., soil properties at nearby depths are usually positively correlated. Ching and Phoon (2019c) proposed a modified Gibbs sampler method for constructing the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ that not only considers the cross correlation at the same depth but also considers the spatial correlation among different depths. The method is quite general, because it can construct the site-specific PDF $f_s(\underline{x}|\mathbf{D})$ based on spatially correlated site-specific data. It can also simulate conditional cross-correlated random fields based on the site-specific data. The hybridization idea can also be implemented to this method, which is illustrated in the current paper.

According to Ching and Phoon (2019c), the spatial-correlation (or auto-correlation) structure of the site-specific data needs to be first identified. This is achieved by analyzing the CPT data at the Onsøy site (Fig. 8) (Lacasse and Lunne 1982) by adopting the single exponential model (Vanmarcke 1983). The scale of fluctuation (δ) is identified to be about 1 m. The site-specific PDF $f_s(\underline{x}|\mathbf{D})$ can be estimated based on the spatially correlated data using the modified Gibbs sampler method proposed in Ching and Phoon (2019c). In the current paper, $f_s(\underline{x}|\mathbf{D})$ is further hybridized with $f_{db}(\underline{x})$ to obtain the hybrid PDF $f_{hb}(\underline{x}|\mathbf{D})$. Together with the identified auto-correlation model, this hybrid PDF $f_{hb}(\underline{x}|\mathbf{D})$ can simulate conditional cross-correlated random field samples for all soil properties. These conditional random field samples can be used to obtain the 95% confidence interval of the soil property profiles. The solid lines in Fig. 9 are the conditional cross-correlated random field samples for the σ'_p and s_u profiles, whereas the dashed lines are the resulting 95% confidence intervals. The random field samples are conditioning on the site-specific data in Table 2, so these random field samples pass through the site-specific data (the measured data in Fig. 9).

There are additional s_u data in Lacasse and Lunne (1982) not included in Table 2. These s_u data are shown as the validation data in Fig. 9. From Fig. 9, it can be seen that 25 out of the 28 validation data lie within the 95% confidence interval. Because 25/28 ≈ 90%, which is close to 95%, the 95% confidence interval seems to be effective for this particular example.

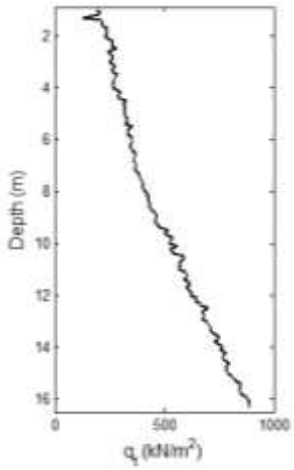


Fig. 8. Cone resistance profile at the Onsøy site.

CONCLUSION

Geotechnical engineering has two features that are in significant contrast with each other. On one hand, site-specific data are sparse and incomplete. On the other hand, generic (non-site-specific) data in the literature are abundant. There is a dilemma to choose between these two scenarios. By only relying on the

sparse and incomplete site-specific data, it is typically not feasible to construct site-specific transformation model for the purpose of soil property estimation. By only relying on the generic database, the constructed transformation model is not site-specific and the transformation uncertainty is large resulting in potentially very conservative lower bound estimates. This paper introduces two strategies that take advantage of both the site-specific data and the generic database to construct a quasi-site-specific transformation model.

It is worth noting that the methods introduced in this paper are purely data-driven. Therefore, their application is not limited to soil property estimation. They can be applied to other types of datasets, such as load test and monitoring data, as well.

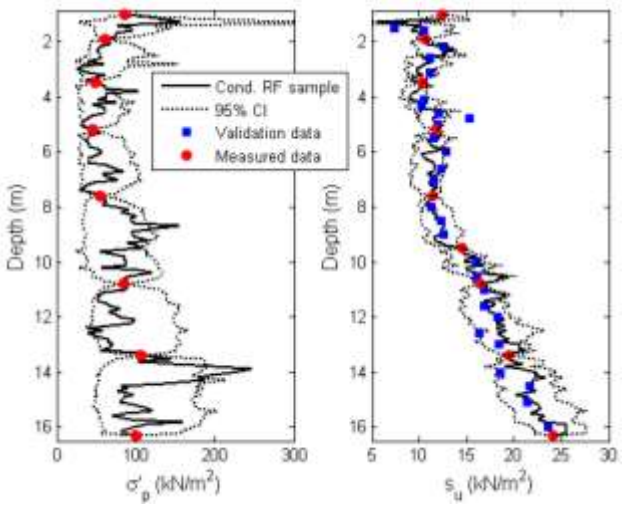


Fig. 9. Conditional cross-correlated random field samples and 95% confidence intervals for the σ'_p and s_u profiles.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the TC304 Committee (Risk) of the International Society of Soil Mechanics and Geotechnical Engineering for developing the database 304dB (http://140.112.12.21/issmge/Database_2010.htm) used in this study and making it available for public.

REFERENCES

Ching, J. and Phoon, K.K. (2012). Modeling parameters of structured clays as a multivariate normal distribution. Canadian Geotechnical Journal, 49(5), 522-545.

Ching, J. and Phoon, K.K. (2013). Multivariate distribution for undrained shear strengths under various test procedures. Canadian Geotechnical Journal, 50(9), 907-923.

Ching, J. and Phoon, K.K. (2014a). Transformations and correlations among some parameters of clays – the global database. Canadian Geotechnical Journal, 51(6), 663-685.

Ching, J. and Phoon, K.K. (2014b). Correlations among some clay parameters–the multivariate distribution. Canadian Geotechnical Journal, 51(6), 686-704.

Ching, J. and Phoon, K.K. (2019a). 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, under review.

Ching, J. and Phoon, K.K. (2019b). Constructing site-specific probabilistic transformation model by Bayesian machine learning. ASCE Journal of Engineering Mechanics, 145(1), 04018126.

Ching, J. and Phoon, K.K. (2019c). Constructing a site-specific multivariate probability distribution using sparse, incomplete, and spatially variable data. ASCE Journal of Engineering Mechanics, in review.

Ching, J., Li, K.H., Phoon, K.K., and Weng, M.C. (2018). Generic transformation models for some intact rock properties. Canadian Geotechnical Journal, 55(12), 1702-1741.

Ching, J., Lin, G.H., Chen, J.R., and Phoon, K.K. (2017). Transformation models for effective friction angle and relative density calibrated based on a multivariate database of coarse-grained soils. Canadian Geotechnical Journal, 54(4), 481-501.

Ching, J., Phoon, K.K., and Chen, C.H. (2014). Modeling CPTU parameters of clays as a multivariate normal distribution. Canadian Geotechnical Journal, 51(1), 77-91.

Ching, J., Phoon, K.K., Li, K.H. and Weng, M.C. (2019). Multivariate probability distribution for some intact rock properties. Canadian Geotechnical Journal, in press.

D'Ignazio, M., Phoon, K.K., Tan, S.A., and Lansivaara, T. (2016). Correlations for undrained shear strength of Finnish soft clays. Canadian Geotechnical Journal, 53(10), 1628-1645.

Feng, S., and Vardanega, P. J. (2019a). Correlation of the hydraulic conductivity of fine-grained soils with water content ratio using a database. Environmental Geotechnics, in press.

Feng, S., and Vardanega, P. J. (2019b). A database of saturated hydraulic conductivity of fine-grained soils: probability density functions. Georisk, in press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell., 6, 721-741.

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

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 36, 149-176.

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

Lacasse, S. and Lunne, T. (1982). Penetration tests in two Norwegian clays. Proc. 2nd Eur. Symp. on Penetration Testing, Amsterdam, 661-670.

Liu, S., Zou, H., Cai, G., Bheemasetti, B.V., Puppala, A.J., and Lin, J. (2016). Multivariate correlation among resilient modulus and cone penetration test parameters of cohesive subgrade soils. Engineering Geology, 209, 128–142.

Mesri, G. and Huvaj, N. (2007). Shear strength mobilized in undrained failure of soft clay and silt deposits. Advances in Measurement and Modeling of Soil Behaviour (GSP 173), Ed. D.J. DeGroot et al., ASCE, 1-22.

Phoon, K.K. (2018). Editorial for Special Collection on Probabilistic Site Characterization. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 4(4), 02018002.

Phoon, K.K., Ching, J. and Wang, Y. (2019). Managing risk in geotechnical engineering – from data to digitalization. Proceedings, 7th International Symposium on Geotechnical Safety and Risk (ISGSR 2019), Taipei, Taiwan.

Vanmarcke, E.H. (1983). Random Fields: Analysis and Synthesis. The MIT Press, Cambridge, Massachusetts.