Journal of GeoEngineering, Vol. 15, No. 4, pp. 173-182, December 2020
http://dx.doi.org/10.6310/jog.202012_15(4).1

2020 TGS Geotechnical Lecture:

VALUE OF GEOTECHNICAL BIG DATA AND ITS APPLICATION IN SITE-SPECIFIC SOIL PROPERTY ESTIMATION

Jianye Ching [1*]

ABSTRACT

Site uniqueness is a well-known feature in geotechnical engineering. Site investigation data obtained from one site cannot be directly used for another site. However, it is not uncommon that non-site-specific (generic) data is used to support site-specific decision-making. For instance, engineers routinely adopt transformation models to estimate design soil parameters, and most transformation models are calibrated by generic data. It is quite extreme and unrealistic to prohibit the use of such models. In contrast, the success of such transformation models indicates that generic data may have certain value for site-specific decision-making. In this era of BIG DATA, it is timely for geotechnical engineering people to ponder the value of generic databases. This paper first introduces some exiting BIG DATA in the geotechnical literature. Then, the paper introduces some advanced methods developed by the author and his colleagues that can exploit value from BIG DATA to facilitate site-specific estimation for soil properties. Without BIG DATA and the advanced methods, such site-specific estimation can be very challenging.

Key words: Geotechnical BIG DATA, soil properties, databases, site characterization, data-driven analysis.

1. INTRODUCTION

Recently, the term "BIG DATA" has attracted significant attention. The definition of BIG DATA can be found from Wikipedia (Wikipedia 2020a), quoted as follows:

"Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three vs. Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before."

BIG DATA is characterized by three Vs: variety, volume, and velocity. Traditional methods cannot manage BIG DATA. In the current paper, we focus on soil/rock property databases. These geotechnical databases have significant variety and volumes, but they do not have velocity. They may be qualified as "slow" BIG DATA. In the current paper, we simply call them as BIG DATA for brevity.

Moreover, the geotechnical BIG DATA that we discuss in the current paper may be further classified as "dark" BIG DATA, defined by Wikipedia as follows (Wikipedia 2020b):

"Dark data is data which is acquired through various computer network operations but not used in any manner to derive insights or for decision making."

Similar to "dark matter" in the universe, dark data contains lots of information, but it is not used because the data is not well structured. The soil/rock property presented in the current paper are "dark" in the sense that even though they exist in the literature, they do not have a clear structure so that they are not used to derive insights for decision making.

Does geotechnical BIG DATA have value in site-specific decision making? Notice that BIG DATA is non-local (generic) data. Site uniqueness is a well-known fact in geotechnical engineering, which suggests that generic data may not be directly applicable to a local target site. On the other hand, it is also well known that transformation models (Phoon and Kulhawy 1999a) such as the q_t-s_u model (q_t = cone tip resistance from cone penetration test; s_u = undrained shear strength of a clay) constructed by generic data are popular among practitioners. The popularity of transformation models indicates that generic data has certain value in site-specific decision making.

In contrast to BIG DATA, local data that is obtained from a local target site is called "small data" in the current paper. For instance, a typical site investigation plan may include a few boreholes and CPT soundings. Each borehole is used to derive index properties of the soil (*e.g.,* unit weight, Atterberg limits, water content, grain size distribution, etc.) as well as standard penetration test (SPT) N values, and each CPT sounding is used to derive q_t, f_s, and u_2 profiles (f_s = sleeve friction; u_2 = behind-cone pore pressure). Small data is the collection of all borehole and sounding data. Can we solely rely on small data for site-specific decision making? Phoon *et al.* (2019) has described small data as MUSIC (Multivariate, Unique, Sparse, Incomplete, and possibly Corrupted). Recently, Phoon (2020) coined the term MUSIC-X for small data, where X stands for spatial variability, to indicate that small data is also spatial variable. It may

Manuscript received and accepted November 9, 2020.
[1*] Professor (corresponding author), Department of Civil Engineering, Dept of Civil Engineering, National Taiwan University, Taipei, Taiwan (e-mail: jyching@gmail.com).

be questionable if we solely rely on sparse and incomplete small data for site-specific decision making.

The purpose of the current paper is as follows:

1. To review recent developments for geotechnical BIG DATA (soil/rock property databases).
2. To address the question "Can we solely rely on small data for site-specific decision making?" by a real case study, to illustrate that solely relying on small data may cause significant uncertainty in decision making.
3. To address the question "Does geotechnical BIG DATA have value in site-specific decision making?" by the same real case study, to illustrate the plain use of BIG DATA may not necessarily produce value.
4. To address the same question in #3 by the same real case study, but now a more advanced data model is adopted, to illustrate an advanced model together with BIG DATA may produce value.

The structure of the paper follows the above four steps. At the end, there is a conclusion.

2. GEOTECHNICAL BIG DATA

Soil/rock property datasets are abundant, but most of them are "dark" data, hidden in journal papers, conference papers, technical reports, dissertations, *etc.* Even though these generic dark datasets exist in the literature, they do not have a clear structure. Also, it is not clear how they are linked to a local target site. Nonetheless, many useful transformation models are constructed by such generic databases. Kulhawy and Mayne (1990) compiled lots of such models for soil properties, exemplified by the models shown in Fig. 1, whereas Zhang (2016) compiled lots of such models for rock properties.

Since 1990, many generic databases have been compiled. Table 1 shows some databases, labelled as (material type)/(number of parameters of interest)/(number of cases). Most databases in Table 1 are for soils, but there are five databases for rock or rock mass. Some databases are univariate (*i.e.,* only a single parameter is known for each soil/rock case). Many recent databases in Table 1 are multivariate (*i.e.,* multiple parameters are known for each case). The multivariate databases usually have a structure similar to a spreadsheet table with m rows and n columns, where m is the number of cases, and n is the number of parameters of interest. For instance, the CLAY/10/7490 database can be visualized as a spreadsheet table with m = 7490 rows and n = 10 columns. If the table is fully populated by data, the database is a complete database. There are a few complete multivariate databases in Table 1, but most multivariate databases are incomplete, and % of completeness in the table is defined as (# of populated cells)/(total # of cells). For many databases in Table 1, the site for each case is known. This site information is necessary for the hierarchical Bayesian model that is to be introduced later.

The multivariate soil/rock databases are especially valuable for the purpose of constructing transformation models. To illustrate multivariate databases, the dark dots in Fig. 2 show the data points in CLAY/5/345, which contains n = 5 clay parameters (LI, s_u, s_u^{re}, σ'_v, σ'_p) for m = 345 cases, where LI is the liquidity index, s_u is the undrained shear strength, s_u^{re} is the remolded undrained shear strength, σ'_v is the vertical effective stress, and σ'_p is the preconsolidation stress. The correlation behaviors among different clay parameters are sensible.

There are also some regional or municipal databases in Table 1. For instance, JS-CLAY stands for clay of Jiangsu province in China, and SH-CLAY stands for Shanghai clay. Regional or municipal databases usually exhibit less scatter than generic databases. Figure 3 shows the correlation plot between s_u/σ'_v and σ'_v/P_a (P_a = 101.3 kPa = one atmosphere pressure) for SH-CLAY/11/4051. The background grey dots are from CLAY/10/7490, a generic database that contains cases worldwide. It is clear that the data points for Shanghai exhibit less scatters than those for worldwide.

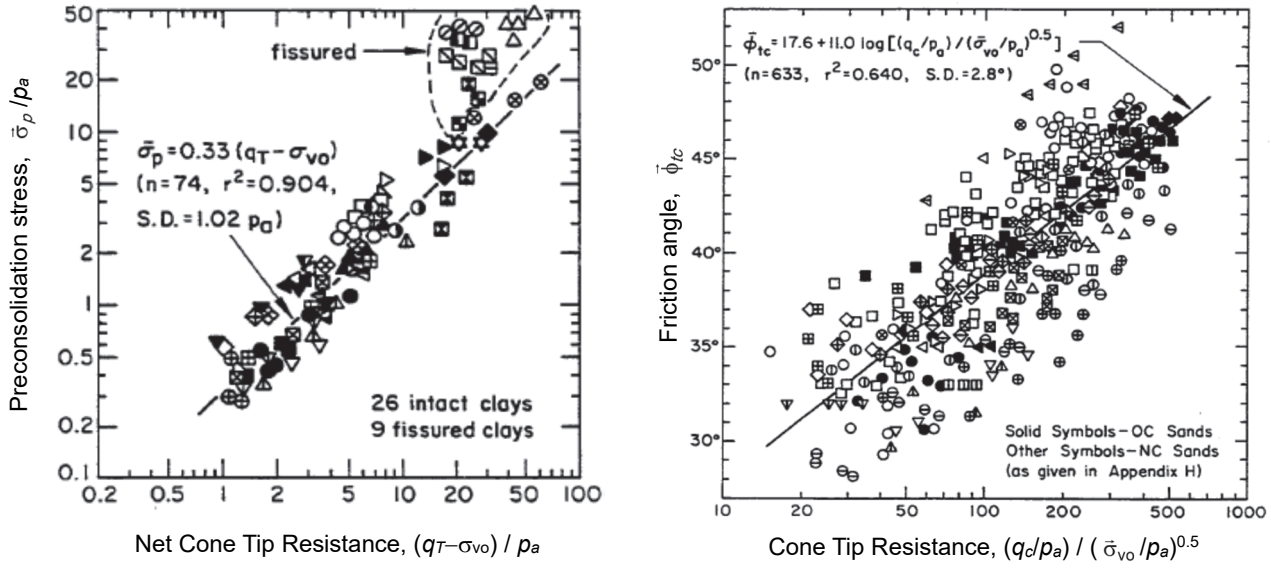


Fig. 1 Examples of transformation models in Kulhawy and Mayne (1990)

Table 1 Soil/rock parameter databases

	Database	Reference	Parameters of interest	# data points	# sites/ studies	% of completeness
Univariate	CLAY/16	Phoon and Kulhawy (1999b)	γ, γ_d, w_n, PL, LL, PI, LI, ϕ', s_u, s_u^{FV}, q_c, q_t, SPT-N, DMT (A, B), PMT p_L			
	SAND/11	Phoon and Kulhawy (1999b)	ϕ', D_r, q_c, SPT-N, DMT (A, B, I_D, K_D, E_D), PMT (p_L, E_{PMT})			
	ROCK/8	Prakoso (2002)	γ (or γ_d), n, R, S_h, σ_{bt}, I_s, σ_c, E			
	ROCK/13	Aladejare and Wang (2017)	ρ, G_s, I_{d2}, n, w_c, γ, R_L, S_h, σ_{bt}, I_{s50}, σ_c, E, ν			
Multivariate	CLAY/5/345	Ching and Phoon (2012)	LI, s_u, s_u^{re}, σ'_p, σ'_v	345	37 sites	100%
	CLAY/7/6310	Ching and Phoon (2013)	s_u from 7 different test procedures	6310	164 studies	17.7%
	CLAY/6/535	Ching *et al.* (2014)	s_u/σ'_v, OCR, q_{tc}, q_{tu}, $(u_2-u_0)/\sigma'_v$, B_q	535	40 sites	100%
	CLAY/10/7490	Ching and Phoon (2014a)	LL, PI, LI, σ'_v/P_a, σ'_p/P_a, s_u/σ'_v, S_t, q_{tc}, q_{tu}, B_q	7490	251 studies	34.1%
	FI-CLAY/7/216	D'Ignazio *et al.* (2016)	s_u^{FV}, σ'_v, σ'_p, w_n, LL, PL, S_t	216	24 sites	100%
	JS-CLAY/5/124	Liu *et al.* (2016)	M_r, q_c, f_s, w_n, γ_d	124	16 sites	100%
	JS-CLAY/7/372	Zou *et al.* (2017)	σ_v, σ'_v, q_{tc}, f_s/σ'_v, B_q, V_{s1}, s_u/σ'_v	372	25 sites	100%
	SAND/7/2794	Ching *et al.* (2017a)	D_{50}, C_u, D_r, σ'_v/P_a, ϕ', q_{t1}, $(N_1)_{60}$	2794	176 studies	60.0%
	EMI-ROCK/8/26000+	Kim and Hunt (2017)	σ_c, σ_{bt}, ρ, CAI, PPI, cohesion, direction shear, triaxial confining	26000+	–	–
	FG/5/1000 (FG: fine grain)	Kootahi and Moradi (2017)	e, w_n, LL, PI, C_c	1000	170 sites	100%
	ROCK/9/4069	Ching *et al.* (2018)	γ, n, R_L, S_h, σ_{bt}, I_{s50}, V_p, σ_{ci}, E_i	4069	184 studies	34.2%
	FG-KSAT/6/1358 (FG: fine grain)	Feng and Vardanega (2019)	e, k, LL, PL, PI, G_s	1358	33 studies	91.4%
	SH-CLAY/11/4051	Zhang *et al.* (2020)	LL, PI, LI, e, K_0, σ'_v/P_a, s_u/σ'_v, S_t, q_c/σ'_v	4051	50 sites	39.5%
	ROCKMass/9/5876	Ching *et al.* (2020)	RQD, RMR, Q, GSI, E_m, E_{em}, E_{dm}, E_i, σ_{ci}	5784	225 studies	29.3%

Note: ρ = density; ν = Poisson ratio; γ = unit weight; ϕ' = effective friction angle; σ'_p = preconsolidation stress; σ'_v = vertical effective stress; σ_{bt} = Brazilian tensile strength; σ_{ci} = uniaxial compressive strength of intact rock; γ_d = dry unit weight; $(N_1)_{60} = N_{60}/(\sigma'_v/P_a)^{0.5}$; $(u_2-u_0)/\sigma'_v$ = normalized excess pore pressure; B_q = pore pressure ratio = $(u_2-u_0)/(q_t-\sigma_v)$; CAI = Cerchar abrasivity index; C_c = compression index; C_u = coefficient of uniformity; D_{50} = median grain size; DMT (A, B, I_D, K_D, E_D) = dilatometer A & B readings, material index, horizontal stress index, modulus; D_r = relative density; e = void ratio; E_{dm} = dynamic modulus of rock mass; E_{em} = elasticity modulus of rock mass; E_i = Young's modulus of intact rock; E_m = deformation modulus of rock mass; f_s = sleeve frictional resistance; G_s = specific gravity; GSI = geological strength index; I_{d2} = slake durability index; I_s = point load strength index ($I_{s50} = I_s$ for diameter 50 mm); k = hydraulic conductivity; K_0 = at-rest lateral earth pressure coefficient; LI = liquidity index; LL = liquid limit; M_r = subgrade resilience modulus; n = porosity; N_{60} = corrected SPT-N; OCR = overconsolidation ratio; P_a = atmospheric pressure = 101.3 kPa; PI = plasticity index; PMT (p_L, E_{PMT}) = pressuremeter limit stress, modulus; PPI = punch penetration index; Q = Q-system; q_c = cone tip resistance; q_t = corrected cone tip resistance; $q_{t1} = (q_t/P_a)/(\sigma'_v/P_a)^{0.5}$; $q_{tc} = (q_t-\sigma_v)/\sigma'_v$ = normalized cone tip resistance; $q_{tu} = (q_t-u_2)/\sigma'_v$ = effective cone tip resistance; R = Schmidt hammer hardness (R_L = L-type Schmidt hammer hardness); RMR = rock mass rating; RQD = rock quality designation; S_h = Shore scleroscope hardness; SPT-N = standard penetration test blow count; S_t = sensitivity; s_u = undrained shear strength for clay; s_u^{FV} = field vane s_u; s_u^{re} = remolded s_u; u_0 = hydrostatic pore pressure; V_p = P-wave velocity; V_s = S-wave velocity; $V_{s1} = V_s(P_a/\sigma'_v)^{0.25}$; w_n = water content.

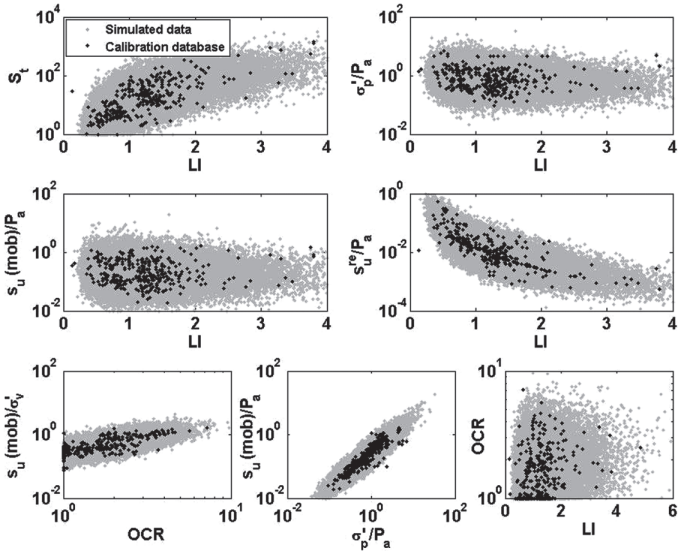


Fig. 2 Data points in CLAY/5/345 (grey dots are simulated data) (source: Ching and Phoon 2012)

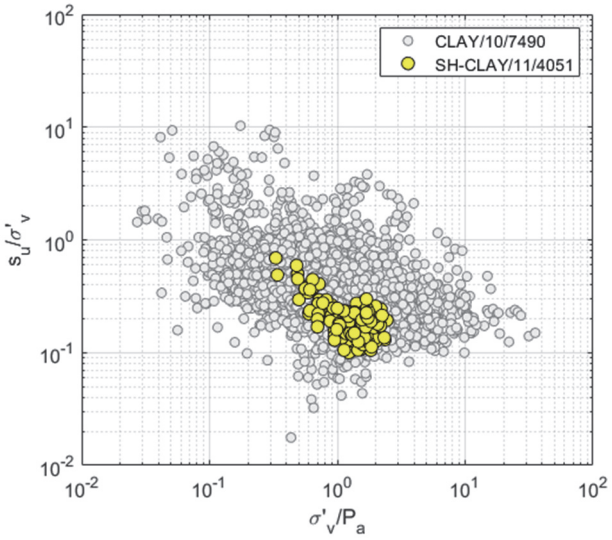


Fig. 3 s_u/σ'_v vs. σ'_v/P_a data in SH-CLAY/11/4051 (grey dots are data from CLAY/10/7490)

3. SITE-SPECIFIC ESTIMATION BASED ON SMALL DATA ONLY

This section addresses the question: "Can we solely rely on small data for site-specific decision making?" Consider the site investigation data (small data) of a Taipei site: Table 2 summarizes the site investigation data at several depths in a silty clay layer at the Taipei site. Figure 4 shows the CPT data at one sounding at the site. The small data is MUSIC-X; in particular, it is sparse (only a handful of depths are measured) and incomplete (there are missing entries in Table 2). This sparse and incomplete small data poses a significant challenge in data analysis, as traditional statistical methods usually require abundant and complete data.

3.1 Gibbs Sampler

Recently, Ching and Phoon (2019) proposed a Bayesian framework of analyzing MUSIC data. In this framework, the soil/rock properties ($\underline{Y}$) at a site are transformed to multivariate normal $\underline{X}$ with site-specific mean vector = $\underline{\mu}$ and site-specific covariance matric = $\mathbf{C}$. The Gibbs sampler (Geman and Geman 1984; Gilks *et al.* 1996) is adopted to estimate $\underline{\mu}$ and $\mathbf{C}$ and, at the meantime, to simulate the missing entries in Table 2 to resolve the issue of incomplete data. To further address the issue of sparse data, non-informative priors are adopted for $\underline{\mu}$ and $\mathbf{C}$. Because of the non-informative priors, the Bayesian method can properly reflect the large statistical uncertainty due to sparse data. Figure 5 shows the learning outcomes for a bivariate (n = 2) simulated example. The underlying distribution that generates the simulated data is a bivariate standard normal distribution. When there are abundant data (m = 50), Fig. 5(a) shows that the trained bivariate distribution is close to the underlying bivariate standard normal distribution. Nonetheless, when there are very sparse data (m = 2), Fig. 5(b) shows that the trained bivariate distribution is quite flat, reflecting the large statistical uncertainty. This large statistical uncertainty is desirable as there are only two data points.

3.2 Gibbs Sampler Considering Spatial Variability

The Bayesian framework proposed by Ching and Phoon (2019) does not address spatial variability. Ching and Phoon (2020) further extended the framework to address spatial variability (MUSIC-X data). This framework is adopted to learn the small data in Table 2 and Fig. 4. Only three s_u values in Table 2 are treated as known, whereas the other six s_u values (those in the parentheses) are treated as unknown for the purpose of validation. The multivariate model trained by the small data is referred to as a "site-specific multivariate model" in this paper.

The black dots in Fig. 6 show the scatter of the samples simulated by the trained site-specific multivariate model. The significant scatters indicate the large statistical uncertainty due to the sparse data. The model not only can simulate correlation samples among different clay parameters (cross-correlation samples) in Fig. 6, but it can also simulate the spatial variabilities of clay parameters (*i.e.,* conditional random fields). The dashed lines in Fig. 7

Table 2 Site investigation data for a silty clay layer at a Taipei site (source: Ou and Liao 1987)

Depth (m)	LL (%)	PI (%)	LI	σ'_v (kPa)	σ'_p (kPa)	s_u (kPa)	q_t (kPa)
12.8	30.1	9.1	1.20	128.0	172.7	46.9	894.3
14.8	32.8	12.8	1.43	144.9	–	(52.9)	881.2
16.1	36.4	14.5	1.24	155.6	–	(51.7)	933.9
17.8	41.9	18.9	0.90	169.9	181.8	42.8	1009.5
18.3	–	–	–	174.5	–	(59.3)	1252.9
20.2	38.1	17.3	0.70	190.0	–	(60.5)	1228.8
22.7	37.0	16.0	0.58	210.9	–	(64.4)	1417.9
24.0	38.0	16.2	0.75	221.7	221.7	67.5	1573.1
26.6	34.8	13.8	0.80	243.7	–	(82.1)	1779.9

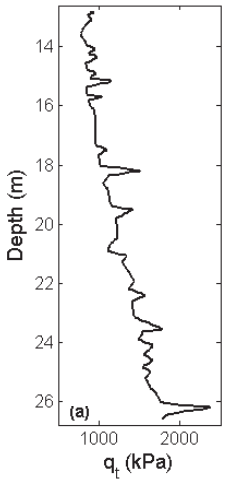
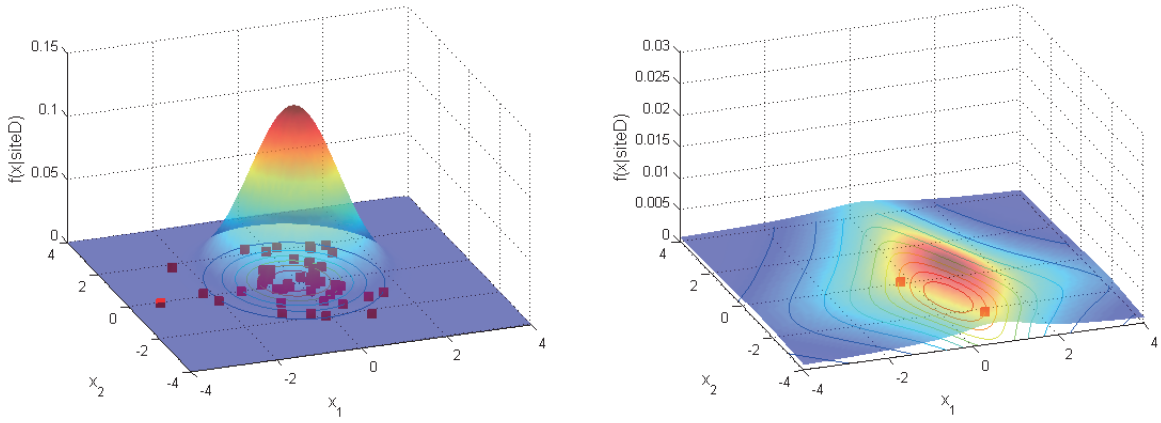


Fig. 4 CPT data of one sounding at the Taipei site



(a) m = 50 (b) m = 2

Fig. 5 Bivariate distribution trained by simulated data: (a) m = 50 and (b) m = 2

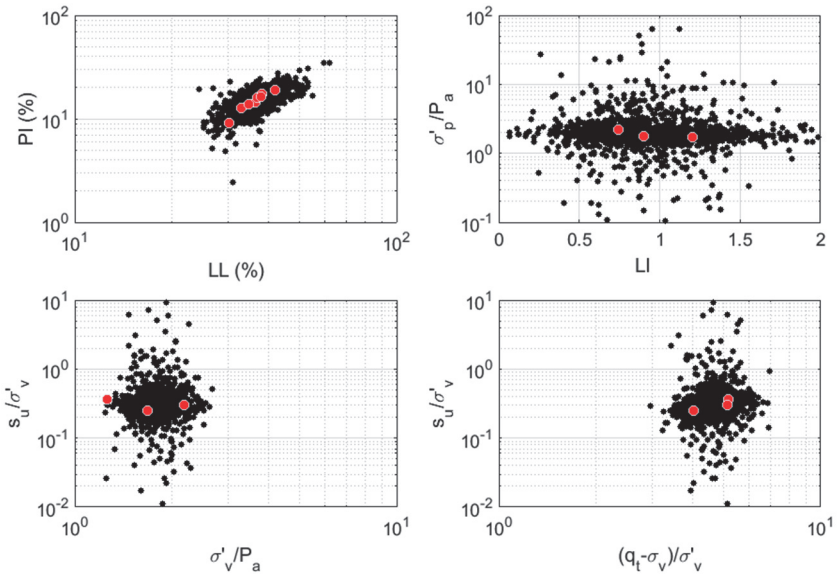


Fig. 6 Samples simulated by the trained site-specific multivariate model (red dots are the measured data)

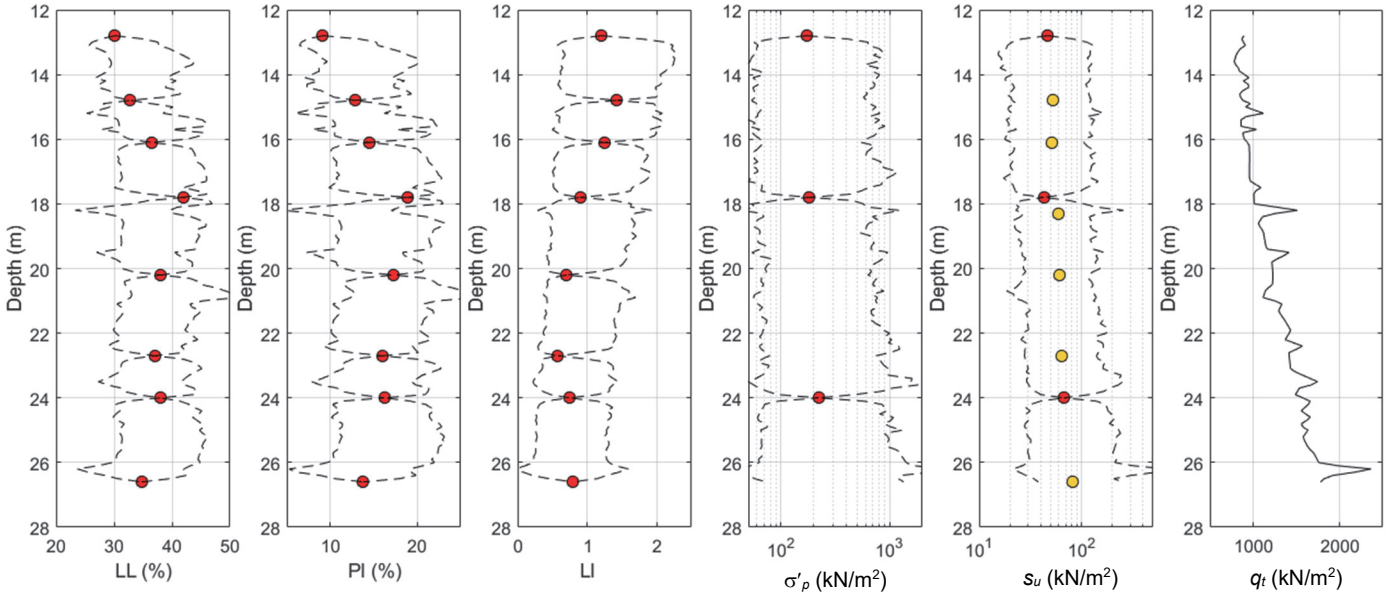


Fig. 7 95% confidence intervals for the conditional random field samples simulated by the trained site-specific multivariate model (red dots are the measured data; orange dots are the verification data)

shows the 95% confidence intervals (95% CIs) of the conditional random field samples. Note that the 95% CI has zero width at the depths where the soil parameter is measured. The orange dots in the s_u profile are the validation data (treated as unknown during learning). It is noteworthy that the 95% CI in Fig. 7 for s_u is very wide, as low as 10 kPa and as high as 200 kPa. This range is not very useful in practice because it spans the full range of soft, medium, stiff, and very stiff clays. The significant scatters in Fig. 6 and large 95% CIs in Fig. 7 both reflect the large statistical uncertainty due to the sparse data.

3.3 Summary

Can we solely rely on small data for site-specific decision making? Typically, NO! because in practice, small data is typically sparse and incomplete. Solely relying on small data may lead to a difficult situation where the resulting site-specific model has too large statistical uncertainty to support site-specific decision making.

4. USE OF BIG DATA IN SITE-SPECIFIC DECISION MAKING

Given that the site-specific model trained by small data may have large statistical uncertainty, one possible strategy of reducing statistical uncertainty is to incorporate BIG DATA. Note that BIG DATA is not sparse. For instance, CLAY/10/7490 in Table 1 has thousands of clay cases. The use of BIG DATA in site-specific estimation is not new. Generic transformation models such as those in Figure 1 are for this purpose. The fact that practical engineers have implemented such transformation models in daily

design indicates that there is certain value in BIG DATA for site-specific estimation.

4.1 Generic Multivariate Model

BIG DATA can be used to construct a generic transformation model as well as the 95% CI. For instance, the dark line in Fig. 8 shows the generic q_t-s_u model constructed by CLAY/10/7490, whereas the dashed lines indicate the generic 95% CI. Note that the generic 95% CI is quite wide, because it needs to accommodate generic cases with different clay types (*e.g.,* structured *vs.* non-structured), geology (marine *vs.* lacustrine), and geographic locations. In the literature, the uncertainty associated with a transformation model is called the transformation uncertainty (Phoon and Kulhawy 1999a), and the 95% CI in Fig. 8 reflects this transformation uncertainty.

The transformation model in Fig. 8 is bivariate (q_t *vs.* s_u). It is possible to construct a multivariate model using multivariate generic data (Ching and Phoon 2012, 2013, 2014b; Ching *et al.* 2014, 2017b, 2019), and such a model is referred to as a "generic multivariate model" in this paper. For the Taipei data in Table 2, the generic multivariate model for (LL, PI, LI, σ'_v, σ'_p, s_u, q_t) is constructed by CLAY/10/7490, and the 95% CI of s_u can be estimated by multiple information of (LL, PI, LI, σ'_v, σ'_p, q_t). The dashed line in Fig. 9(b) shows the 95% CI of s_u produced by the generic multivariate model, whereas Fig. 9(a) shows the 95% CI produced by the site-specific multivariate model (constructed by the small data in Table 2). It is clear that the 95% CI produced by the generic model is even wider than the 95% CI produced by the site-specific model. This wide 95% CI is not very useful in practice.

4.2 Hierarchical Bayesian Model

Recently, Ching *et al.* (2020) proposed a hierarchical Bayesian model (HBM) that can learn the site-specific statistics from BIG DATA. In CLAY/10/7490, the generic cases are from many sites. Cases from the same site are considered to belong to the same group by the HBM. Figure 10(a) re-plots the generic q_t-s_u cases in Fig. 8 but now with "site labels". The cases with the same label are from the same site. Each site has distinct site-specific statistics: the i-th site has its own site-specific mean vector and covariance matrix (in the X space), denoted by $\underline{\mu}_i$ and $\mathbf{C}_i$. Suppose that there are M sites in CLAY/10/7490.

The HBM proposed by Ching *et al.* (2020) has the model structure in Fig. 11. The mean vectors of different sites in CLAY/10/7490 ($\underline{\mu}_1$, $\underline{\mu}_2$, …, $\underline{\mu}_M$) are distinct but assumed to follow a common distribution governed by ($\underline{\mu}_0$, $\mathbf{C}_0$). Similarly, the covariance matrices of different sites ($\mathbf{C}_1$, $\mathbf{C}_2$, …, $\mathbf{C}_M$) are also assumed to follow a common distribution governed by (ν_0, $\mathbf{\Sigma}_0$). ($\underline{\mu}_0$, $\mathbf{C}_0$, ν_0, $\mathbf{\Sigma}_0$) are called the hyper-parameters. The HBM can learn the site-specific statistics in BIG DATA through these hyper-parameters. Figure 10(b) shows the behavior of the HBM trained by CLAY/10/

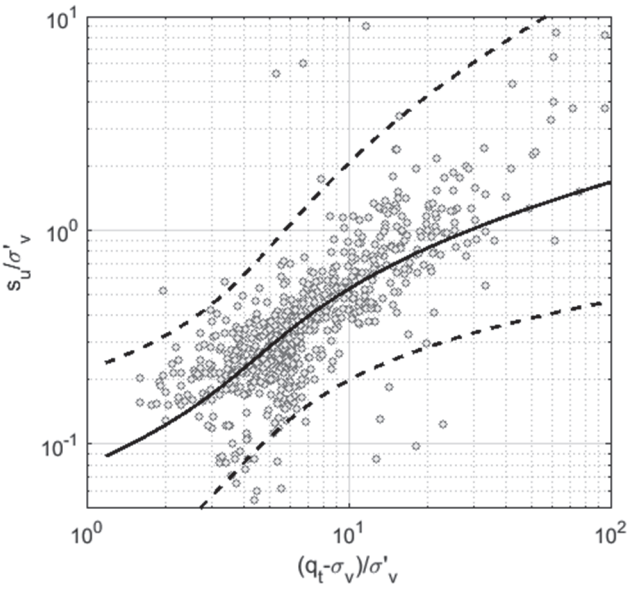
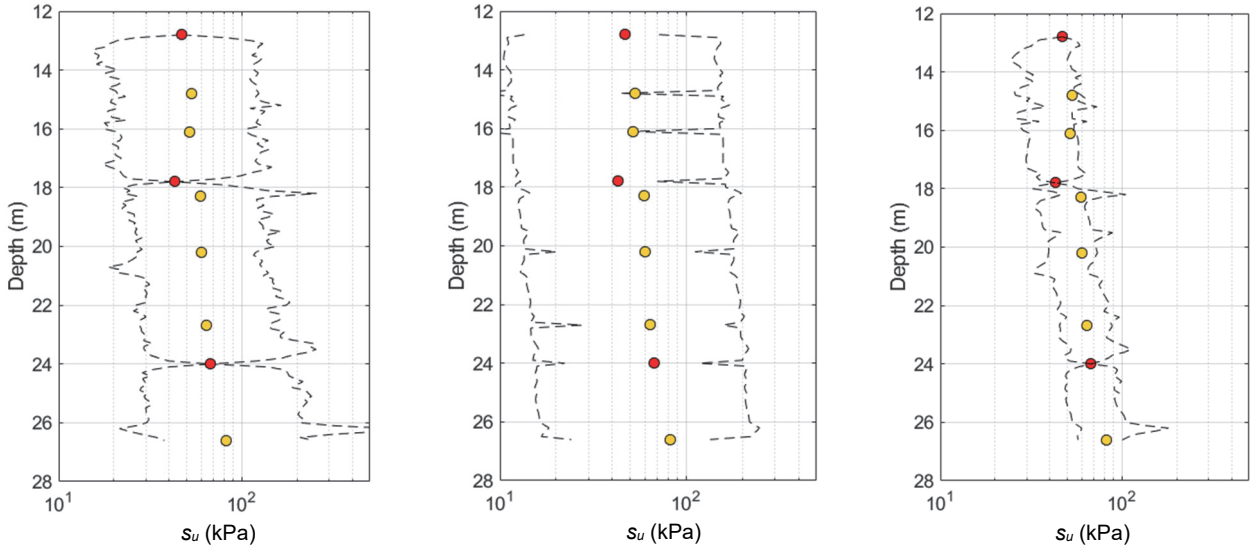
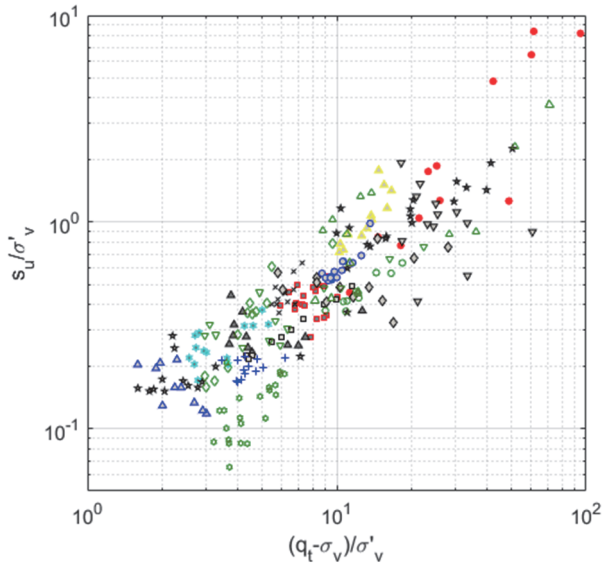


Fig. 8. q_t-s_u transformation model (grey dots are the data points in CLAY/10/7490)

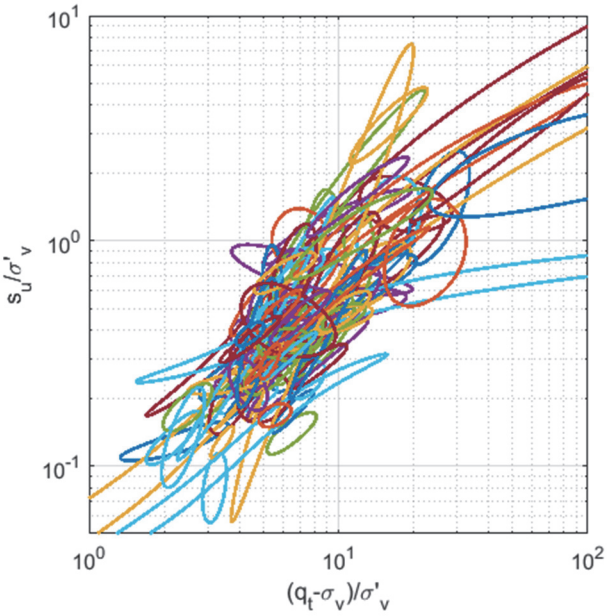


(a) Site-specific model (b) Generic model (c) Quasi-site-specific HBM

Fig. 9 95% confidence intervals for s_u based on (a) site-specific model, (b) generic model, and (c) quasi-site-specific HBM (red dots are the measured data; orange dots are the verification data)



(a) q_t-s_u cases with site labels



(b) Behavior of the trained HBM

Fig. 10 Generic q_t-s_u cases and the behavior of the trained HBM (in Fig. 10(a), markers with the same symbol are from the same site; in Fig. 10(b), the colored ellipses represent the simulated site-specific models)

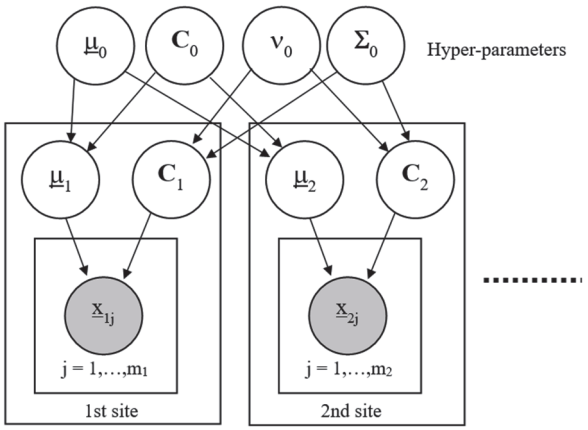


Fig. 11 Model structure of the HBM (source: Ching *et al.* 2020)

7490. The trained HBM can simulate ($\underline{\mu}_i$, $\mathbf{C}_i$) for a future site, and each (skewed) ellipse represents such a future site. Note that the center and shape of each ellipse in Fig. 10(b) mimic the centers and local correlation behaviors for the sites in CLAY/10/7490. The trained HBM in Fig. 10(b) is quite different from the trained generic model in Fig. 8: the trained HBM can mimic a future site, but the trained generic model can only provide the generic trend.

The trained HBM can be further conditioned by the small data in Table 2. By conditioning on the small data, most ellipses in Fig. 10(b) are incompatible to the small data, and only few ellipses are compatible. Figure 12 shows the incompatible ellipses in grey and compatible ellipses in colors. The red dots in Fig. 12 show the three cases in Table 2 with q_t-s_u information. Note that there are many cases in Table 2 with q_t information only (s_u is missing), and these cases are not shown in Fig. 12. The colored (compatible) ellipses illustrate the model after the conditioning. This model is referred to as a "quasi-site-specific HBM" in this paper, because it is first trained by CLAY/10/7490 and subsequently conditioned by the small data. This quasi-site-specific HBM is in strong contrast with the site-specific model trained by the small data (Section 3). It is also in strong contrast with the generic model trained by BIG DATA (Section 4.1). This quasi-site-specific HBM incorporates both BIG DATA and small data.

The quasi-site-specific HBM in Fig. 12 is bivariate (q_t *vs.* s_u). The HBM is also applicable to multivariate BIG DATA (CLAY/10/7490) and multivariate small data (Table 2). The resulting model is a quasi-site-specific multivariate HBM. This quasi-site-specific multivariate HBM can also simulate cross-correlation samples (Fig. 13) as well as spatial variability (conditional random fields) (Fig. 14). The quasi-site-specific HBM can be compared with the site-specific model by comparing Fig. 13 with Fig. 6 and by comparing Fig. 14 with Fig. 7. The general observation is that the quasi-site-specific HBM has significantly less statistical uncertainty than the site-specific model. Figure 9 compares the s_u estimation results based on the three models, site-specific model, generic model, and quasi-site-specific HBM. It is clear that the quasi-site-specific HBM produces the least uncertainty (narrowest 95% CI), yet the six validating s_u data (orange dots) are still within its narrow 95% CI.

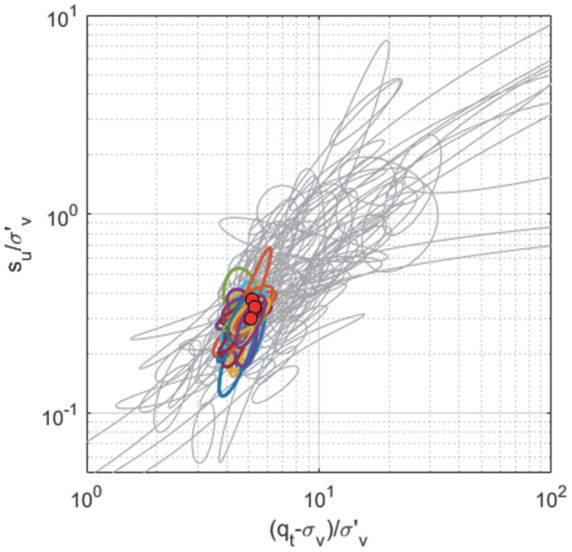


Fig. 12 Illustration of the quasi-site-specific HBM (the colored ellipses represent the compatible site-specific models)

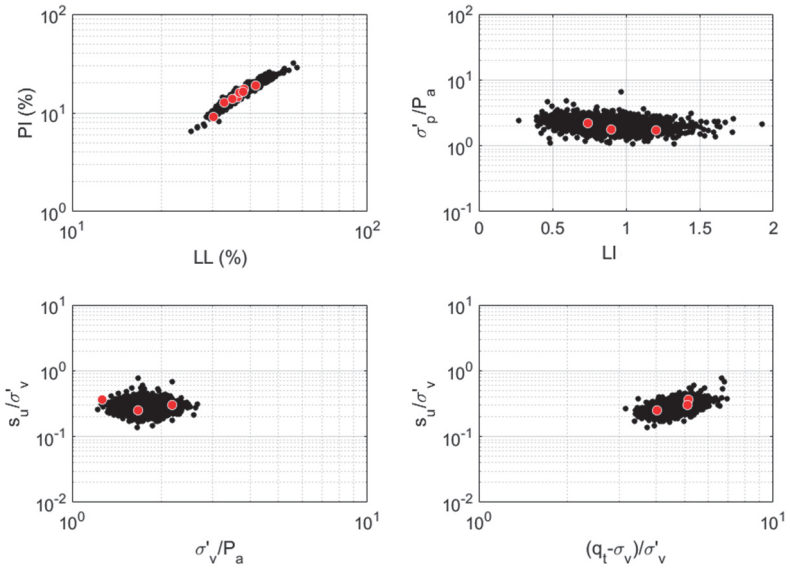


Fig. 13 Samples simulated by the quasi-site-specific HBM (red dots are the measured data)

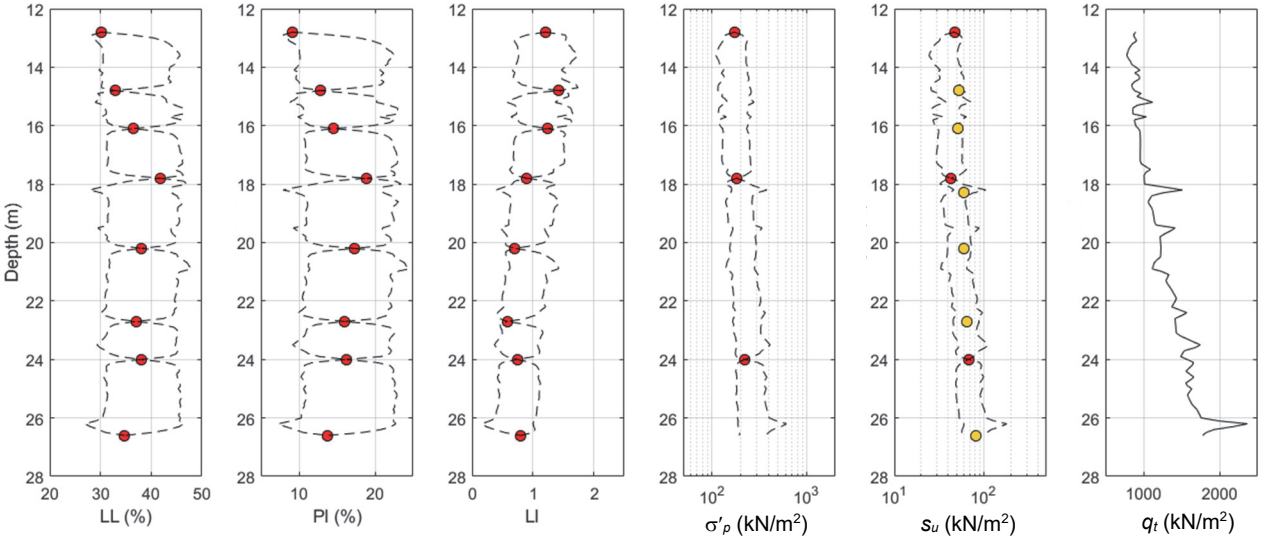


Fig. 14 95% confidence intervals for conditional random field samples (quasi-site-specific HBM) (red dots are the measured data; orange dots are the verification data)

4.3 Summary

Does geotechnical BIG DATA have value in site-specific decision making? In principle, YES! However, it requires a suitable model such as the hierarchical Bayesian model to effectively exploit the value of BIG DATA. If BIG DATA is implemented in an ineffective way (*e.g.,* merely to construct a generic model), BIG DATA may not produce value in site-specific decision making.

5. CONCLUSIONS

In this paper, some geotechnical soil/rock property databases are reviewed. These are called geotechnical BIG DATA, and the value of BIG DATA in supporting site-specific decision is demonstrated in this paper. It is first shown that small data (site investigation data at a local target site) is typically sparse and incomplete such that the site-specific model constructed by small data usually has large statistical uncertainty. The resulting 95% confidence interval (95% CI) for the design soil/rock parameter is usually quite wide and not very useful in practice. On the other hand, BIG DATA can be used to construct a generic model, but the generic model usually has large transformation uncertainty. This large transformation uncertainty also leads to wide 95% CI that may render the generic model not useful. In general, it is not trivial to effectively exploit value in BIG DATA to support site-specific decision making.

This paper demonstrates that in order to effectively exploit the value in BIG DATA to support site-specific decision making, a model with a suitable structure is needed. For this purpose, a hierarchical Bayesian model was proposed by the author to learn the site-specific statistics in BIG DATA. The resulting model is a quasi-site-specific model because the model is first trained by BIG DATA and subsequently conditioned by the small data. It is shown that the 95% CI produced by this quasi-site-specific model is significantly narrower than those produced by site-specific and generic models.

ACKNOWLEDGEMENTS

This work was presented in the 18th National Conference in Geotechnical Engineering, Taiwan as the 2020 TGS Geotechnical Lecture. The author would like to thank Taiwan Geotechnical Society for giving him the opportunity of presenting this work.

FUNDING

The author would like to thank the funding support by the Ministry of Science and Technology of Taiwan (107-2221-E-002-053-MY3 & 109-2221-E-002-029-MY3).

DATA AVAILABILITY

This study does not generate new data and/or new computer codes.

CONFLICT OF INTEREST STATEMENT

The author certifies that there is no conflict of interest for this work.

REFERENCES

Aladejare, A.E. and Wang, Y. (2017). "Evaluation of rock property variability." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, **11**(1), 22-41. https://doi.org/10.1080/17499518.2016.1207784

Ching, J. and Phoon, K.K. (2012). "Modeling parameters of structured clays as a multivariate normal distribution." *Canadian Geotechnical Journal*, **49**(5), 522-545. https://doi.org/10.1139/t2012-015

Ching, J. and Phoon, K.K. (2013). "Multivariate distribution for undrained shear strengths under various test procedures." *Canadian Geotechnical Journal*, **50**(9), 907-923. https://doi.org/10.1139/cgj-2013-0002

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. https://doi.org/10.1139/cgj-2013-0262

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

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

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

Ching, J., 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. https://doi.org/10.1139/cgj-2017-0537

Ching, J., Lin, G.H., Chen, J.R., and Phoon, K.K. (2017a). "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. https://doi.org/10.1139/cgj-2016-0318

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

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. https://doi.org/10.1139/cgj-2012-0259

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

Ching, J., Phoon, K.K., Li, K.H., and Weng, M.C. (2019). "Multivariate probability distribution for some intact rock properties." *Canadian Geotechnical Journal*, **56**(8), 1080-1097. https://doi.org/10.1139/cgj-2018-0175

Ching, J., Wu, S., and Phoon, K.K. (2020). "Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model." *Journal of Engineering Mechanics*, ASCE (in review).

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. https://doi.org/10.1139/cgj-2016-0037

Feng, S. and Vardanega, P.J. (2019). "A database of saturated hydraulic conductivity of fine-grained soils: probability density functions." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, **13**(4), 255-261. https://doi.org/10.1080/17499518.2019.1652919

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741. https://doi.org/10.1109/TPAMI.1984.4767596

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

Kim, E. and Hunt, R. (2017). "A public website of rock mechanics database from Earth Mechanics Institute (EMI) at Colorado School of Mines (CSM)." *Rock Mechanics and Rock Engineering*, **50**(12), 3245-3252. https://doi.org/10.1007/s00603-017-1292-1

Kootahi, K. and Moradi, G. (2017). "Evaluation of compression index of marine finegrained soils by the use of index tests." *Marine Georesources and Geotechnology*, **35**(4), 548-570. https://doi.org/10.1080/1064119X.2016.1213775

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

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. https://doi.org/10.1016/j.enggeo.2016.05.018

Ou, C.Y. and Liao, J.T. (1987). *Geotechnical Engineering Research Report*, GT96008, National Taiwan University of Science and Technology, Taipei.

Phoon, K.K. (2020). "The story of statistics in geotechnical engineering." *Georisk: Assessment and Management of Risk for*

Engineered Systems and Geohazards, **14**(1), 3-25. https://doi.org/10.1080/17499518.2019.1700423

Phoon, K.K. and Kulhawy, F.H. (1999a). "Evaluation of geotechnical property variability." *Canadian Geotechnical Journal*, **36**(4), 625-639. https://doi.org/10.1139/t99-039

Phoon, K.K. and Kulhawy, F.H. (1999b). "Characterization of geotechnical variability." *Canadian Geotechnical Journal*, **36**(4), 612-624. https://doi.org/10.1139/t99-038

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

Prakoso, W.A. (2002). *Reliability-based Design of Foundations on Rock Masses for Transmission Line and Similar Structures*. Ph.D. Dissertation, Cornell University, Ithaca, NY.

Wikipedia (2020a). https://en.wikipedia.org/wiki/Big_data

Wikipedia (2020b). https://en.wikipedia.org/wiki/Dark_data

Zhang, D.M., Zhou, Y., Phoon, K.K., and Huang, H.W. (2020). "Multivariate probability distribution of Shanghai clay properties." *Engineering Geology*, **273**, 105675. https://doi.org/10.1016/j.enggeo.2020.105675

Zhang, L. (2016). *Engineering Properties of Rocks*, 2nd Edition. Elsevier Ltd., Cambridge, MA, U.S.A. https://doi.org/10.1016/C2014-0-02645-7

Zou, H., Liu, S., Cai, G., Puppala, A.J., and Bheemasetti, T.V. (2017). "Multivariate correlation analysis of seismic piezocone penetration (SCPTU) parameters and design properties of Jiangsu quaternary cohesive soils." *Engineering Geology*, **228**, 11-38. https://doi.org/10.1016/j.enggeo.2017.07.005