Journal of GeoEngineering, Vol. 18, No. 4, pp. 215-223, December 2023
http://dx.doi.org/10.6310/jog.202312_18(4).5

DATA-CENTRIC QUASI-SITE-SPECIFIC PREDICTION FOR SOIL MODULUS

Jianye Ching [1*] and Ming-Chieh Kuo [2]

ABSTRACT

This paper compiles a new soil property database called SOIL-DMT/8/7186 that contains 8 parameters of 7,186 soil records from 701 sites worldwide, including the modulus parameter obtained by the dilatometer test (DMT). With this database, the paper demonstrates how to construct a quasi-site-specific model that can predict the soil modulus for a target site (the Bothkennar test site, UK). First, the hierarchical Bayesian model (HBM) developed by the first author is used to learn the site-specific characteristics of the 701 sites in the database. The learned HBM can produce a prior model for the target site. Then, this prior model is further updated by sparse target-site data into a (posterior) quasi-site-specific model. The quasi-site-specific model can then be used to predict the soil modulus of the target site based on inexpensive site investigation data such as SPT blow count and CPT cone tip resistance. A remarkable observation is that the effectiveness of the quasi-site-specific model generally depends on the soil property database: the quasi-site-specific model based on a database more relevant to the target site can produce a narrower 95% confidence interval for the soil modulus. Moreover, it is found that a small database that contains only a few sites relevant to the target site may still perform well if these sites are properly chosen.

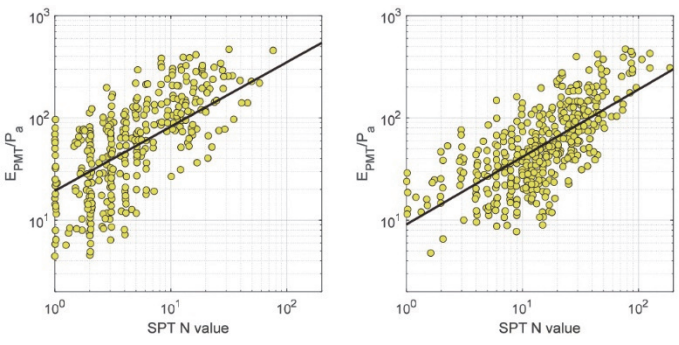
Key words: Probabilistic site characterization; soil modulus; SOIL-DMT/8/7186; hierarchical Bayesian model (HBM); data-centric geotechnics.

1. INTRODUCTION

The serviceability limit state in geotechnical design concerns about the deformation of a geotechnical structure. The soil properties that affect the deformation of a geotechnical structure include the soil modulus (such as Young's modulus) and compression index (C_c). The data-centric quasi-site-specific prediction for C_c has been investigated by Ching *et al.* (2022). The term "data-centric prediction" refers to any prediction methods that solely rely on data, which can include the site-specific investigation data (*e.g.,* boreholes, cone penetration tests, vane shear tests, *etc.* conducted at the target site) and/or non-site-specific (generic) data from a database. The current paper focuses on the data-centric quasi-site-specific prediction of soil modulus.

The direct determination of soil modulus usually requires undisturbed samples and costly laboratory tests. In practice, it is common to adopt a transformation model (Phoon and Kulhawy 1999) derived from generic (non-site-specific) datasets to correlate inexpensive site investigation data, *e.g.,* standard penetration test (SPT) blow count (N value), to soil modulus. However, such a generic transformation model is not exact and usually has significant transformation uncertainty. For instance, Fig. 1 shows the generic transformation models between SPT N value and the soil modulus obtained from the pressuremeter test (denoted by E_{PMT}) for clays and sands. The solid lines in the figure are the mean trends of the transformation models proposed by Ohya *et al.* (1982). The scatter of the data points around the mean trends is large, indicating that the transformation uncertainty is significant. The large scatter in Fig. 1 may be due to the fact that the generic transformation models were developed by data from different sites. Each site may have its unique site-specific local trend, so it is imprecise to fit all data using a single generic trend. Ching *et al.* (2021a, 2021b, 2022) showed that the transformation uncertainty of a site-specific transformation model can be significantly less than that of a generic transformation model. However, a site-specific transformation model requires site-specific data, which are usually sparse. Due to the data sparsity, a site-specific transformation model often has significant statistical uncertainty.

The statistical uncertainty of a site-specific model for a target site may be reduced by considering the past experiences from other sites. If the target site is not an anomaly, it is expected that the characteristics of the site-specific model of the target site (*e.g.,* the intersect and gradient of the mean trend, degree of scatter around



(a) clay (b) sand

Fig. 1 Transformation models between E_{PMT}/P_a (P_a = 101.3 kPa = one atmosphere pressure) and SPT N value (data points from Japan, source: Ohya *et al.* 1982)

Manuscript received May 25, 2023; revised August 12, 2023; accepted August 31, 2023.
[1*] Professor (corresponding author), Dept of Civil Engineering, National Taiwan University, Taiwan (e-mail: jyching@gmail.com).
[2] Graduate Student, Dept of Civil Engineering, National Taiwan University, Taiwan.

the mean trend, *etc.*) should be also not anomalies compared to other sites. As a result, by learning the characteristics from other sites, one can get a rough idea what the characteristics of the target site are. For this purpose, Ching *et al.* (2021a) developed the hierarchical Bayesian model (HBM) to learn the characteristics of sites in a soil property database. The HBM learning outcome can be used to construct a "prior model" for the target site, and this prior model can be further updated into a "posterior model" by the target-site data. The resulting posterior model can then be used to predict the soil modulus of the target site. This posterior model is not entirely site-specific because it not only depends on the target site but also depends on the prior model constructed based on the database: the posterior model is "quasi-site-specific". The purpose of the current paper is to demonstrate the process of constructing such a quasi-site-specific model.

First, a soil property database named SOIL-DMT/8/7186 is compiled in the current paper. This new soil property database contains 8 parameters of 7,186 soil records from 701 sites worldwide, including 2 index properties: liquidity index (LI) and median grain size (D_{50}), and 6 in-situ test parameters: undrained shear strength for the vane shear test ($s_{u,\mathrm{VST}}$; VST stands for "vane shear test"), corrected blow count (N_{60}; corrected for energy ratio of 60%) for the standard penetration test (SPT), corrected cone tip resistance (q_t; corrected for pore pressure) for the cone penetration test (CPT), CPT pore pressure coefficient (B_q), soil modulus obtained by the pressuremeter test (E_{PMT}; PMT stands for "pressuremeter test"), and soil modulus obtained by the dilatometer test (E_{DMT}; DMT stands for "dilatometer test"). Among the 8 parameters, E_{PMT} and E_{DMT} are soil moduli, and the remaining 6 parameters have certain correlations to the soil moduli. However, the soil modulus data are mainly from DMT, not from PMT.

Then, the data-centric quasi-site-specific prediction for soil modulus is demonstrated. The target site for demonstration is the Bothkennar test site (UK) (Nash *et al.* 1992). First, the HBM is adopted to learn the characteristics of the 701 sites in SOIL-DMT/8/7186. Then, a prior model is constructed based on the learned HBM, and this prior model is further updated into the (posterior) quasi-site-specific model by the target-site data. This quasi-site-specific model can then predict the soil modulus of the Bothkennar site. The same database SOIL-DMT/8/7186 is also adopted to develop a generic transformation model to predict the soil modulus of the Bothkennar site. The effectiveness of the quasi-site-specific and generic models can then be compared.

Finally, it is shown that the effectiveness of the quasi-site-specific model generally depends on the soil property database that trains the HBM. It is shown that a database relevant to the target site is more preferable than an irrelevant one (*e.g.,* if the target site is a clay site, a clay database is more preferable than a sand database). More interestingly, it is observed that the quasi-site-specific model based on a small database with only a few sites (*e.g.,* 10+ sites) may still perform well compared to the one based on a large database with hundreds of sites if the sites in the small database are properly chosen. The implication of this new observation (the quasi-site-specific model based on a small database can perform well) will be discussed.

2. SOIL-DMT/8/7186 DATABASE

This study compiles a new database (SOIL-DMT/8/7186) from the literature for 8 soil properties, including liquidity index (LI), median grain size (D_{50}), VST undrained shear strength ($s_{u,\mathrm{VST}}$), SPT blow count (N_{60}), CPT cone tip resistance (q_t), CPT pore pressure coefficient (B_q), PMT modulus (E_{PMT}), and DMT modulus (E_{DMT}). This database contains 7,186 records from 701 sites worldwide. All records are for in-situ natural soils (no laboratory reconstituted soils). The SOIL-DMT/8/7186 database is extracted from the datasets in Kuo (2020). The 8 soil properties are categorized into 3 categories:

1. Index properties (LI, D_{50}). For the records in SOIL-DMT/8/7186, 74% of them are clayey soils and 26% of them are granular soils. For clay records, LI values are usually reported, but D_{50} values are not reported in most cases. In contrast, for granular soil records, LI values are typically unknown, and D_{50} values are sometimes reported.
2. VST, SPT, and CPT parameters ($s_{u,\mathrm{VST}}$, N_{60}, q_t, B_q). For the clay records in SOIL-DMT/8/7186, 56% of them have $s_{u,\mathrm{VST}}$ information. For the granular soil records, the $s_{u,\mathrm{VST}}$ information is not available. 19% of all soil records in SOIL-DMT/8/7186 have SPT N_{60} information. The SPT N value is corrected to an energy ratio of 60% whenever possible (*e.g.,* the energy ratio for Japan records is assumed to be 75%, so $N_{60} = 1.25 \times N$ for Japan records). 45% of all soil records have CPT q_t information, where $q_t = q_c + (1 - a)u_2$ is the cone tip resistance corrected for pore pressure, q_c is the uncorrected cone tip resistance, u_2 is the behind-cone pore water pressure, and a is the area ratio of the cone. Only 20% of all soil records have CPT B_q information, because B_q is typically not reported for granular soils.
3. Soil moduli (E_{PMT}, E_{DMT}). Soil modulus can be obtained by the pressuremeter test (PMT) (*e.g.,* Baguelin *et al.* 1978) and by the dilatometer test (DMT) (Marchetti 1980). 41% of all soil records in SOIL-DMT/8/7186 have E_{DMT} information, whereas only 7% have E_{PMT} information. Namely, the soil modulus data in SOIL-DMT/8/7186 are mainly from DMT. The pressuremeter modulus (E_{PMT}) quantifies the in-situ horizontal modulus measured during the expansion of the PMT cylindrical probe. It is customarily assumed that E_{PMT} roughly represents the Young's modulus (E) of the soil: $E \approx E_{\mathrm{PMT}}$ (Kulhawy and Mayne 1990). The dilatometer modulus (E_{DMT}) quantifies the in-situ horizontal modulus measured during the inflation of the DMT membrane. It is customarily assumed that E_{DMT} is related to the Young's modulus by $E \approx E_{\mathrm{DMT}} \times (1 - \nu^2)$ (Kulhawy and Mayne 1990), where ν is the Poisson ratio of the soil. For clays, the drainage condition during PMT or DMT is generally unknown (Kulhawy and Mayne 1990), so it is uncertain whether E_{PMT} (or E_{DMT}) represents drained or undrained modulus. Also, it is uncertain whether or not ν for clays should be taken as 0.5 (undrained). In the current paper, ν is assumed to be 0.5 for clays and 0.2 for granular soils. It is noteworthy that the soil modulus generally decreases with increasing strain, and the strain levels for PMT and DMT are different. The average strain level for DMT is smaller than that for PMT (Mayne 2001). The former (the average strain level for DMT) is in the middle of typical strain range for the deformation analysis in foundation design (Mayne 2001).

Each soil record in SOIL-DMT/8/7186 is stored as one row in an Excel worksheet. There are 7,186 rows. The 8 soil properties are usually not completely observed for each record, *i.e.,* there are usually empty entries. If the database is complete, it contains 7,186

× 8 = 57,488 full entries. SOIL-DMT/8/7186 only contains 22,651 full entries, so 72% of them are incomplete. The basic marginal statistics of the 8 soil properties are listed in Table 1. In this table, E_{DMT} is multiplied by $(1-\nu^2)$ to represent E (Young's modulus) because $E \approx E_{DMT} \times (1-\nu^2)$. In contrast, E_{PMT} is directly adopted to represent E because $E \approx E_{PMT}$. All strength and modulus parameters are normalized by one atmosphere pressure (P_a = 101.3 kPa). Note that the cone tip resistance (q_t) is subtracted by the total vertical stress (σ_v) to obtain the net cone tip resistance. For sand records, σ_v and u_2 are usually negligible compared to q_c, so $(q_t - \sigma_v) \approx q_c$ when σ_v and/or u_2 information is not available. Table 2 shows the numbers of records with bivariate information. Note that the statistics in Tables 1 and 2 are not for a specific site but for the entire database, *i.e.,* data from multiple sites are pooled. Table 3 shows the group-level statistics. One "data group" constitutes one site or one project. To obtain meaningful second-order group-level statistics, only data groups with no less than 5 records are considered.

Table 1 Marginal statistics of SOIL-DMT/8/7186 at the generic level

Parameter	Number of records	Mean	COV*	Min	Max
LI	3347	1.03	0.71	–2.17	7.0
D_{50} (mm)	406	0.29	3.77	0.01	11.0
$s_{u,VST}/P_a$	2984	0.32	0.89	0.002	3.62
N_{60}	1354	18.71	0.89	0.99	124.1
$(q_t - \sigma_v)/P_a$	3198	28.22	1.36	0.38	378.2
B_q	1408	0.50	0.51	–0.10	1.56
E_{PMT}/P_a	514	280.44	1.08	7.73	1888.7
$(1-\nu^2) \times E_{DMT}/P_a$	2966	174.06	1.09	1.12	1733.2

* COV stands for the coefficient of variation.

2.1 Cross Correlations Among Parameters

Figures 2-5 show the cross-correlation plots between some soil parameters in SOIL-DMT/8/7186. In Figs. 2-4, it is clear that there is a positive correlation between the soil modulus and other in-situ test parameters such as N_{60}, $(q_t - \sigma_v)$, and $s_{u,VST}$. Although the soil modulus and N_{60} are positively correlated in Fig. 2, the scatter is large. This is expected because SPT N values are subjected to large variabilities (Kulhawy and Mayne 1990). The E_{PMT}/P_a *vs.* SPT N transformation models developed by Ohya *et al.* (1982) for clays and sands are shown in Fig. 2(a) to compared with the data in SOIL-DMT/8/7186. These transformation models were calibrated by Japanese data. The E_{PMT}/P_a *vs.* N_{60} data in SOIL-DMT/8/7186 mostly lie above these Japan-based transformation models.

Table 1 The numbers of records with bivariate information

	LI	D_{50}	$s_{u,VST}/P_a$	N_{60}	$(q_t - \sigma_v)/P_a$	B_q	E_{PMT}/P_a	$(1-\nu^2)\times E_{DMT}/P_a$
LI	3347	–	2447	230	1272	984	64	512
D_{50}		406	–	161	314	–	21	358
$s_{u,VST}/P_a$			2984	45	718	626	0	252
N_{60}				1354	652	125	342	662
$(q_t - \sigma_v)/P_a$					3198	1406	126	1796
B_q						1408	22	443
E_{PMT}/P_a							514	55
$(1-\nu^2) \times E_{DMT}/P_a$								2966

Table 2 Summary statistics of SOIL-DMT/8/7186 at the group level

Property	No. of data groups	No. of records /group		Property mean value		Property COV	
		Range	Mean	Range	Mean	Range	Mean
LI	256	5 ~ 49	12.0	–0.49 ~ 3.54	1.03	0.02 ~ 6.49	0.37
D_{50} (mm)	36	5 ~ 31	10.5	0.012 ~ 0.67	0.21	0.07 ~ 0.62	0.36
$s_{u,VST}/P_a$	240	5 ~ 50	12.0	0.034 ~ 2.98	0.34	0.03 ~ 1.82	0.38
N_{60}	102	5 ~ 57	12.2	1.56 ~ 73.2	16.95	0.08 ~ 1.53	0.43
$(q_t - \sigma_v)/P_a$	231	5 ~ 50	12.8	1.50 ~ 170.4	29.24	0.07 ~ 1.61	0.44
B_q	97	5 ~ 46	13.5	–0.05 ~ 1.05	0.49	0.06 ~ 2.75	0.36
E_{PMT}/P_a	31	5 ~ 57	14.2	13.2 ~ 902.0	289.38	0.11 ~ 1.07	0.54
$(1-\nu^2) \times E_{DMT}/P_a$	216	5 ~ 50	12.9	4.62 ~ 671.6	167.23	0.08 ~ 1.21	0.46

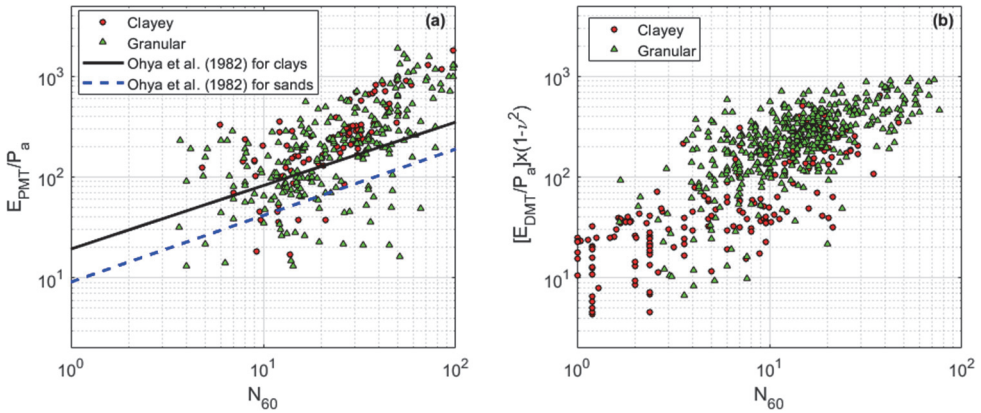


Fig. 2 Soil modulus *vs.* N_{60}: (a) E_{PMT}/P_a *vs.* N_{60}; (b) $(E_{DMT}/P_a) \times (1-\nu^2)$ *vs.* N_{60}

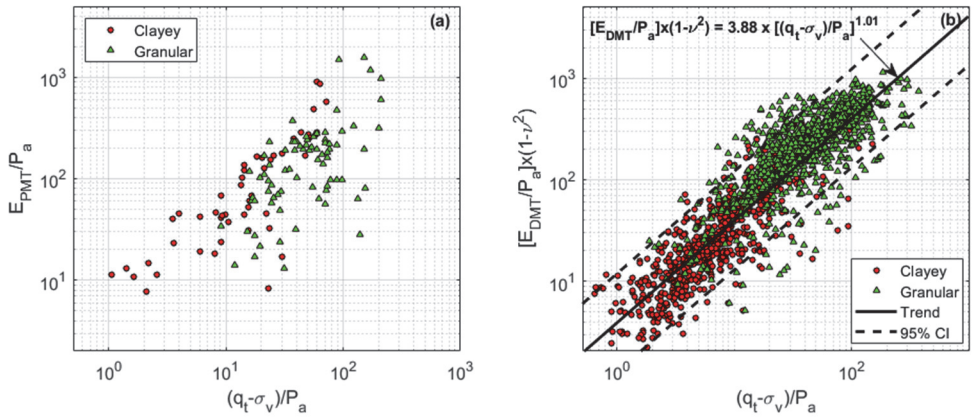


Fig. 3 Soil modulus *vs.* $(q_t - \sigma_v)/P_a$: (a) E_{PMT}/P_a *vs.* $(q_t - \sigma_v)/P_a$; (b) $(E_{DMT}/P_a) \times (1 - \nu^2)$ *vs.* $(q_t - \sigma_v)/P_a$

In contrast to the large scatter for modulus *vs.* N_{60}, the scatters for the relationships of modulus *vs.* $(q_t - \sigma_v)$ and modulus *vs.* $s_{u,VST}$ are smaller (Figs. 3 and 4). It is remarkable that modulus *vs.* in-situ parameter for clayey and granular soils seem to follow a unique continuous trend. This is quite evident in Fig. 3(b), where the $E_{DMT} \times (1 - \nu^2)$ *vs.* $(q_t - \sigma_v)$ relationship for clays seems to follow a unique continuous trend with that for granular soils. For clays, there seems to be a negative correlation between the modulus ratio [defined as (soil modulus)/(in-situ test parameter)] and the liquidity index (LI). The negative correlation is quite evident between $[E_{DMT} \times (1 - \nu^2)]/(q_t - \sigma_v)$ and LI in Fig. 5(b).

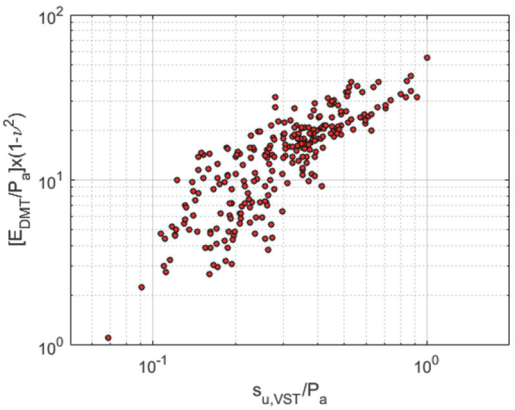


Fig. 4 $(E_{DMT}/P_a) \times (1 - \nu^2)$ *vs.* $s_{u,VST}/P_a$ for clayey soil records

3. CONSTRUCTION OF QUASI-SITE-SPECIFIC MODEL

3.1 Generic Transformation Model

The SOIL-DMT/8/7186 database can be used to construct generic transformation models for soil modulus. For instance, the data in Fig. 3(b) can be used to construct a generic $(E_{DMT}/P_a) \times (1 - \nu^2)$ *vs.* $(q_t - \sigma_v)/P_a$ model. The mean trend and 95% confidence interval (95% CI) of this model are shown in Fig. 3(b). The 95% CI is fairly wide: the coefficient of variation (COV) of the transformation uncertainty is about 60%. This large transformation uncertainty is due to the overlook of site uniqueness. Records from different sites are lumped to develop the generic model. In reality, each site may have its own site-specific model. In principle, a site-specific model is more relevant to design than a generic model, but as mentioned earlier, a site-specific model may suffer from significant statistical uncertainty due to sparse site-specific data. When there are limited site-specific data, which is usually the case in geotechnical design, design engineers may adopt a generic model with the cost of large transformation uncertainty. There is a dilemma: a generic model usually has large transformation uncertainty, whereas a site-specific model may have large statistical uncertainty. Nonetheless, the quasi-site-specific model presented in the following sections can overcome this dilemma. The quasi-site-specific requires the hierarchical Bayesian model (HBM) proposed by Ching *et al.* (2021a). This HBM is briefly reviewed in the section below.

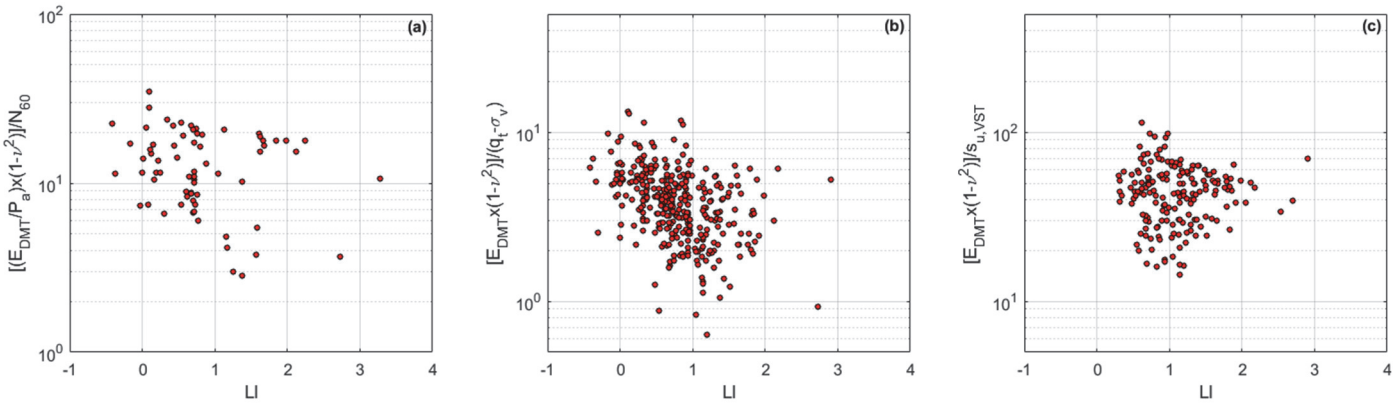


Fig. 5 Modulus ratio *vs.* LI for clayey soil records: (a) $[(E_{DMT}/P_a) \times (1 - \nu^2)]/N_{60}$ *vs.* LI; (b) $[E_{DMT} \times (1 - \nu^2)]/(q_t - \sigma_v)$ *vs.* LI; (c) $[E_{DMT} \times (1 - \nu^2)]/s_{u,VST}$ *vs.* LI

3.2 Hierarchical Bayesian Model

Ching *et al.* (2021a) proposed the hierarchical Bayesian model to learn the characteristics of sites in a soil property database. Figure 6 shows the model structure of the HBM. The (transformed) soil parameters for the j-th record at the i-th site in a database are denoted by the vector $\underline{X}_{ij} \in \mathbf{R}^{n\times 1}$ (n is the soil parameters of interest; n = 8 for SOIL-DMT/8/7186). The vector $\underline{X}_{ij}$ is assumed to follow a multivariate normal probability density function (PDF) with site-specific mean vector = $\underline{\mu}_i \in \mathbf{R}^{n\times 1}$ and site-specific covariance matrix = $\mathbf{C}_i \in \mathbf{R}^{n\times n}$, denoted by N($\underline{x}$; $\underline{\mu}_i$, $\mathbf{C}_i$). The parameters ($\underline{\mu}_i$, $\mathbf{C}_i$) quantify the "intra-site" variability within the i-th site. Because of site uniqueness, $\underline{\mu}_i \neq \underline{\mu}_j$ and $\mathbf{C}_i \neq \mathbf{C}_j$ if $i \neq j$, *i.e.,* there is also "inter-site" variability. In the HBM, $\{\underline{\mu}_i: i = 1, \ldots, n_s\}$ (n_s is the number of sites) are assumed to follow the same multivariate normal PDF N($\underline{\mu}$; $\underline{\mu}_0$, $\mathbf{C}_0$), where $\underline{\mu}_0 \in \mathbf{R}^{n\times 1}$ is the (hyper) mean vector and $\mathbf{C}_0 \in \mathbf{R}^{n\times n}$ is the (hyper) covariance matrix. Similarly, $\{\mathbf{C}_i: i = 1, \ldots, n_s\}$ are assumed to follow the same inverse-Wishart (IW) distribution (James 1964), denoted by IW($\mathbf{C}$; $\mathbf{\Sigma}_0$, ν_0), where $\mathbf{\Sigma}_0 \in \mathbf{R}^{n\times n}$ is the (hyper) scale matrix and $\nu_0 \in \mathbf{R}$ is the (hyper) degree of freedom. The parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0), called the hyper-parameters of the HBM, quantify both inter-site and intra-site variabilities in the soil property database.

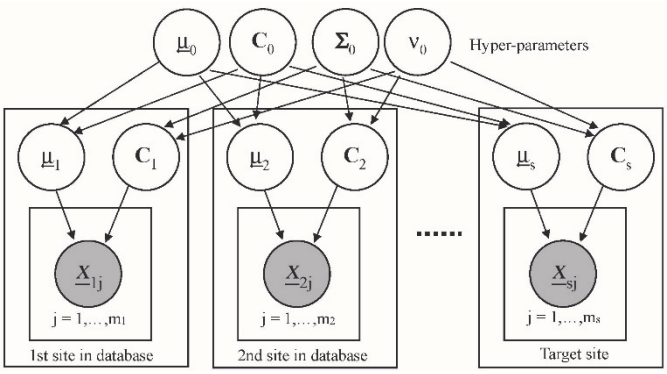


Fig. 3 Model structure of the HBM

3.3 Construction of Quasi-Site-Specific Model by HBM

The HBM has two stages: learning stage and inference stage. In the learning stage, the hyper-parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0) are calibrated to learn the inter-site and intra-site characteristics in the soil property database. The calibrated hyper-parameters can produce a "prior model" for the target site. In the inference stage, this prior model is updated into a posterior model by further conditioning on the target-site data. This posterior model is quasi-site-specific because not only the target-site data are used to construct the model but also the soil property database is used to develop its prior. In the following sub-sections, the learning and inference stages of the HBM are demonstrated.

3.3.1 HBM Learning Stage

To begin with, the following 8 soil parameters are considered, denoted by (Y_1, Y_2, …, Y_8):

$$\begin{aligned} Y_1 &= \text{LI} \\ Y_2 &= \ln(D_{50}) \qquad (D_{50} \text{ in mm}) \\ Y_3 &= \ln(s_{u,\text{VST}}/P_a) \\ Y_4 &= \ln(N_{60}) \\ Y_5 &= \ln[(q_t - \sigma_v)/P_a] \\ Y_6 &= B_q \\ Y_7 &= \ln(E_{\text{PMT}}/P_a) \\ Y_8 &= \ln[(1-\nu^2)\times E_{\text{DMT}}/P_a] \end{aligned} \tag{1}$$

The natural logarithm is taken for all non-negative soil parameters. There are 7,186 records in the SOIL-DMT/8/7186 database. Each record contains $\underline{Y} = (Y_1, Y_2, \ldots, Y_8)$ of a soil with possibly missing entries (some Y values are not measured). These 7,186 records are grouped into 701 groups (i = 1, …, 701) because there are 701 sites. The i-th group has m_i records $\{\underline{Y}_{ij}: j = 1, \ldots, m_i\}$. Because the HBM proposed by Ching *et al.* (2021a) operates in the standard normal space, certain transforms (*e.g.,* the Johnson transform is adopted in Ching *et al.* 2021a) are needed to convert $\underline{Y}$ into standard normal variable $\underline{X}$. The purpose of the HBM learning stage is to calibrate the hyper-parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0) such that the calibrated ones can mimic the inter-site and intra-site variabilities exhibited in SOIL-DMT/8/7186.

To demonstrate the HBM learning outcome, the behaviors of the calibrated hyper-parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0) are compared with the actual site-specific statistics of SOIL-DMT/8/7186. The calibrated hyper-parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0) can generate "new sites" by the following simulation: $\underline{\mu}_{\text{new}} \sim N(\underline{\mu}; \underline{\mu}_0, \mathbf{C}_0)$ and $\mathbf{C}_{\text{new}} \sim \text{IW}(\mathbf{C}; \mathbf{\Sigma}_0, \nu_0)$. Figure 7 compares the distribution of ($\underline{\mu}_{\text{new}}$, $\mathbf{C}_{\text{new}}$) generated

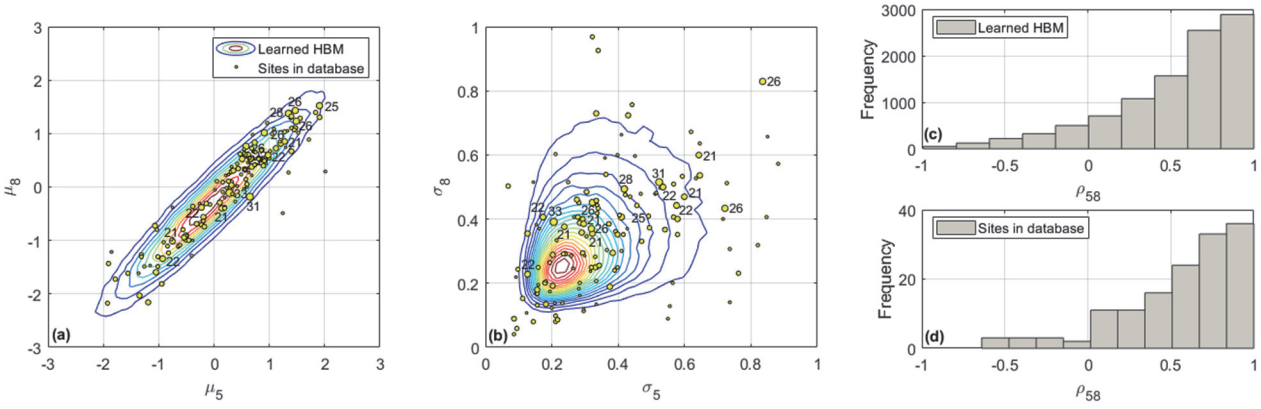


Fig. 4 ($\underline{\mu}_{\text{new}}$, $\mathbf{C}_{\text{new}}$) generated by the learned HBM *vs.* actual statistics in SOIL-DMT/8/7186: (a) μ_5 and μ_8; (b) σ_5 and σ_8; (c) ρ_{58} of the learned HBM; (d) actual statistics of ρ_{58}

by the learned HBM with the actual statistics of the 701 sites in SOIL-DMT/8/7186. The actual statistics for the sites in the database are shown as yellow dots in Figs. 7(a) and 7(b), where a larger dot denotes a site with more records (the numbers of records for some prominent sites are annotated in the plot). For visualization, only the statistics of X_5 [corresponding to $(q_t - \sigma_v)/P_a$] *vs.* X_8 [corresponding to $(1 - \nu^2) \times E_{DMT}/P_a$] are demonstrated. The conclusions for other pairs of soil parameters are similar. Figure 7(a) compares the means, Fig. 7(b) compares the standard deviations, and Figs. 7(c) and 7(d) compare the correlation.

Figure 8 further illustrates each new site generated by the calibrated hyper-parameters as a 1-σ ellipse (one standard deviation in each eigenvector direction in the X space). The ellipses in the Y_5-Y_8 space are skewed because they experience a nonlinear X-to-Y transformation. From Figs. 7 and 8, it is evidence that calibrated hyper-parameters ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{\Sigma}_0$, ν_0) have somewhat captured the inter-site and intra-site characteristics in SOIL-DMT/8/7186. In fact, Figs. 7 and 8 illustrate the behaviors of the "prior model" produced by the learned HBM. This prior model will be later updated into the posterior model by the target-site data during the HBM inference stage.

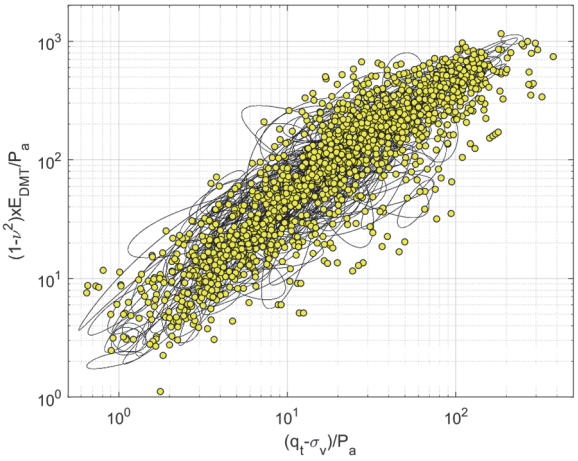


Fig. 8 ($\underline{\mu}_{new}$, $\mathbf{C}_{new}$) generated by the learned HBM (each ellipse represents a new site). The yellow dots in the background are the records in SOIL-DMT/8/7186.

3.3.2 HBM Inference Stage

Now consider the Bothkennar test site (UK) (Nash *et al.* 1992) as the target site. The Bothkennar test site is not within the SOIL-DMT/8/7186 database. There are many in-situ and laboratory tests conducted at this test site, but only those relevant to (Y_1, …, Y_8) are investigated in this section. The site investigation data are digitalized from the plots in Nash *et al.* (1992) with a 1-m depth interval, as shown in Table 4. The Bothkennar test site is a soft clay site, and the D_{50} information is not reported. The SPT N blow count is also not reported in Nash *et al.* (1992). Self-bored PMT was conducted but the E_{PMT} data are not directly available from the plots. The values of $(1 - \nu^2) \times E_{DMT}/P_a$ are available for all depths in Table 4 (ν is assumed to be 0.5).

To illustrate the HBM inference stage, the $(1 - \nu^2) \times E_{DMT}/P_a$ values for some depths are deliberately treated as unknowns (see the numbers in the parentheses in the table). The known soil modulus data are sparse such that a site-specific transformation model may not be reliably constructed based on the sparse soil modulus data. The "unknown" soil modulus values will later serve as the validation data for the inference stage. The following data scenarios with sparse soil modulus data are considered:

1. 0-known scenario; see the 3rd column from the right of Table 4. All E_{DMT} data are treated as unknown. This is the data scenario with the most sparsity.
2. 3-known scenario; see the 2nd column from the right of Table 4. The E_{DMT} data for most depths are treated as unknown except 3 depths.
3. 7-known scenario; see the rightmost column of Table 4. The E_{DMT} data for most depths are treated as unknown except 7 depths.

The remaining target-site data in the table [LI, $s_{u,VST}/P_a$, $(q_t - \sigma_v)/P_a$, B_q, and the "known" $(1 - \nu^2) \times E_{DMT}/P_a$ values] are used to update the prior model obtained from the HBM learning stage into the posterior model. Recall that the posterior model is quasi-site-specific. The resulting (posterior) quasi-site-specific model is shown as the contour lines in Fig. 9.

Table 4 Data of the Bothkennar test site (data digitalized from the plots in Nash *et al.* 1992)

Depth (m)	LI (Y_1)	D_{50} (Y_2)	$s_{u,VST}/P_a$ (Y_3)	N_{60} (Y_4)	$(q_t - \sigma_v)/P_a$ (Y_5)	B_q (Y_6)	E_{PMT}/P_a (Y_7)	$(1 - \nu^2) \times E_{DMT}/P_a$ (Y_8)		
								0 known	3 known	7 known
1.5	0.36	–	0.21	–	2.92	0.32	–	(10.94)*	(10.94)	(10.94)
2.5	0.63	–	0.22	–	3.39	0.36	–	(9.48)	9.48	9.48
3.5	0.92	–	0.23	–	2.35	0.53	–	(9.48)	(9.48)	(9.48)
4.5	0.84	–	0.25	–	2.25	0.61	–	(10.21)	(10.21)	10.21
5.5	0.82	–	0.27	–	2.70	0.62	–	(10.39)	(10.39)	(10.39)
6.5	0.97	–	0.29	–	2.76	0.65	–	(12.95)	(12.95)	12.95
7.5	0.66	–	0.30	–	3.05	0.65	–	(14.59)	(14.59)	(14.59)
8.5	0.84	–	0.32	–	3.15	0.68	–	(16.41)	16.41	16.41
9.5	1.08	–	0.35	–	3.33	0.70	–	(16.96)	(16.96)	(16.96)
10.5	0.84	–	0.38	–	3.87	0.62	–	(20.79)	(20.79)	20.79
11.5	0.73	–	0.42	–	4.08	0.69	–	(18.97)	(18.97)	(18.97)
12.5	0.84	–	0.43	–	4.53	0.62	–	(22.25)	(22.25)	22.25
13.5	0.80	–	0.47	–	4.86	0.68	–	(23.34)	(23.34)	(23.34)
14.5	0.87	–	0.47	–	5.51	0.59	–	(23.34)	23.34	23.34
15.5	0.69	–	0.48	–	5.26	0.72	–	(24.80)	(24.80)	(24.80)

* Values inside parentheses "(.)" are treated as unknowns during the HBM inference.

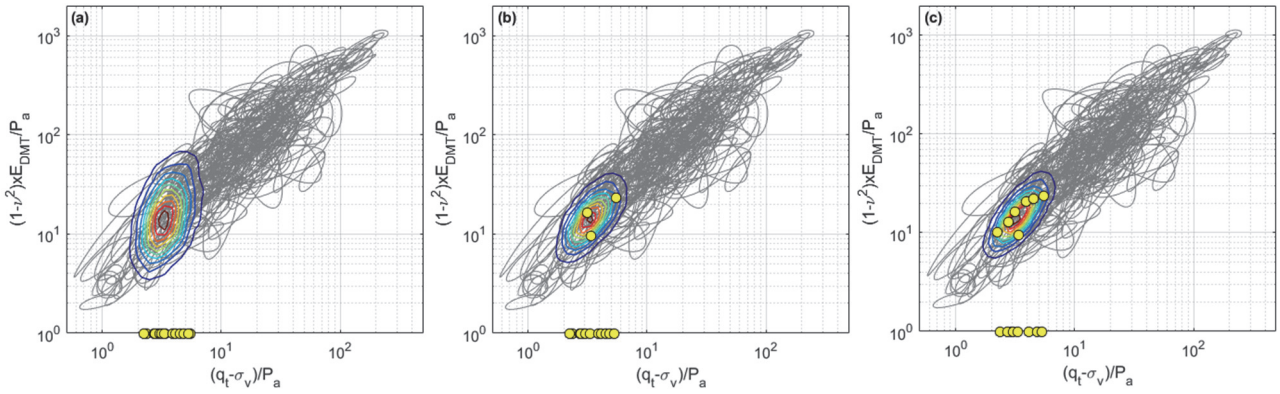


Fig. 9 The quasi-site-specific model obtained by the HBM inference stage: (a) 0 known; (b) 3 known; (c) 7 known. The background grey ellipses illustrate the prior model. The yellow dots are the known target-site data.

The quasi-site-specific model can be used to predict the unknown $(1-\nu^2)\times E_{DMT}/P_a$. Recall that these "unknown" $(1-\nu^2)\times E_{DMT}/P_a$ are deliberately treated as unknown during the inference stage. In fact, they are known and can serve as the validation data. The vertical line segments in Fig. 10 show the 95% confidence intervals (95% CIs) of the quasi-site-specific prediction. The validation data are shown as the red dots in the figure. For comparison, the prediction provided by the generic transformation model in Fig. 3(b) is also shown in Fig. 10. The vertical interval between the two dashed lines is the 95% CI for the generic prediction. It is clear that the 95% CIs for the quasi-site-specific prediction (vertical bars) are narrower than that for the generic prediction (vertical interval between the two dashed lines), yet the validation data (red dots) still lie within the vertical bars. This suggests that the both models provide consistent predictions (the validation data are within the 95% CIs), yet the quasi-site-specific model is more effective than the generic model in terms of the size of 95% CI.

4. EFFECT OF CHOICE OF DATABASE ON QUASI-SITE-SPECIFIC MODEL

This section illustrates that the effectiveness of the quasi-site-specific model generally depends on the adopted soil property database: a database more relevant to the target site is more effective (produce narrower 95% CI). To illustrate this, various soil property databases are adopted to train the HBMs, and the prior models produced by the HBM learning stage are updated into their (posterior) quasi-site-specific models in the HBM inference stage. The effectiveness of each quasi-site-specific model is quantified by the size of its 95% CI. A quasi-site-specific model with a narrower 95% CI (yet all validation data are still within the CI) is a more effective one. Let us first consider the following databases:

1. The entire SOIL-DMT/8/7186 is adopted. There are 701 sites in this database. Among the 701 sites, 514 sites are clay sites, and 187 sites are granular soil sites. The resulting quasi-site-specific model has been presented in the previous section.
2. The database that only contains the 514 clay sites is adopted. There are 5,318 records in total in this database. This database is denoted by C514-DMT/8/5318.
3. The database that only contains the 187 granular soil sites is adopted. There are 1,868 records in total in this database. This database is denoted by G187-DMT/8/1868.

Figure 11 shows the inference results for the 7-known scenario (the rightmost column in Table 4) of the Bothkennar test site. It is clear that the quasi-site-specific model based on C514-DMT/8/5318 (Fig. 11(b)) are roughly as effective as that based on SOIL-DMT/8/7186 (Fig. 11(a)) because the sizes of the 95% CIs are comparable. This suggests that removing granular soil sites from the database (namely, reducing SOIL-DMT/8/7186 to C514-DMT/8/5318) does not seem to degrade the effectiveness of the resulting quasi-site-specific model. This result is reasonable because the target site (Bothkennar test site) is a soft clay site, so the

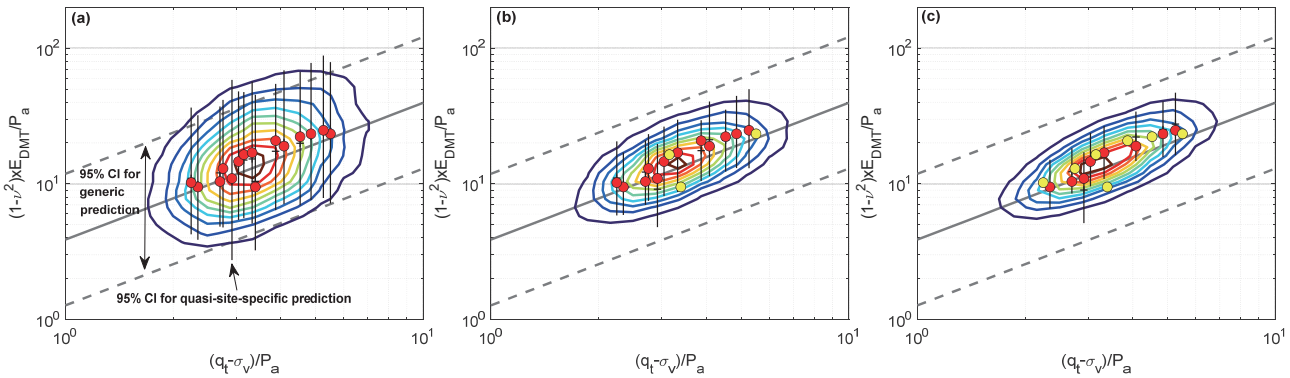


Fig. 10 The $(1-\nu^2)\times E_{DMT}/P_a$ prediction results by the quasi-site-specific model: (a) 0 known; (b) 3 known; (c) 7 known. The yellow dots are the target-site data with known $(1-\nu^2)\times E_{DMT}/P_a$. The red dots are the validation data.

granular soil sites are less relevant to the target site. The removal of these granular soil sites should not have much negative impact. In contrast, the quasi-site-specific model based on G187-DMT/8/1868 (Fig. 11(c)) are less effective with wider 95% CIs. This suggests that removing clay sites from the database (namely, reducing SOIL-DMT/8/7186 to G187-DMT/8/1868) does seem to degrade the effectiveness of the resulting quasi-site-specific model. This result is also reasonable.

Note that the Bothkennar test site has multivariate information of (Y_1, Y_3, Y_5, Y_6, Y_8) (see Table 4). Among the 701 sites in SOIL-DMT/8/7186, there are 12 clay sites with complete multivariate information of (Y_1, Y_3, Y_5, Y_6, Y_8). Now consider the database that contains these 12 clay sites. There are only 197 records in total in this small database, denoted by C12-DMT/8/197. Figure 11(d) shows the inference results for the 7-known scenario for the quasi-site-specific model based on C12-DMT/8/197. The quasi-site-specific model based on C12-DMT/8/197 (Fig. 11(d)) are slightly more effective than those based on SOIL-DMT/8/7186 (Fig. 11(a)) and C514-DMT/8/5318 (Fig. 11(b)) in terms of slightly narrower 95% CIs. It is remarkable that a small database that contains only 12 clay sites relevant to the Bothkennar test site [relevant in the sense that the 12 sites contain complete multivariate information of (Y_1, Y_3, Y_5, Y_6, Y_8)] can slightly outperform databases containing hundreds of sites.

5. CONCLUSIONS AND REMARKS

In this paper, a new soil property database named SOIL-DMT/8/7186 is compiled. This database contains 8 parameters of 7186 soil records from 701 sites, including 2 modulus parameters obtained by the pressuremeter test (PMT) and dilatometer test (DMT). However, the soil modulus data are mainly from DMT, not from PMT. The HBM is adopted to learn the characteristics of the 701 sites in SOIL-DMT/8/7186. The learned HBM can produce a prior model for the target site, and this prior model can be further updated by sparse target-site data into a (posterior) quasi-site-specific model. The quasi-site-specific model can be used to predict the soil modulus of the target site based on inexpensive site investigation data such as SPT blow account, CPT cone tip resistance, VST undrained shear strength, index properties, etc. It is shown that the quasi-site-specific model is superior to the generic transformation model constructed based on the same database in terms of narrower 95% CIs.

A remarkable observation obtained in the current paper is that the effectiveness of the quasi-site-specific model depends on the database. The quasi-site-specific model based on a database more relevant to the target site is more effective (produce narrower 95% CIs). For a clay target site, a clay database is more relevant than a granular soil database. More interestingly, it is found that a small database that contains only a few sites (*e.g.,* 10+ sites) relevant to the target site may still perform well if these sites are properly

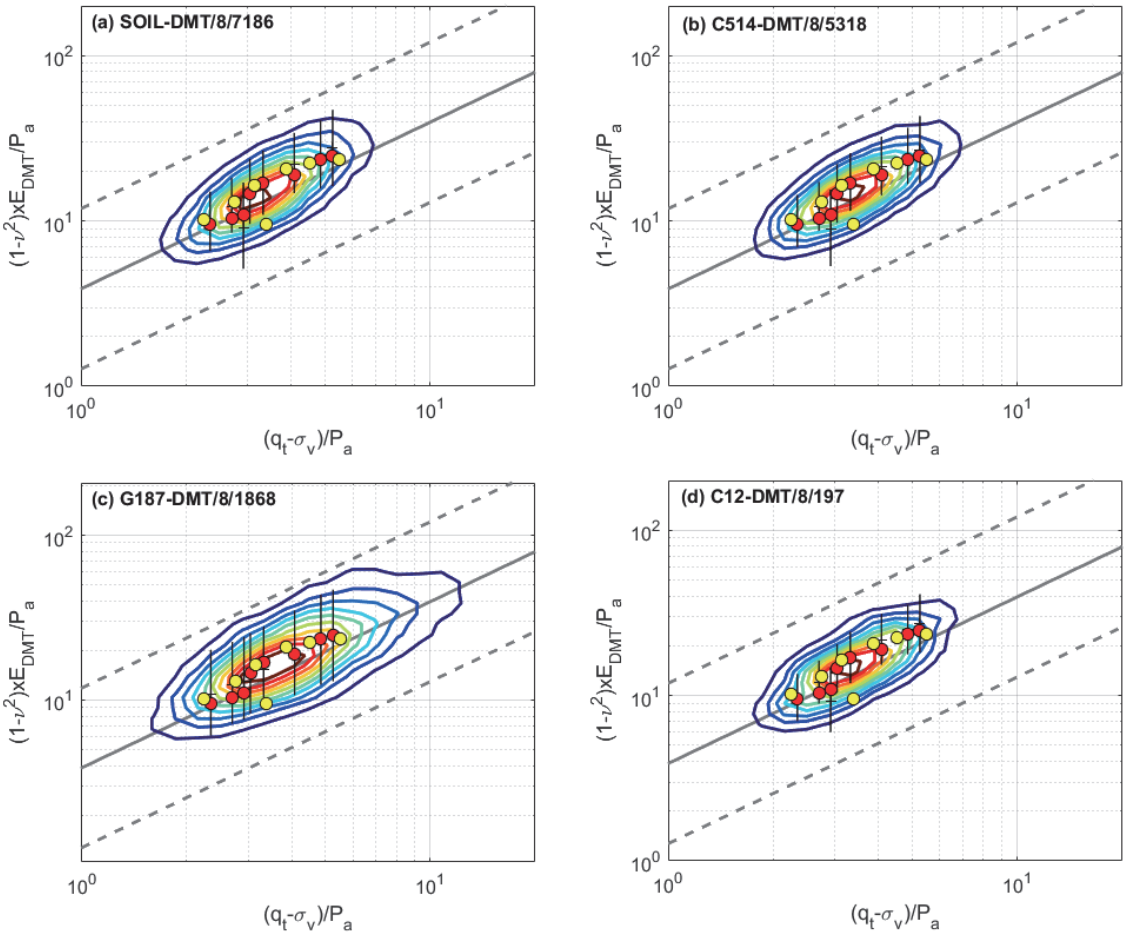


Fig. 11 The $(1-\nu^2)\times E_{DMT}/P_a$ prediction results by the quasi-site-specific models based on various databases: (a) SOIL-DMT/8/7186; (b) C514-DMT/8/5318; (c) G187-DMT/8/1868; (d) C12-DMT/8/197.

chosen. Although the current paper does not make effort in defining or quantifying the concept of "the sites relevant to the target sites", Sharma *et al.* (2022) recently proposed the concept of "similarity between sites". They also proposed a HBM-based framework to quantify the similarity between sites. It is possible to identify "the sites relevant to the target sites" by using the site similarity proposed by Sharma *et al.* (2022). Once these "relevant sites" are identified, they can form a small database to train the HBM and may produce a quasi-site-specific model that is effective for the target site. This new line of research, however, is not pursued in the current paper and is left as a future research direction.

FUNDING

The first author would like to thank the funding support from the National Council of Science and Technology of Taiwan (109-2221-E-002-029-MY3).

DATA AVAILABILITY

The database SOIL-DMT/8/7186 can be downloaded at the following ISSMGE TC304 (risk) webpage: http://140.112.12.21/issmge/tc304.htm?=6. The computer codes generated in this study are available from the corresponding author upon reasonable request.

CONFLICT OF INTEREST STATEMENT

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

REFERENCES

Baguelin, F., Jezequel, J.F., and Shields, D.H. (1978). "The pressuremeter and foundation engineering." *Trans Tech Publications*, Clausthal.

Ching, J., Phoon, K.K., and Wu, C.T. (2022). "Data-centric quasi-site-specific prediction for compressibility of clays." *Canadian Geotechnical Journal*, **59**(12), 2033-2049. https://doi.org/10.1139/cgj-2021-0658

Ching, J., Phoon, K.K., Ho, Y.H., and Weng, M.C. (2021b). "Quasi-site-specific prediction for deformation modulus of rock mass." *Canadian Geotechnical Journal*, **58**, 936-951. https://doi.org/10.1139/cgj-2020-0168

Ching, J., Wu, S., and Phoon, K.K. (2021a). "Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model." *ASCE Journal of Engineering Mechanics*, ASCE, **147**(10), 04021069. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001964

James, A. (1964). "Distributions of matrix variates and latent roots derived from normal samples." *Annals of Mathematical Statistics*, **35**(2), 475-501. http://dx.doi.org/10.1214/aoms/1177703550

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

Kuo, M.C. (2020). *Big Data Analysis for Soil Parameters: Focusing on Modulus and Coefficient of Earth Pressure at Rest*. M.S. Thesis, Department of Civil Engineering, National Taiwan University, Taipei, Taiwan (in Chinese).

Marchetti, S. (1980). "In situ tests by flat dilatometer." *ASCE Journal of Geotechnical Engineering Division*, **106**(3), 299-321. https://doi.org/10.1061/AJGEB6.0000934

Mayne, P.W. (2001). "Stress-strain-strength-flow parameters from enhanced in-situ tests." *Proceedings of the International Conference on In-Situ Measurement of Soil Properties & Case Histories*, Bali, Indonesia, 27-48.

Nash, D.F.T., Powell, J.J.M., and Lloyd, I.M. (1992). "Initial investigation of the soft clay test site at Bothkennar." *Geotechnique*, **42**(2), 163-181. https://doi.org/10.1680/geot.1992.42.2.163

Ohya, S., Imai, T., Matsubara, M. (1982). "Relation between N value by SPT and LLT pressuremeter results." *Proceedings of the 2nd European Symposium on Penetration Testing*, Amsterdam, **1**, 125-130.

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

Sharma, A., Ching, J., and Phoon, K.K. (2022). "A hierarchical Bayesian similarity measure for geotechnical site retrieval." *ASCE Journal of Engineering Mechanics*, ASCE, **148**(10), 04022062. https://doi.org/10.1061/(ASCE)EM.1943-7889.0002145