Proc. of the 6th Intl. Symposium on Reliability Engineering and Risk Management (6ISRERM)
31 May – 1 June 2018, Singapore
Editor(s) Xudong Qian, Sze Dai Pang, Ghim Ping Raymond Ong, Kok-Kwang Phoon

BAYESIAN DATA MINING FOR A GENERIC GEOTECHNICAL DATABASE

JIANYE CHING[1] and KOK-KWANG PHOON[2]

[1]*Department of Civil Engineering, National Taiwan University, Taipei, Taiwan.*
E-mail: jyching@gmail.com
[2]*Department of Civil & Environmental Eng., National University of Singapore, Singapore.*
E-mail: kkphoon@nus.edu.sg

This paper proposes a Bayesian data mining approach that searches a generic database for data points with soil characteristics similar to a set of site-specific data. A similarity index between the generic and site-specific data points is proposed based on the Bayesian analysis. The effectiveness of the proposed approach is illustrated by considering a generic clay database and a specific site in Sweden. The generic data points identified as "similar" can be combined with the limited site-specific data to construct a transformation model more relevant to a specific site.

Keywords: geotechnical database, data mining, site characterization, transformation model.

1 Introduction

Geotechnical design is site-specific, because every site has its unique geological characteristics. This site-specific aspect of geotechnical design can be found in transformation models (Phoon and Kulhawy 1999). The transformation model suitable for one site may not be suitable for another site. It is obviously desirable to adopt a site-specific transformation model constructed by site-specific site investigation data in a design project. However, for small projects where the budget does not justify extensive site investigation, it is typically not possible to construct the site-specific transformation model with sufficient confidence. Under this circumstance, the engineer is forced to adopt a generic transformation model constructed by data from other sites under the assumption that the geology is broadly similar. Useful compilations of generic transformation models are available in the literature (e.g., Djoenaidi 1985; Kulhawy and Mayne 1990; Mayne et al. 2001). When applied to a specific site, the transformation uncertainty of a generic transformation model can be excessively large, because it is intended to accommodate a wide range of soil types and site conditions. However, if we narrow down to a single site, the site investigation data can be too sparse to construct the site-specific transformation model with any acceptable degree of statistical significance.

Although the site investigation data in a small project may be insufficient to construct a reliable site-specific transformation model, the data may be sufficient to reveal certain soil property characteristics for the site of interest. For instance, a limited sample of data at a Taipei site may reveal that the Taipei clay is primarily lightly over consolidated (LOC) and medium plastic (MP). In this case, it is reasonable to argue that the data points in the generic database that are also LOC and MP may be more relevant to this site than other generic data points. Transformation models constructed by these LOC and MP generic data points, albeit not site-specific, may be more suitable for Taipei than those constructed by generic data points with wide ranges of OCR and plasticity.

The approach of extracting relevant data points from a generic database can be broadly categorized as a data mining approach. The basic idea is to search for generic data that have similar soil characteristics as the site-specific data. However, this is not straightforward because geotechnical data are usually incomplete. For instance, if the properties of concern include (PI, σ'_v, σ'_p, s_u, SPT-N) (PI = plasticity index; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; s_u = undrained shear strength; SPT-N = N value for standard penetration test), in principle we need complete multivariate data with simultaneous knowledge of (PI, σ'_v, σ'_p, s_u, SPT-N) for both generic and site-specific data to define "similarity" in a quantitative way. However, it is common that a small project is lacking in such complete multivariate data points. It is more common to measure incomplete multivariate data points at different depths and locations, for instance, some data points have (PI, σ'_v, SPT-N) information or some have (σ'_v, SPT-N) information. One simple method to circumvent this incompleteness difficulty is to look for the most commonly occurring pairwise information, e.g., (PI, SPT-N), and perform data mining using pairwise information only. However, it is possible that data points classified as "similar" based on (PI, SPT-N) may not be applicable to another transformation model involving other parameters, e.g., (PI, s_u).

If the data points are visualized as a spreadsheet table of size (m × 4), where m is the number of data points, incomplete multivariate data means there are missing entries in the spreadsheet table. The purpose of the current paper is to propose a data mining approach that can handle incomplete multivariate geotechnical data for the purpose of constructing a transformation model. A Bayesian data mining approach that can characterize the statistical uncertainty associated with sparse site-specific data will be proposed in the current paper. It will be shown that a data mining approach based on incomplete multivariate data is more robust than one based on bivariate information.

2 Generic Database versus Site-specific Data

2.1 *Generic database*

The proposed method requires a generic database. The word "generic" is in the sense that the database covers a wider range of conditions than those encountered at a single site. In the current paper, a global clay database named CLAY/10/7490 (Ching and Phoon 2014a) is adopted. The CLAY/10/7490 database consists of 7490 data points for ten dimensionless clay parameters from 251 studies in the literature that cover 30 countries/regions worldwide. The ten clay parameters are denoted by (Y_1, Y_2, …, Y_{10}):

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

where LL = liquid limit; PI = plasticity index; LI = liquidity index; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; 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). Note that CLAY/10/7490 is not a complete multivariate database. If the data points are visualized as a spreadsheet table of size (m_g × n), where m_g = 7490 is the total number of data points in the generic database and n = 10 is the dimension of each data point, there are lots of missing entries in the spreadsheet table. It is worth mentioning that each multivariate data

"point" is a row of numbers containing results from different tests conducted in close proximity at the same depth. A missing number in this row means that a particular test has not been carried at this location and depth. The key capability of this proposed Bayesian data mining approach is that it can identify generic data points that are "similar" to the site of interest under this incomplete multivariate context.

2.2 *Site-specific data*

Table 1 shows the site investigation results for a clay site in Stora an (Sweden) (D'Ignazio et al. 2016). The site-specific data will be denoted by **Y** from here on. **Y** can be visualized as a spreadsheet table of size ($m_s \times n$), where $m_s = 7$ is the total number of data points (rows) in Table 1 and n = 10 to match information available in CLAY/10/7490 even though the columns Y_7 to Y_{10} are empty for all depths (see grey boxes). Moreover, some (Y_5, Y_6) data are deliberately removed (they are available in D'Ignazio et al. 2016) to demonstrate the ability of the proposed Bayesian data mining approach in handling incomplete data. They are shown as crossed-out numbers in grey boxes in Table 1. Let the observed data in Table 1 be denoted by $\mathbf{Y}_o$ (normal entries) and the unobserved data be denoted by $\mathbf{Y}_u$ (grey entries). The original s_u data in D'Ignazio et al. (2016) are based on field vane. They are converted to mobilized s_u using the empirical equation proposed by Bjerrum (1972).

Table 1. Site investigation results for a clay site in Stora an (Sweden) (Source: D'Ignazio et al. 2016).

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})
1.5	113.8	73.8	0.92	0.103	0.433	0.657				
2.0	115.3	74.6	0.92	0.111	~~0.256~~*	~~0.532~~*				
2.3	125.0	70.8	0.97	0.118	0.237	0.475				
3.1	118.3	76.1	0.99	0.139	0.185	0.342				
3.8	123.5	85.8	0.89	0.162	0.200	0.276				
4.6	104.1	58.2	1.05	0.193	~~0.286~~*	~~0.340~~*				
5.3	104.9	63.4	0.98	0.225	0.313	0.355				

* Entries are known in D'Ignazio et al. (2016) but are made empty in this study to demonstrate the ability of the proposed Bayesian data mining approach in handling incomplete data.

2.3 *Multivariate normality*

The proposed approach operates in the multivariate normal space, but soil data are typically non-normal. It is desirable to convert the Y_i data to normal variable X_i by a certain transform. The transform based on the cumulative density function (CDF) of the Johnson distribution (Johnson 1949) used by Ching and Phoon (2014b, 2018) is adopted in the current paper to maintain the consistency between the current paper and our past works. This CDF transform is adopted to transform Y_i to X_i for both generic and site-specific data. After the transformation, the generic X dataset is still a spreadsheet table of size ($m_g \times n$), and $\mathbf{X}_o$ and $\mathbf{X}_u$ still correspond to normal and grey entries similar to Table 1. Moreover, it is further *assumed* that site-specific property $\underline{X} = (X_1, X_2, \ldots, X_n)$ is multivariate normal:

$$f\left(\underline{x} \mid \underline{\mu}_s, \mathbf{C}_s\right) = N\left(\underline{x} \mid \underline{\mu}_s, \mathbf{C}_s\right) = \left|\mathbf{C}_s\right|^{-\frac{1}{2}} \left(2\pi\right)^{-\frac{n}{2}} \exp\left[-0.5 \times \left(\underline{x} - \underline{\mu}_s\right)^{\mathrm{T}} \mathbf{C}_s^{-1} \left(\underline{x} - \underline{\mu}_s\right)\right] \tag{2}$$

where n is the dimension of the multivariate PDF (n = 10 for our example); $N(\underline{x}|\underline{\mu}_s,\mathbf{C}_s)$ denotes a multivariate normal PDF for $\underline{x}$ with mean vector = $\underline{\mu}_s$ and covariance matrix = $\mathbf{C}_s$; $\underline{\mu}_s \in \mathbf{R}^{n\times 1}$ is the mean vector for $\underline{X}$; $\mathbf{C}_s \in \mathbf{R}^{n\times n}$ is the covariance matrix for $\underline{X}$ that characterizes the site-specific correlation among $(X_1, X_2, \ldots, X_n)$. The site-specific mean and covariance ($\underline{\mu}_s$, $\mathbf{C}_s$) are treated as unknowns and will be inferred from $\mathbf{X}_o$ using the Bayesian approach. The multivariate normality for $\underline{X} = (X_1, X_2, \ldots, X_n)$ is a key assumption adopted by the current paper.

3 Proposed Bayesian Data Mining Approach

The proposed Bayesian data mining approach contains two steps. In the first step, the Bayesian analysis is adopted to construct the posterior PDF of $\underline{X} = (X_1, X_2, \ldots, X_n)$ conditioning on $\mathbf{X}_o$, denoted by $f(\underline{x}|\mathbf{X}_o)$. In its essence, $f(\underline{x}|\mathbf{X}_o)$ summarizes the soil characteristics at the site of interest as a multivariate PDF so that the random sample $\underline{X} \sim f(\underline{x}|\mathbf{X}_o)$ has a multivariate distribution similar to that for $\mathbf{X}_o$. In the second step, $f(\underline{x}|\mathbf{X}_o)$ is further adopted to quantify the similarity between generic data and $\mathbf{X}_o$. Let $\underline{x}_g^{(k)}$ be the k-th generic data point (k = 1, 2, …, m_g), i.e., $\underline{x}_g^{(k)} \in \mathbf{R}^{n\times 1}$ corresponds to the k-th row in the ($m_g \times n$) spreadsheet, where m_g = 7490 and n = 10 for CLAY/10/7490. The posterior probability of $\underline{x}_g^{(k)}$, denoted as $P(k|\mathbf{X}_o)$, which can be computed based on $f(\underline{x}|\mathbf{X}_o)$, quantifies the plausibility of $\underline{x}_g^{(k)}$ (k^{th} row in CLAY/10/7490) given $\mathbf{X}_o$. For $\underline{x}_g^{(k)}$ whose characteristics are similar to those for $\mathbf{X}_o$, $P(k|\mathbf{X}_o)$ is large, and the converse is also true. The generic data points with larger $P(k|\mathbf{X}_o)$ may be more relevant to the site of interest than other generic data points.

3.1 *Construction of* $f(\underline{x}|\mathbf{X}_o)$

For the construction of $f(\underline{x}|\mathbf{X}_o)$, it suffices to estimate $(\underline{\mu}_s,\mathbf{C}_s)$. The main challenge for estimating ($\underline{\mu}_s$, $\mathbf{C}_s$) is that $\mathbf{X}_o$ is incomplete, because most parameter estimation techniques require complete $\mathbf{X}_o$. For incomplete $\mathbf{X}_o$, Ching and Phoon (2018) showed that it is possible to draw ($\underline{\mu}_s$, $\mathbf{C}_s$) samples from $f(\underline{\mu}_s,\mathbf{C}_s|\mathbf{X}_o)$ in an analytical manner by adopting the Gibbs sampler (GS) (Geman and Geman 1984; Gilks et al. 1996) in conjunction with the assumed conjugate prior PDFs. Moreover, unobserved entries, denoted by $\mathbf{X}_u$, can be also sampled in an analytical manner (Ching and Phoon 2018). The basic idea is to divide the random variables into three groups, ($\underline{\mu}_s$, $\mathbf{C}_s$, $\mathbf{X}_u$), and the GS is adopted to sequentially sample them from the following conditional PDFs:

$$\underline{\mu}_s \sim f\left(\underline{\mu}_s \mid \mathbf{C}_s, \mathbf{X}_u, \mathbf{X}_o\right) \qquad \mathbf{C}_s \sim f\left(\mathbf{C}_s \mid \underline{\mu}_s, \mathbf{X}_u, \mathbf{X}_o\right) \qquad \mathbf{X}_u \sim f\left(\mathbf{X}_u \mid \underline{\mu}_s, \mathbf{C}_s, \mathbf{X}_o\right) \tag{3}$$

Due to the assumed multivariate normality for $\underline{X}$, conjugate prior PDFs for ($\underline{\mu}_s$, $\mathbf{C}_s$) exist: the conjugate prior for $f(\underline{\mu}_s)$ is multivariate normal, and that for $f(\mathbf{C}_s)$ is inverse-Wishart. The posterior PDFs $f(\underline{\mu}_s|\mathbf{C}_s,\mathbf{X}_u,\mathbf{X}_o)$ and $f(\mathbf{C}_s|\underline{\mu}_s,\mathbf{X}_u,\mathbf{X}_o)$ will be still multivariate normal and inverse-Wishart. Moreover, $f(\mathbf{X}_u|\underline{\mu}_s,\mathbf{C}_s,\mathbf{X}_o)$ is also multivariate normal (Ching and Phoon 2018) due to the assumed multivariate normality for $\underline{X}$. As a result, ($\underline{\mu}_s$, $\mathbf{C}_s$, $\mathbf{X}_u$) can be sampled in an analytical manner. The details for this GS algorithm can be found in Ching and Phoon (2018). Let us denote the samples obtained using the GS by ($\underline{\mu}_s^{gb}$, $\mathbf{C}_s^{gb}$, $\mathbf{X}_u^{gb}$). The GS starts with an initial sample of ($\underline{\mu}_{s,0}^{gb}$, $\mathbf{C}_{s,0}^{gb}$, $\mathbf{X}_{u,0}^{gb}$) (time step t = 0), then it sequentially draws samples ($\underline{\mu}_{s,t}^{gb}$, $\mathbf{C}_{s,t}^{gb}$, $\mathbf{X}_{u,t}^{gb}$) (t = 1, 2, …, T) from the conditional PDFs in Eq. (3) based on the latest parameter values. The ($\underline{\mu}_{s,t}^{gb}$, $\mathbf{C}_{s,t}^{gb}$, $\mathbf{X}_{u,t}^{gb}$) samples after the burn-in period are collected. It can be shown that these samples are distributed as $f(\underline{\mu}_s, \mathbf{C}_s, \mathbf{X}_u|\mathbf{X}_o)$. It is noteworthy that the scatter of the ($\underline{\mu}_{s,t}^{gb}$, $\mathbf{C}_{s,t}^{gb}$) samples quantifies the site-specific statistical uncertainty. It is essential to quantify this statistical uncertainty rigorously if $\mathbf{X}_o$ is sparse. Based on the total probability theorem, the posterior PDF $f(\underline{x}|\mathbf{X}_o)$ can be approximated as a mixture of multivariate normal PDFs:

$$f\left(\underline{\mathrm{x}} \mid \mathbf{X}_o\right) \approx \frac{1}{T-t_b}\left[\sum_{t=t_b+1}^{T}\left|\mathbf{C}_{s,t}^{gb}\right|^{-\frac{1}{2}}(2\pi)^{-\frac{n}{2}}\exp\left[-0.5\times\left(\underline{x}-\underline{\mu}_{s,t}^{gb}\right)^{\mathrm{T}}\left(\mathbf{C}_{s,t}^{gb}\right)^{-1}\left(\underline{x}-\underline{\mu}_{s,t}^{gb}\right)\right]\right] \tag{4}$$

where t_b is the end of the burning-period. It is desirable that the prior PDFs $f(\underline{\mu}_s)$ and $f(\mathbf{C}_s)$ are non-informative. The multivariate normal prior $f(\underline{\mu}_s)$ can be made non-informative by adopting large variances. However, it is challenging to make the inverse-Wishart prior $f(\mathbf{C}_s)$ non-informative. Ching and Phoon (2018) adopted the hierarchical inverse-Wishart model proposed by Huang and Wand (2013). By adopting a set of hyperparameters, this hierarchical model makes $f(\mathbf{C}_s)$ roughly non-informative, yet the prior conjugacy required by the GS is preserved.

3.2 *Evaluation of* $\mathbf{P(k|X_o)}$

Recall that $\underline{\mathrm{x}}_g^{(k)}$ corresponds to the k-th data point or row (k = 1, 2, …, m_g) in the generic database. The posterior probabilities $P(k|\mathbf{X}_o)$ can be computed using the Bayes' rule:

$$\begin{aligned} P\left(k \mid \mathbf{X}_o\right) &= \left[f\left(\mathbf{X}_o \mid \underline{\mathrm{x}}_g^{(k)}\right)P(k)\right] \Big/ \left[\sum_{k=1}^{m_g} f\left(\mathbf{X}_o \mid \underline{\mathrm{x}}_g^{(k)}\right)P(k)\right] \\ &= \left[f\left(\underline{\mathrm{x}}_g^{(k)} \mid \mathbf{X}_o\right) \Big/ f\left(\underline{\mathrm{x}}_g^{(k)}\right)\right] \Big/ \left[\sum_{k=1}^{m_g} f\left(\underline{\mathrm{x}}_g^{(k)} \mid \mathbf{X}_o\right) \Big/ f\left(\underline{\mathrm{x}}_g^{(k)}\right)\right] \end{aligned} \tag{5}$$

where P(k) is the prior probability of $\underline{\mathrm{x}}_g^{(k)}$, taken to be $P(k) = 1/m_g$ for k = 1, 2, …, m_g; $f(\mathbf{X}_o|\underline{\mathrm{x}}_g^{(k)})$ is the likelihood of $\underline{\mathrm{x}}_g^{(k)}$; $f(\underline{\mathrm{x}}_g^{(k)})$ is the prior PDF of $\underline{\mathrm{x}}_g^{(k)}$. The posterior probability $P(k|\mathbf{X}_o)$ quantifies the probability of $\underline{\mathrm{x}}_g^{(k)}$ conditioning on $\mathbf{X}_o$, whereas the prior P(k) quantifies the probability of $\underline{\mathrm{x}}_g^{(k)}$ without $\mathbf{X}_o$. For $\underline{\mathrm{x}}_g^{(k)}$ with characteristics similar to those for $\mathbf{X}_o$, the posterior probability $P(k|\mathbf{X}_o)$ in Eq. (5) will be large, and the converse is also true. Therefore, $P(k|\mathbf{X}_o)$ quantifies whether $\underline{\mathrm{x}}_g^{(k)}$ has similar soil characteristics as $\mathbf{X}_o$. Equation (5) is evaluated for all data points (rows) in the generic database (k = 1, 2, …, m_g) to obtain m_g posterior probabilities that measure the similarity between the generic data points and the site-specific data $\mathbf{X}_o$.

Note that $\underline{\mathrm{x}}_g^{(k)}$ is usually an incomplete (n×1) vector with empty entries. Let us denote the observed entries in $\underline{\mathrm{x}}_g^{(k)}$ by $\underline{\mathrm{x}}_{go}^{(k)}$ and also denote the sub-mean vector by $\underline{\mu}_{so}$ corresponding to the observed entries and the sub-covariance matrix by $\mathbf{C}_{so}$. $f(\underline{\mathrm{x}}_g^{(k)}|\mathbf{X}_o)$ can be estimated as:

$$f\left(\underline{\mathrm{x}}_g^{(k)} \mid \mathbf{X}_o\right) \approx \frac{1}{T-t_b}\left[\sum_{t=t_b+1}^{T}\left|\mathbf{C}_{so,t}^{gb}\right|^{-\frac{1}{2}}(2\pi)^{-\frac{n_o}{2}}\exp\left[-\frac{1}{2}\left(\underline{x}_{go}^{(k)}-\underline{\mu}_{so,t}^{gb}\right)^{\mathrm{T}}\left(\mathbf{C}_{so,t}^{gb}\right)^{-1}\left(\underline{x}_{go}^{(k)}-\underline{\mu}_{so,t}^{gb}\right)\right]\right] \tag{6}$$

where n_o is the number of the observed entries in $\underline{\mathrm{x}}_g^{(k)}$ ($n_o \le n$). $f(\underline{\mathrm{x}}_g^{(k)})$ can also be estimated by a similar equation based on the Monte Carlo samples drawn from the prior PDFs $f(\underline{\mu}_s)$ and $f(\mathbf{C}_s)$.

3.3 *Example: the Stora an (Sweden) site*

Consider CLAY/10/7490 as the generic database and the Stora an (Sweden) site (Table 1) as the site of interest. The proposed Bayesian data mining approach is implemented to search for the generic data points in CLAY/10/7490 that have similar soil characteristics as the observed site-specific data $\mathbf{Y}_o$ in Table 1. For demonstration, Table 2 shows the top ten generic data points with relatively large $P(k|\mathbf{X}_o)$ as well as other three data points with very small $P(k|\mathbf{X}_o)$. It is

evidence that the data points with relatively large $P(k|\mathbf{X}_o)$ have soil characteristics similar to those in Table 1, e.g., high (LL, PI), LI around 1, NC to LOC, etc. In fact, some are Sweden cases, whereas others are from Bangkok, Thailand. The soil characteristics for the three cases with very small $P(k|\mathbf{X}_o)$ are not similar to those in Table 1. They are presented for illustration.

Table 2. $P(k|\mathbf{X}_o)$ values for chosen generic data points

Rank	$P(k\|\mathbf{X}^o)$	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	0.023	129.7	82.7	0.91	0.18	0.21	0.30		1.03	8.62	0.76	1.15	Sweden (Upplands-Vasby)
2	0.013	126.1	67.6	0.95	0.12	0.24	0.64	8.0				1.94	Sweden (Stockholm)
3	0.008	132.7	79.1	0.98	0.14	0.19	0.44					1.43	Thailand (Bangkok)
4	0.007	124.2	80.5	0.88	0.22	0.25	0.26		0.98	7.91	1.17	1.16	Sweden (Upplands-Vasby)
5	0.004	129.7	82.2	1.01	0.15	0.21	0.39		0.98	10.28	1.16	1.40	Sweden (Upplands-Vasby)
6	0.003	110.0	71.8	0.98	0.28	0.32	0.25		1.06	6.32	0.61	1.15	Sweden (Upplands-Vasby)
7	0.003	119.3	78.3	0.90	0.24	0.28	0.26		1.04	7.08	0.71	1.16	Sweden (Upplands-Vasby)
8	0.001	146.0	83.0	0.96	0.15	0.21	0.38					1.47	Thailand (Bangkok)
9	0.001	146.2	87.1	0.93	0.14	0.19	0.47					1.43	Thailand (Bangkok)
10	0.001	105.1	69.0	0.94	0.32	0.35	0.24		1.02	6.45	0.90	1.12	Sweden (Upplands-Vasby)
2000	2.1E-16						0.47						Taiwan
4000	1.9E-19		14.6				0.22						Taiwan (Taipei)
6000	1.4E-61	19.5	4.5	3.93									Norway (Manglerud)

Not all generic data points with large $P(k|\mathbf{X}_o)$ are useful for constructing a transformation model. For instance, if the purpose is to construct an OCR versus s_u/σ'_v transformation model, only the generic data with simultaneous information (σ'_v/P_a, σ'_p/P_a, s_u/σ'_v) are eligible. There are 1408 eligible generic data points in CLAY/10/7490. Note that the $P(k|\mathbf{X}_o)$ values for the 1408 eligible data points no longer sum up to unity. It is desirable to re-normalize them:

$$P'(k \mid \mathbf{X}_o) = P(k \mid \mathbf{X}_o) \Big/ \sum_{\text{Eligible } k} P(k \mid \mathbf{X}_o) \tag{7}$$

The re-normalized $P'(k|\mathbf{X}_o)$ for all eligible data now sum up to unity.

Figure 1a shows the OCR-s_u/σ'_v plot. In this plot, the generic data points with relatively large $P'(k|\mathbf{X}_o)$, called the "relevant" generic data points from here on, are plotted as solid squares. Moreover, a darker square corresponds to a larger $P'(k|\mathbf{X}_o)$. The less "relevant" generic data, e.g., those with $P'(k|\mathbf{X}_o) < 1\times10^{-4}$, are plotted as light cross markers. The relevant generic data seem to cluster around the site-specific data, indicating that their OCR-s_u/σ'_v characteristics are indeed similar. It is noteworthy that not only their OCR-s_u/σ'_v characteristics are similar, their other characteristics are also similar. For instance, some generic data points not only have (σ'_v/P_a, σ'_p/P_a, s_u/σ'_v) information but also have (LL, PI) information, e.g., many rows in Table 2 have it. Figure 1b shows the LL-PI correlation plot for the relevant generic data. Again, relevant generic data are plotted as solid squares, whereas less relevant ones as light crosses. It is clear that the LL-PI characteristics for the relevant generic data also similar to those for the site-specific data. This is because when the posterior probabilities $P'(k|\mathbf{X}_o)$ are computed, all information (X_1, X_2, …, X_n) are considered. As a result, the posterior probability measures the similarity in the (n×1) space, not just in the (σ'_v/P_a, σ'_p/P_a, s_u/σ'_v) space. This also explains why in Figure 1a there are generic data with small posterior probabilities clustering around $\mathbf{Y}_o$:

although these data have (σ'_v/P_a, σ'_p/P_a, s_u/σ'_v) similar to $\mathbf{Y}_o$, their other properties such as LL, PI, LI, etc. are not.

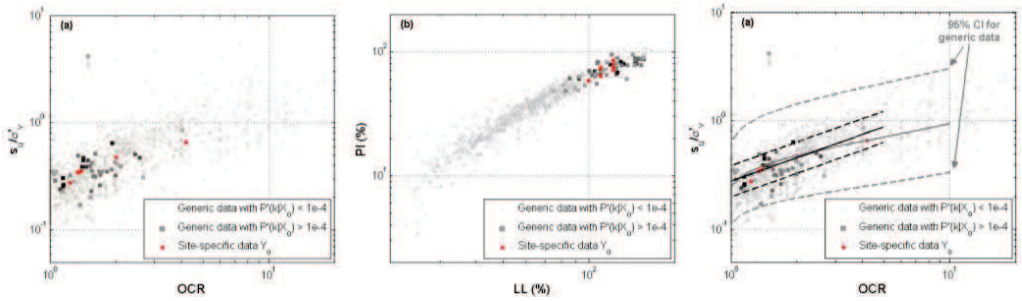

Figure 1. Results for Bayesian data mining: (a) OCR-s_u/σ'_v plot; (b) LL-PI plot; (c) median estimate and 95% confidence interval for s_u/σ'_v based on the WML method.

4 Construction of a Transformation Model

Because the observed site-specific data points $\mathbf{Y}_o$ are usually sparse, they alone may be insufficient to construct a reliable transformation model. Because the relevant generic data have similar soil characteristics as $\mathbf{Y}_o$, they can be combined with $\mathbf{Y}_o$ to construct a transformation model. While there are several methods of combining the two sources of data, only the weighted maximum likelihood (WML) method (e.g., Karampatziakis and Langford 2011) is demonstrated in this section. Also, only the construction of the OCR-s_u/σ'_v transformation model for the Stora an site is demonstrated. Consider the following transformation model:

$$\ln\left(s_u/\sigma'_v\right) = a + b\cdot\ln\left(\mathrm{OCR}\right) + \varepsilon \tag{8}$$

where (a,b) are unknown 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. The WML method determines (a,b,σ) by maximizing the weighted log-likelihood:

$$\left(a^*,b^*,\sigma^*\right) = \underset{a,b,\sigma}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \cdot \left[-\frac{1}{2}\ln\left(2\pi\right) - \ln\left(\sigma\right) - \frac{1}{2\sigma^2}\left(\ln\left[\left(s_u/\sigma'_v\right)_i\right] - a - b\cdot\ln\left(OCR_i\right)\right)^2\right] \tag{9}$$

where (a^*,b^*,σ^*) are the WML estimates; w_i is the importance weight for the i-th data point; N is the total number of data points. There is no strict rule for assigning the importance weights (w_1, w_2, …, w_N). For our case, the data are the combination of the observed site-specific data $\mathbf{Y}_o$ and the generic data. Let N_s be the number of the site-specific data points and N_g be the number of the generic data points. There are N_s = 5 data points with (OCR, s_u/σ'_v) information in Table 1 and N_g = 1408 generic data points with (OCR, s_u/σ'_v) information. As a result, N = N_s + N_g = 1413. The weight of each site-specific data point is taken to be 0.5×(1//N_s) so that the total weight for all N_s site-specific data is 0.5. The weight for each generic data point is taken to be 0.5×P′(k|$\mathbf{X}_o$) so that the total weight for all N_g generic data is also 0.5. Therefore, the total weight for all N data is unity. The 0.5-0.5 rule between site-specific and generic data is adopted for demonstration in this section. Other rules such as 0.3-0.7 can be adopted. Figure 1c shows the resulting median estimate and 95% confidence interval for s_u/σ'_v based on the WML

estimates (a^*,b^*,σ^*). It is remarkable that the transformation uncertainty, quantified by the 95% confidence interval, is significantly less than that exhibited in the generic data.

5 Conclusion

This paper proposes a Bayesian data mining approach that searches a generic database for data points similar to a set of site-specific data $\mathbf{X}_o$. The similarity between the k-th generic data point and the site-specific data $\mathbf{X}_o$ is quantified by the posterior probability $P(k|\mathbf{X}_o)$. Basically, a generic data point with soil characteristics similar to those for $\mathbf{X}_o$, the posterior probability $P(k|\mathbf{X}_o)$ will be large, and the converse is also true. For illustration, a generic clay database CLAY/10/7490 is adopted as the generic database and the Stora an site (Sweden) is adopted as the local site of interest is this paper. The Bayesian data mining approach seems effective in the sense that the generic data points identified as "similar" indeed have characteristics (plasticity, degree of over-consolidation, water content, etc.) similar to the Stora an site. A weighted maximum likelihood method is further used to construct the transformation model for the Stora an site using the combination between the searched generic data and the site-specific data $\mathbf{X}_o$. The resulting transformation uncertainty is much less than that exhibited in the generic data.

References

Bjerrum, L., Embankments on Soft Ground, *Proc. Specialty Conference on Performance of Earth and Earth-Supported Structures*, ASCE, Purdue University, Lafayette, USA, 2, 1-54, 1972.

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

Ching, J. and Phoon, K.K., Correlations among Some Clay Parameters – the Multivariate Distribution, *Canadian Geotechnical Journal*, 51(6), 686-704, 2014b.

Ching, J. and Phoon, K.K., Constructing Site-specific Probabilistic Transformation Model by Bayesian Machine Learning, *ASCE Journal of Engineering Mechanics* (in review), 2018.

D'Ignazio, M., Phoon, K.K., Tan, S.A., and Lansivaara, T., Correlations for Undrained Shear Strength of Finnish Soft Clays, *Canadian Geotechnical Journal*, 53(10), 1628-1645, 2016.

Djoenaidi, W.J., *A Compendium of Soil Properties and Correlations*, MEng Thesis, University of Sydney, 1985.

Geman, S. and Geman, D., Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images, *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-741, 1984.

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

Huang, A. and Wand, M.P., Simple Marginally Noninformative Prior Distributions for Covariance Matrices, *Bayesian Analysis*, 8(2), 439-452, 2013.

Johnson, N.L., Systems of Frequency Curves Generated by Methods of Translation, *Biometrika*, 36, 149-176, 1949.

Karampatziakis, N. and Langford, J., Online Importance Weight Aware Updates, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, Barcelona, Spain, 392-399, 2011.

Kulhawy, F.H. and Mayne, P.W., *Manual on Estimating Soil Properties for Foundation Design*, Report EL6800, Electric Power Research Institute, Palo Alto, CA, 1990.

Mayne, P.W., Christopher, B.R., and DeJong, J., *Manual on Subsurface Investigations*, National Highway Institute Publication No. FHWA NHI-01-031, Federal Highway Administration, Washington, D.C., 2001.

Mesri, G. and Huvaj, N., Shear Strength Mobilized in Undrained Failure of Soft Clay and Silt Deposits, *Advances in Measurement and Modeling of Soil Behavior* (GSP 173), Ed. D.J. DeGroot et al., ASCE, 1-22, 2007.

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