Similarity Measure for Soil Properties between Different Sites

Zhiyong Yang[1] and Jianye Ching[1]

[1] Department of Civil Engineering, National Taiwan University, Taipei, Taiwan.
E-mail: zyyang@whu.edu.cn; jyching@gmail.com

Abstract: Soil material property has a significant influence on stability analysis of geotechnical structure. Consequently, geotechnical structures located in different sites with similar soil materials might share similar construction process and/or have same failure mechanisms. It is thus more rational to pay extra attention to the successful experiences or failed lessons of the geotechnical structures from those sites that has similar soil material, geological condition and hydraulic condition with the target site. This necessitate the effective recognition of historical similar sites, which is difficult for common statistical distance-based feature classification methods and deep learning methods because soil property parameter measured from general sites has the multivariate, unique, uncertain, sparse and incomplete (abbreviated as "MUSIC") characteristics. This paper proposes a hypothesis test-based method to address the similarity recognition issue for "MUSIC" site data. Database CLAY/10/7490 is employed to demonstrate the performance of the proposed method. It is found that the proposed method can identify the similar sites fairly well.

Keywords: Soil parameters; site characterization; hypothesis test; site similarity; geotechnical database

1 Introduction

Stability of geotechnical structure is highly dependent on the mechanical properties of the supporting soil such as shear strength, elastic module, compressibility, and permeability. As a result, geotechnical structures constructed at different sites with similar soils are likely to behave similarly. The experiences obtained from geotechnical structures constructed at sites similar to the current site of interest can be valuable. Geotechnical engineers should exercise their judgments based on the previous experiences. However, the experiences may not be transferrable because different sites have different soil properties. In principle, it is more desirable to learn experiences from sites with soil properties similar to the current site. This leads to a practical problem of how to identify sites with similar soil properties.

Soil behaviors manifest themselves as a series of soil properties such as index properties (unit weight, grain size distribution, Atterberg limits, relative density, water content, etc.), stress states (overburden stress, preconsolidation stress, lateral stress, etc.), shear strengths (cohesion, friction angle, undrained shear strength, sensitivity, etc.), deformation parameters (Young's modulus, Poisson ratio, consolidation parameters, etc.), permeability parameters, and in-situ test results (standard penetration test N, cone tip resistance, vane shear test results, etc.). Similarity between two sites can be investigated by comparing the soil property datasets of the two sites. Identification of similarity between two datasets is a classical problem in machine leaning, such as image identification (He and Zhang 2016), voice identification (Muda et al. 2010), signal identification (Luo et al. 2019), and other data-driven identification problems (Witten et al. 2016). Numerous methods have been developed to quantify the similarity, such as the Kullback–Leibler divergence (Goldberger et al. 2003; Sfikas et al. 2005; Hershey and Olsen 2007), Bhattacharyya distance (Møllersen et al. 2016), and deep learning algorithms (Krizhevsky et al. 2012; He et al. 2016). However, these methods may not be suitable for geotechnical data because geotechnical data are "MUSIC" – Multivariate, Unique & Uncertain, Sparse, and InComplete (Phoon 2018). A more concrete way of describing MUSIC data is to imagine an EXCEL spreadsheet containing n columns and m rows. Each column represents one test parameter (e.g., liquid limit LL, plasticity index PI, cone tip resistance q_t, etc.). Each row represents the values produced by different tests conducted in close proximity at roughly the same depth. This data structure is "multivariate" because $n > 1$ in most site investigation programs, is "sparse" because m is small, and is "incomplete" because there will be blank cells in the spreadsheet. MUSIC data poses significant challenges for statistical characterization, because most statistical methods require complete data.

To address the challenges in analyzing MUSIC data, Ching and Phoon (2019a) proposed a Gibbs sampler (GS) method that can handle incomplete data and quantify the statistical uncertainty associated with sparse data. Based on this GS method, Ching and Phoon (2019b) further proposed a similarity measure to quantify the similarity between a collection of site records and an individual record in a soil database. Here, a collection of site records can be an EXCEL spreadsheet containing multiple rows, each row (record) representing the investigation result for a certain depth at the site of interest. An individual record in a soil database is then a single row in another large EXCEL spreadsheet containing numerous records from multiple depths in generic

Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR)
Editors: Jianye Ching, Dian-Qing Li and Jie Zhang

Published by Research Publishing, Singapore.
ISBN: 978-981-11-2725-0; doi:10.3850/978-981-11-2725-0_MS2-9-cd

sites. The similarity measure reflects how similar an individual record in the generic soil database is to the investigation data at the site of interest.

An equally interesting research question is how to measure the similarity between two sites of interest, which is about the similarity measure between two collections of records. Ching and Phoon (2019b) did not address this kind of similarity. The purpose of this paper is to propose a new similarity measure that quantifies the similarity between two sites with MUSIC data. The new similarity measure is based on Ching and Phoon (2019b)'s similarity measure but a novel method of "leave-one-out hypothesis testing" is proposed. A real case history is used to demonstrate its application.

2 Review of the Gibbs Sampler Method Proposed by Ching and Phoon (2019a)

The proposed method requires the Gibbs sampler (GS) method to construct the site-specific joint probability density function (PDF) based on MUSIC site data. This section reviews the GS method proposed by Ching and Phoon (2019a). The GS method operates in the multivariate normal space, but soil data are typically non-normal. It is desirable to convert the soil data Y to normal variable X by a certain transform. Although many transforms are possible, the transform based on the cumulative distribution function (CDF) of the Johnson distribution (Johnson 1949) used by Ching and Phoon (2019a) is adopted in the current paper to maintain the consistency with our past works. Moreover, it is further assumed that the transformed site-specific property $\underline{X} = (X_1, X_2, \ldots, X_n)$ is multivariate normal:

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

where n is the dimension of the multivariate probability density function (PDF); $\underline{\mu}_s$ is the site-specific mean vector for $\underline{X}$; $\mathbf{C}_s$ is the covariance matrix for $\underline{X}$ that characterizes the site-specific correlation among (X_1, X_2, …, X_n). Consider that the site data contain n data points (the number of measured depths = n). The site data can then be viewed as an n×m matrix, denoted by $\mathbf{X}$, possibly with some empty entries. The observed site data are denoted by $\mathbf{X}_o$, and the unobserved data (empty entries) are denoted by $\mathbf{X}_u$. To construct the site-specific PDF, denoted by $f(\underline{x}|\mathbf{X}_o)$, it suffices to estimate the mean vector $\underline{\mu}_s$ and covariance matrix $\mathbf{C}_s$. Most parameter estimation methods of estimating $\underline{\mu}_s$ and $\mathbf{C}_s$ are not applicable if $\mathbf{X}_o$ is sparse (i.e., n is small) and incomplete (i.e., $\mathbf{X}$ contains empty entries). Ching and Phoon (2019a) showed that, based on non-informative conjugate priors, it is possible to draw $\underline{\mu}_s$, $\mathbf{C}_s$, and $\mathbf{X}_u$ samples sequentially using the Gibbs sampler:

$$\underline{\mu}_s \sim f(\underline{\mu}_s \mid \mathbf{C}_s, \mathbf{X}_u, \mathbf{X}_o) \qquad \mathbf{C}_s \sim f(\mathbf{C}_s \mid \underline{\mu}_s, \mathbf{X}_u, \mathbf{X}_o) \qquad \mathbf{X}_u \sim f(\mathbf{X}_u \mid \underline{\mu}_s, \mathbf{C}_s, \mathbf{X}_o) \tag{2}$$

The above sampling can be done in an analytical manner because of the use of the conjugate priors. The detailed implementation of the GS method can be found in Ching and Phoon (2019a). The GS starts with a set of initial samples of ($\underline{\mu}_s$, $\mathbf{C}_s$, $\mathbf{X}_u$), followed by repetitive implementation of Eq. (2) until a large number of GS samples of ($\underline{\mu}_s$, $\mathbf{C}_s$, $\mathbf{X}_u$) are generated. Suppose that there are K sets of ($\underline{\mu}_s$, $\mathbf{C}_s$, $\mathbf{X}_u$) samples (denoted by $\underline{\mu}_{s,k}$, $\mathbf{C}_{s,k}$ and $\mathbf{X}_{u,k}$, k = 1, 2,…, K) after the burn-in period. According to the total probability theorem, $f(\underline{x}|\mathbf{X}_o)$ can be approximated as a multivariate mixture Gaussian PDF:

$$f(\underline{x} \mid \mathbf{X}_o) \approx \frac{1}{K}\sum_{k=1}^{K} (2\pi)^{-\frac{n}{2}} \left|\mathbf{C}_{s,k}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\underline{x}-\underline{u}_{s,k}\right)^T \mathbf{C}_{s,k}^{-1}\left(\underline{x}-\underline{u}_{s,k}\right)\right] \tag{3}$$

The above GS method can be also implemented to a generic soil database. It is also assumed that the transformed property for a generic record follows the multivariate normal PDF in Eq. (1) with ($\underline{\mu}_s$, $\mathbf{C}_s$) being replaced by ($\underline{\mu}_g$, $\mathbf{C}_g$), where $\underline{\mu}_g$ is the generic mean vector, and $\mathbf{C}_g$ is the generic covariance matrix. A generic soil database, denoted by $\mathbf{D}$, can also be viewed as an N×m matrix (N is the number of generic records in $\mathbf{D}$) possibly with some empty entries. The observed generic data are denoted by $\mathbf{D}_o$, and the unobserved data (empty entries) are denoted by $\mathbf{D}_u$. The GS can be used to obtains samples for ($\underline{\mu}_{g,k}$, $\mathbf{C}_{g,k}$, $\mathbf{D}_{u,k}$), k = 1, 2,…, K. The generic PDF, denoted by $g(\underline{x}|\mathbf{D}_o)$, can be also represented by an equation similar to Eq. (3), with ($\underline{\mu}_{s,k}$, $\mathbf{C}_{s,k}$) being replaced by ($\underline{\mu}_{g,k}$, $\mathbf{C}_{g,k}$). For the current paper, because the CLAY/10/7490 database (Ching and Phoon 2014a, 2014b) is adopted as the generic database, the fixed $\underline{\mu}_g$ vector and $\mathbf{C}_g$ matrix developed in Ching and Phoon (2014b) are adopted.

3 Review of the Similarity Measure Proposed by Ching and Phoon (2019b)

The site-specific PDF $f(\underline{x}|\mathbf{X}_o)$ constructed by Eq. (3) represents the site of interest. Based on this $f(\underline{x}|\mathbf{X}_o)$, Ching and Phoon (2019b) proposed a method of measuring the similarity between the site of interest and a record in a generic database. The idea is simple: a record $\underline{x}$ in the generic database is similar to the site of interest if $f(\underline{x}|\mathbf{X}_o)$ is high. Consider two records in the generic database, $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$. If $f(\underline{x}^{(1)}|\mathbf{X}_o) > f(\underline{x}^{(2)}|\mathbf{X}_o)$, $\underline{x}^{(1)}$ is more similar to

the site of interest than $\underline{x}^{(2)}$. However, the main challenge lies in the fact that the generic database is incomplete such that $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$ may contain different observed components, e.g., $\underline{x}^{(1)} = [x_1^{(1)}, x_2^{(1)}, \text{empty}, x_4^{(1)}]$ and $\underline{x}^{(2)} = [x_1^{(2)}, \text{empty}, \text{empty}, x_4^{(2)}]$. In this scenario, $f(\underline{x}^{(1)}|\mathbf{X}_o)$ is a three-dimensional PDF, whereas $f(\underline{x}^{(2)}|\mathbf{X}_o)$ is two-dimensional. It is not feasible to compare between $f(\underline{x}^{(1)}|\mathbf{X}_o)$ and $f(\underline{x}^{(2)}|\mathbf{X}_o)$ directly. To make the comparison feasible, Ching and Phoon (2019b) proposed a similarity measure, denoted by $S(\underline{x}|\mathbf{X}_o)$, which is normalized with respect to the generic PDF $g(\underline{x}|\mathbf{D}_o)$ in a way that the mean value of $S(\underline{x}|\mathbf{X}_o)$ is always 1 regardless of the number of observed components:

$$S(\underline{x}|\mathbf{X}_o) = \frac{f(\underline{x}|\mathbf{X}_o)}{\int f(\underline{x}|\mathbf{X}_o) g(\underline{x}|\mathbf{D}_o) d\underline{x}} = \frac{\sum_{k=1}^{K} \left|\mathbf{C}_{s,k}^o\right|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\left(\underline{x}^o - \underline{\mu}_{s,k}^o\right)^T \left(\mathbf{C}_{s,k}^o\right)^{-1} \left(\underline{x}^o - \underline{\mu}_{s,k}^o\right)\right]}{\sum_{k=1}^{K} \left|\mathbf{C}_{s,k}^o + \mathbf{C}_g^o\right|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\left(\underline{\mu}_g^o - \underline{\mu}_{s,k}^o\right)^T \left(\mathbf{C}_{s,k}^o + \mathbf{C}_g^o\right)^{-1} \left(\underline{\mu}_g^o - \underline{\mu}_{s,k}^o\right)\right]} \tag{4}$$

where the superscript 'o' denotes the observed components in the record $\underline{x}$. For instance, for $\underline{x} = [x_1, x_2, \text{empty}, x_4]$, the 1st, 3rd, and 4th components of $\underline{x}$ are observed, then $\underline{u}_s^o$ is the sub-vector containing the 1st, 3rd, and 4th components of $\underline{u}_s$. Because the mean value of $S(\underline{x}|\mathbf{X}_o)$ is always 1 regardless of the number of observed components, $S(\underline{x}^{(1)}|\mathbf{X}_o)$ and $S(\underline{x}^{(2)}|\mathbf{X}_o)$ can be compared even if $\underline{x}^{(1)}$ and $\underline{x}^{(2)}$ have different observed components.

4 New Similarity Measure between Two Sites

Ching and Phoon (2019b) proposed a similarity measure that quantifies the similarity between a site of interest and an individual record in a soil database. An equally interesting research question is how to measure the similarity between two sites of interest. Ching and Phoon (2019b) did not address this kind of similarity. This section proposes such a new similarity measure between two sites of interest. Let $\mathbf{X}^{(p)}$ be the investigation data for the p-th site with n_p records (n_p rows in Excel spreadsheet), $\underline{x}^{(p)}{}_1$, $\underline{x}^{(p)}{}_2$, …, and $\underline{x}^{(p)}{}_{np}$. Let $\mathbf{X}^{(q)}$ be the investigation data for the q-th site with n_q records, $\underline{x}^{(q)}{}_1$, $\underline{x}^{(q)}{}_2$, …, and $\underline{x}^{(q)}{}_{nq}$. Let $\mathbf{X}^{(p)}{}_o$ and $\mathbf{X}^{(q)}{}_o$ denote the observable elements in $\mathbf{X}^{(p)}$ and $\mathbf{X}^{(q)}$. The basic idea is to compute the "self-similarity" measure $S(\underline{x}^{(p)}{}_i|\mathbf{X}^{(p)}{}_{o\backslash i})$ for each record in the p-th site, where $\mathbf{X}^{(p)}{}_{o\backslash i}$ is the leave-ith-out data. These self-similarity measures $S(\underline{x}^{(p)}{}_i|\mathbf{X}^{(p)}{}_{o\backslash i})$, i = 1, 2, …, n_p, quantify how similar a record in the p-th site is to the rest of the data in the p-th site. Then, the similarity measure $S(\underline{x}^{(q)}{}_i|\ \mathbf{X}^{(p)}{}_o)$ is computed for each record in the q-th site. These "cross-similarity" measures $S(\underline{x}^{(q)}{}_i|\mathbf{X}^{(p)}{}_o)$, i = 1, 2, …, n_q, quantify how similar a record in the q-th site is to the data of the p-th site. It is hypothesized that if the two sites have identical ($\underline{u}_s$, $\mathbf{C}_s$), the self-similarity measures $S(\underline{x}^{(p)}{}_i|\mathbf{X}^{(p)}{}_{o\backslash i})$ i = 1, …, n_p and cross-similarity measures $S(\underline{x}^{(q)}{}_i|\ \mathbf{X}^{(p)}{}_o)$ i = 1, …, n_q will belong to the same population. Whether the two sites indeed have identical ($\underline{u}_s$, $\mathbf{C}_s$) can be cast into a hypothesis testing problem:

$$H_0: \mu_s^{(1)} = \mu_s^{(2)} \qquad H_1: \mu_s^{(1)} \neq \mu_s^{(2)} \tag{5}$$

Under the null hypothesis H_0, the self-similarity $S(\underline{x}^{(p)}{}_i|\mathbf{X}^{(p)}{}_{o\backslash i})$ and cross-similarity $S(\underline{x}^{(q)}{}_i|\mathbf{X}^{(p)}{}_o)$ are from the same population. We further assume $\ln[S(\underline{x}^{(p)}{}_i|\mathbf{X}^{(p)}{}_{o\backslash i})]$ and $\ln[S(\underline{x}^{(q)}{}_i|\mathbf{X}^{(p)}{}_o)]$ to be normally distributed. The null hypothesis can be tested using the following t-test:

$$t = \frac{m_{self} - m_{cross}}{\sqrt{\left[(n_1-1)\cdot s_{self}^2 + (n_2-1)\cdot s_{cross}^2\right] / (n_1+n_2-2)}} \tag{6}$$

where t is distributed as the t-distribution with degree of freedom (DOF) = n_1+n_2-2;

$$m_{self} = \frac{1}{n_1}\sum_{i=1}^{n_1} \ln\left[S\left(\underline{x}_i^{(1)} \mid \mathbf{X}_{o\backslash i}^{(1)}\right)\right] \qquad m_{cross} = \frac{1}{n_2}\sum_{i=1}^{n_2} \ln\left[S\left(\underline{x}_i^{(2)} \mid \mathbf{X}_o^{(1)}\right)\right]$$
$$s_{self}^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\ln\left[S\left(\underline{x}_i^{(1)} \mid \mathbf{X}_{o\backslash i}^{(1)}\right)\right] - m_{self}\right)^2 \qquad s_{cross}^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(\ln\left[S\left(\underline{x}_i^{(2)} \mid \mathbf{X}_o^{(1)}\right)\right] - m_{cross}\right)^2 \tag{7}$$

The null hypothesis can be rejected if the P-value is less than 0.05. The P-value is equal to $2\times P(T>t)$ if $t > 0$ and equal to $2\times P(T<t)$ if $t < 0$, where T is a t random variable with DOF = n_1+n_2-2. If the null hypothesis is not rejected, it does not necessarily imply that the two sites are similar. It means that there is no sufficient evidence for the rejection. Noted that each site (i.e., site p and site q) under comparison might contain different number of samples and sparsity of their samples are also different, resulting that $S(\underline{x}^{(q)}{}_i|\ \mathbf{X}^{(p)}{}_o)$ and $S(\underline{x}^{(p)}{}_i|\ \mathbf{X}^{(q)}{}_o)$ might be significantly different. Moreover, using only $S(\underline{x}^{(q)}{}_i|\ \mathbf{X}^{(p)}{}_o)$ as similarity measure does not satisfy the symmetry properties. Hence, another hypothesis testing is adopted, but now $\mathbf{X}^{(p)}$ and $\mathbf{X}^{(q)}$ switch roles: the self-similarity is computed for $\mathbf{X}^{(q)}$, and the cross-similarity is computed for $\mathbf{X}^{(p)}$ with respect to $\mathbf{X}^{(q)}$. Another P-value can be computed. The bi-directional P-value is the smaller one between the two P-values and is taken to be the new

similarity measure between two sites. If the bi-directional P-value is greater than 0.05, there is no strong evidence to believe the two sites to be substantially different.

5 Illustrative Examples

This section employs the CLAY/10/7490 database compiled by Ching and Phoon (2014a) to illustrate the proposed method. The database contains 7490 records with 10 dimensional clay property parameters from 598 sites covering 30 countries/regions. The 10 clay property parameters are $Y_1 = \ln(LL)$, $Y_2 = \ln(PI)$, $Y_3 = LI$, $Y_4 = \ln(\sigma'_v/P_a)$, $Y_5 = \ln(\sigma'_p/P_a)$, $Y_6 = \ln(s_u/\sigma'_v)$, $Y_7 = \ln(S_t)$, $Y_8 = B_q$, $Y_9 = q_{t1}$, and $Y_{10} = q_{tu}$, where LL is liquid limit, PI is plasticity index, LI is liquidity index, σ'_v is vertical effective stress, σ'_p is preconsolidation stress, P_a is one atmosphere pressure, s_u is undrained shear strength, S_t is sensitivity, $B_q = (u_2-u_0)/(q_t-\sigma_v)$, $q_{t1} = (q_t-\sigma_v)/\sigma'_v$, $q_{tu} = (q_t-u_2)/\sigma'_v$, q_t is (corrected) cone tip resistance, σ_v is vertical total stress, u_2 is pore pressure behind the cone, and u_0 is hydrostatic pore pressure. A screening is taken to remove sites with very little information (e.g., contain less than two records or have only one dimensional data). After the screening, 420 sites are left. Among these sites, one site (a USA site) is selected to demonstrate its similarity with respect to the remaining 419 sites. 1000 number of Gibbs samples after the burn-in period are collected to construct the f(x|Xo) (i.e., K=1000). Table 1 shows the data for the USA site. Data of this site can be put into a 15×10 matrix (15 records with 10 dimension) with many empty entries (Y_7 to Y_{10} are totally empty). The bi-directional P-value between the USA site and the remaining 419 sites are computed. Among the 419 sites, there are 90 sites with bi-directional P-values with respect to the USA site greater than 0.05. These 90 sites are summarized in Table 2. Table 2 also gives the uni-directional P-values. It is found that sites with fewer records (n) and/or lower data dimensions (m) tend to have higher P-values. This is reasonable because there is no strong evidence to reject the null hypothesis when there is less information. In contrast, there are four sites with more records and/or higher data dimensions (bold in Table 2). These four sites have bi-directional P-values greater than 0.05. This suggests that under sufficient information, there is no strong evidence to reject the null hypothesis that these sites are similar to the USA site. Figure 1 shows the some Y_i-Y_j plots for the USA site and the four sites. It is found that the correlation behaviors for these four sites are indeed similar to those for the USA site.

Table 1. Data of the USA site for illustration.

Depth (m)	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})
1.97	32.85	9.62	2.68	0.07		4.37				
8.05				0.33	0.37					
8.22				0.33						
8.41				0.34						
3.05	34	16	0.63	0.47		0.93				
6.08	31	12	0.83	0.71		0.18				
6.22	35	14	0.57	0.76	3.70	0.66				
10.96	49.50	34.5	0.72	1.04	3.08	0.51				
12.19	52	25	0.72	1.14	2.94	0.46				
17.29	47	25	0.92	1.66	2.37	0.31				
19.29	48	24	0.75	1.85	2.18	0.25				
23.38	47	25	0.84	2.13	2.13	0.17				
28.96	40	20	1.20	2.75	2.75	0.18				
35.06	40	11	1.27	3.32	3.32	0.18				
38.11	45	21	0.76	3.46	3.46	0.17				

Table 2. Sites in CLAY/10/7490 that are similar to the USA site.

Country/region	Site	n	m	P-value[a]	P-value[b]	Bi-directional P-value
USA	Atlantic	6	3	0.66	0.05	0.05
Canada	Cornwall	3	4	0.80	0.05	0.05
Norway		**12**	**6**	**0.64**	**0.06**	**0.06**
Japan	Kawasaki	3	2	0.59	0.06	0.06
Canada		4	3	0.78	0.06	0.06
Canada		7	6	0.56	0.06	0.06
USA	**Boston**	**15**	**6**	**0.12**	**0.06**	**0.06**
England		18	2	0.74	0.07	0.07
USA	Atlantic	6	3	0.90	0.07	0.07
USA		7	2	0.80	0.07	0.07
New Zealand		16	3	0.92	0.08	0.08
Norway		6	2	0.56	0.08	0.08
USA		4	3	0.55	0.08	0.08
USA		9	4	0.91	0.08	0.08
USA		9	3	0.69	0.08	0.08
Singapore	KJ-BH-34	3	9	0.08	0.29	0.08
Thailand		5	5	0.88	0.09	0.09
Venezuela	Lagunillas	9	2	0.55	0.09	0.09

Canada		3	3	0.83	0.09	0.09
USA		9	3	0.50	0.09	0.09
USA		9	3	0.48	0.09	0.09
Norway	Vaterland,Olso	3	4	0.55	0.09	0.09
Norway	Gunnerungate, Olso	6	4	0.28	0.09	0.09
Canada		24	2	0.49	0.10	0.10
UK		5	2	0.76	0.10	0.10
Norway	Oslo	5	4	0.23	0.11	0.11
Norway	Trondheim	6	4	0.91	0.11	0.11
USA	Boston	3	3	0.56	0.11	0.11
USA		15	3	0.18	0.11	0.11
Canada	Ottawa, Ont	10	7	0.78	0.11	0.11
UK		7	3	0.86	0.12	0.12
Singapore	KJ-BH 6	3	9	0.73	0.12	0.12
UK		7	2	0.71	0.13	0.13
Norway		9	6	0.37	0.13	0.13
Norway		4	4	0.83	0.13	0.13
Iraq		6	2	0.47	0.13	0.13
USA		9	2	0.97	0.13	0.13
USA		8	2	0.36	0.14	0.14
UK		10	3	0.46	0.14	0.14
Iraq		6	2	0.73	0.14	0.14
Norway		7	7	0.57	0.14	0.14
Canada		8	3	0.14	0.22	0.14
Canada	Berthierville	6	9	1.00	0.15	0.15
Singapore	KJ-BH 5	3	9	0.17	0.15	0.15
Norway	Gronland, Olso	3	4	0.50	0.16	0.16
USA		5	3	0.80	0.17	0.17
Japan	Kawasaki	3	6	0.94	0.17	0.17
USA		7	3	0.49	0.17	0.17
Taiwan		6	3	0.94	0.18	0.18
Norway		7	4	0.18	0.20	0.18
USA		11	3	0.55	0.19	0.19
USA	**Boston**	**61**	**7**	**0.19**	**0.48**	**0.19**
UK		5	2	0.92	0.21	0.21
Norway	Manglerud	4	3	0.99	0.21	0.21
Norway		5	4	0.22	0.76	0.22
Norway		5	3	0.74	0.27	0.27
Japan	Bentonite, B2	10	3	0.97	0.27	0.27
USA		6	6	0.78	0.29	0.29
Mexico	**Gulf of Mexico**	**32**	**7**	**0.29**	**0.70**	**0.29**
Venezuela		10	2	0.55	0.30	0.30
Norway		9	4	0.86	0.32	0.32
Canada		20	2	0.45	0.32	0.32
Norway	Drammen	5	4	0.33	0.95	0.33
Japan	Ariake clay, A1	6	3	0.96	0.35	0.35
Thailand		3	5	0.36	0.49	0.36
UK	Gosport	7	3	0.67	0.38	0.38
USA		17	4	0.55	0.39	0.39
USA	Boston	4	5	0.50	0.40	0.40
USA		10	3	0.41	0.45	0.41
Canada	Ottawa, Ont	12	5	0.48	0.41	0.41
USA		14	4	0.44	0.50	0.44
USA	Detroit I	3	4	0.71	0.44	0.44
Canada		6	3	0.86	0.46	0.46
Japan	Mixture of Ariake clay and sand, M2	3	3	0.65	0.48	0.48
Norway	Toyen, Olso	5	4	0.93	0.49	0.49
Norway	Sarpsborg	3	4	0.56	0.89	0.56
Japan	NaN	44	3	0.62	0.56	0.56
USA		3	3	0.57	0.76	0.57
USA	Estuarine	3	4	0.62	0.58	0.58
Norway		4	2	0.74	0.61	0.61
Norway		3	2	0.98	0.63	0.63
Japan	Mixture of Ariake clay and sand, M4	8	3	0.85	0.65	0.65
Norway	Brage	3	8	0.66	0.73	0.66
Japan	Mixture of Ariake clay and sand, M3	3	3	0.75	0.70	0.70
Norway		5	3	0.86	0.71	0.71
Norway	Troll2 (North Sea)	5	8	0.84	0.76	0.76
Japan	Bentonite, B1	8	3	1.00	0.76	0.76
Japan	Ariake clay, A2	6	3	0.88	0.79	0.79
Norway		4	5	0.98	0.85	0.85
Japan	Ariake clay, A3	4	3	0.97	0.89	0.89

Note: a-USA site is treated as p-th site; b-USA site is treated as q-th site.

6. Conclusion

This paper proposes a new similarity measure between two sites. The Gibbs sampler method is used to construct the site-specific PDFs of the two sites based on their site investigation data. The similarity between the two sites are characterized by a new similarity measure called the bi-directional P-value. The hypothesis that the two sites

belong to the same population can be rejected if the bi-directional P-value is less than 0.05. The proposed method is demonstrated by a USA site. It is found that 90 sites in the CLAY/10/7490 database are not rejected with a significance level of 0.05.

Acknowledgments

The authors would like to thank TC304 Committee on Engineering Practice of Risk Assessment & Management of the International Society of Soil Mechanics and Geotechnical Engineering for developing the database 304 dB (http://140.112.12.21/issmge/Database_2010.htm) used in this study and making it available for scientific inquiry. The first author would like to thank for the financial support from Higher Education Sprout Project (Project No. 107L4000).

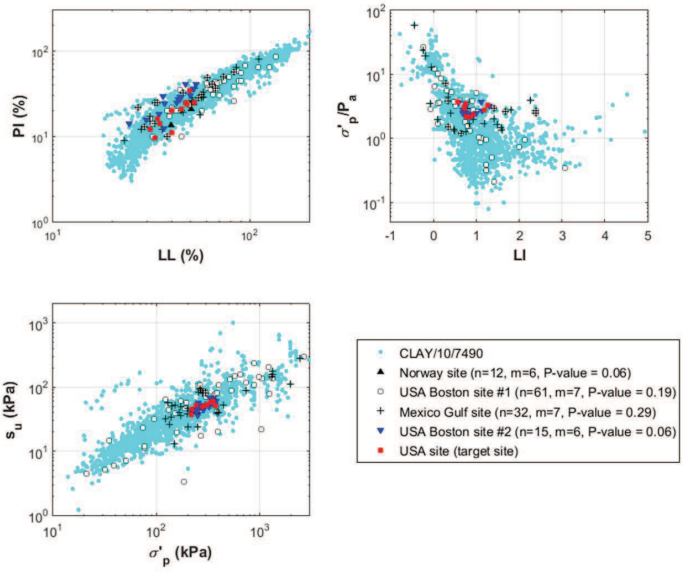

Figure 1. Some Y_i-Y_j plots.

References

Ching, J. and Phoon, K. K. (2019a). Constructing site-specific multivariate probability distribution model using Bayesian machine learning. *Journal of Engineering Mechanics*, 145(1), 04018126.

Ching, J. Y. and Phoon, K. K. (2019b). Measuring similarity between site-specific data and records from other sites. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems*, *Part A: Civil Engineering*, accepted.

Hershey, J. R. and Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. *In 2007 IEEE International Conference on Acoustics*, Speech and Signal Processing-ICASSP.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference onCcomputer Vision and Pattern Recognition*.

Muda, L., Begam, M., and Elamvazuthi, I. (2010). *Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*, arXiv preprint arXiv:1003.4083.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in Neural Information Processing Systems*, 1097-1105).

Luo, G., Yao, C., Tan, Y., and Liu, Y. (2019). Transient signal identification of HVDC transmission lines based on wavelet entropy and SVM. *The Journal of Engineering*, (16), 2414-2419.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

Goldberger, J., Gordon, S., and Greenspan, H. (2003). An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. *Proceedings of the Ninth IEEE International Conference on Computer Vision*.

Møllersen, K., Dhar, S. S., and Godtliebsen, F. (2016). *On Data-Independent Properties for Density-Based Dissimilarity Measures in Hybrid Clustering*, arXiv preprint arXiv:1609.06533.

Sfikas, G., Constantinopoulos, C., Likas, A., and Galatsanos, N. P. (2005). An analytic distance metric for Gaussian mixture models with application in image retrieval. *In International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg.

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

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