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

CONSTRUCTING MULTIVARIATE PROBABILITY DISTRIBUTION FOR SOIL PROPERTIES BASED ON SITE-SPECIFIC DATA

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

It is challenging to construct a multivariate probability distribution for soil properties based on site-specific data, because the construction of a multivariate probability distribution usually requires many data points. Site-specific data are typically sparse for this purpose. The current study investigates the possibility of constructing multivariate probability distribution for soil properties based on site-specific data. To circumvent the difficulty of data sparsity, the Bayesian approach is adopted to quantify the potentially large uncertainties. In particular, a hierarchical Bayesian framework is adopted with suitable non-informative conjugate priors so that the Markov chain Monte Carlo can be conveniently and effectively executed using the Gibbs sampler. The usefulness and effectiveness of this new method will be demonstrated using a real case.

Keywords: site characterization, multivariate probability distribution, statistical uncertainty, Bayesian analysis.

1 Introduction

Multivariate soil data obtained in site investigation are valuable because they can be used to explore the correlation behaviors among soil properties. Multivariate soil data available in the literature have been compiled to construct the multivariate probability distribution for soil properties (Ching and Phoon 2014, Liu et al. 2016, Ching et al. 2017, 2018). However, the resulting multivariate probability distribution is not site-specific and is typically developed using a generic database compiled from a large number of sites. The multivariate probability distribution for a specific site can be fairly different from that constructed by a generic database, because the latter is intended to accommodate a wide range of soil types and site conditions. It is desirable to construct a site-specific multivariate probability distribution based on site-specific data.

There are challenges for constructing a site-specific multivariate probability distribution. If we narrow down a database to a single site, the data points can be too sparse to construct the multivariate probability distribution with acceptable statistical significance. In this case, it is essential to quantify the statistical uncertainty in the parameters for the multivariate probability distribution. Another challenge is for the incomplete multivariate site-specific data. For instance, if a multivariate probability distribution for (LI, σ'_p, s_u) is to be constructed, where LI is the liquidity index, σ'_p is the preconsolidation stress, and s_u is the undrained shear strength, in principle we need multivariate data with simultaneous knowledge of (LL, LI, σ'_p, s_u) at the same

depth and reasonably close borehole/test locations. However, it is very rare that such complete multivariate data points are available during a common site investigation program. It is common to measure incomplete multivariate data points at different depths and locations, for instance, some data points have (LI, σ'_p) information, some have (LI, s_u) information, or even some only have LI information. If the data points are visualized as a spreadsheet table of size (m × 3), where m is the number of data points, incomplete multivariate data means there are missing entries in the spreadsheet table.

The purpose of this paper is to propose a Bayesian method for constructing a site-specific multivariate probability distribution that can accommodate very sparse and incomplete site-specific data while quantifying the associated large statistical uncertainties correctly.

2 Site-specific Multivariate Probability Density Function

Site-specific data are typically sparse and incomplete. Table 1 shows the site investigation results for a silty clay layer in a Taipei (Taiwan) site (Ou and Liao 1987). The depth intervals for the data range from 0.5 m to 2.6 m. Let us denote

$$Y_1 = \ln(LL) \quad Y_2 = \ln(PI) \quad Y_3 = LI \quad Y_4 = \ln(\sigma'_v/P_a) \quad Y_5 = \ln(\sigma'_p/P_a) \quad Y_6 = \ln(s_u/\sigma'_v) \tag{1.}$$

where LL = liquid limit; PI = plasticity index; LI = liquidity index; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; s_u = undrained shear strength; P_a = atmospheric pressure = 101.3 kPa. 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). Let the observed data in Table 1 be denoted by $\mathbf{Y}_o$ and the unobserved data be denoted by $\mathbf{Y}_u$. The $\mathbf{Y}_o$ data will be used to "train" the site-specific multivariate PDF model. It is desirable to convert the Y_i data to normal variable X_i by a certain transform. Although many transforms are possible, the transform based on the cumulative density function (CDF) of the Johnson distribution (Johnson 1949) used by Ching and Phoon (2014, 2018) is adopted in the current paper to maintain the consistency between the current paper and our past works:

$$X_i = \Phi^{-1}\left[F_i\left(Y_i\right)\right] \tag{2.}$$

where F_i = the cumulative distribution function (CDF) of Y_i, modeled as a Johnson distribution; Φ = CDF for a standard normal random variable. Let $\mathbf{X}_o$ and $\mathbf{X}_u$ be transformed from $\mathbf{Y}_o$ and $\mathbf{Y}_u$, respectively.

Table 1 Site investigation results for a silty clay layer at a Taipei site.

Depth (m)	Test results (training data $\mathbf{Y}_o$)					
	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)
12.8	30.1	9.1	1.20	1.26	1.71	0.37
14.8	32.8	12.8	1.43	1.43	n/a	0.36
16.1	36.4	14.5	1.24	1.54	n/a	0.33
17.8	41.9	18.9	0.90	1.68	1.79	0.25
18.3	n/a*	n/a	n/a	1.72	n/a	0.34
20.2	38.1	17.3	0.70	1.88	n/a	0.32
22.7	37.0	16.0	0.58	2.08	n/a	0.31
24.0	38.0	16.2	0.75	2.19	2.19	0.30
26.6	34.8	13.8	0.80	2.41	n/a	0.34

* n/a indicates "not available".

It is further *assumed* that site-specific X_i is normal and that, moreover, site-specific (X_1, X_2, …, X_n) is multivariate normal:

$$f\left(\underline{x}\,|\,\underline{\mu},\mathbf{C}\right)=N\left(\underline{x}\,|\,\underline{\mu},\mathbf{C}\right)=|\mathbf{C}|^{-\frac{1}{2}}(2\pi)^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\left(\underline{x}-\underline{\mu}\right)^{T}\mathbf{C}^{-1}\left(\underline{x}-\underline{\mu}\right)\right] \tag{3.}$$

where $\underline{\mu}$ is the mean vector for site-specific (X_1, X_2, …, X_n), and $\mathbf{C}$ is the covariance matrix. With the multivariate normality, elegant analytical solutions are possible.

3 Bayesian Analysis

The site-specific parameters $\underline{\mu}$ and $\mathbf{C}$ are unknown and are to be inferred by $\mathbf{X}_o$. As mentioned previously, the statistical uncertainty in the inferred $\underline{\mu}$ and $\mathbf{C}$ can be significant if $\mathbf{X}_o$ is sparse. Conjugate prior PDFs for $\underline{\mu}$ and $\mathbf{C}$ exist because (X_1, X_2, …, X_n) is assumed to be multivariate normal. The conjugate prior PDF for $\underline{\mu}_s$ is multivariate normal, whereas the conjugate prior PDF for $\mathbf{C}$ is inverse-Wishart. It is desirable that the prior PDFs $f(\underline{\mu})$ and $f(\mathbf{C})$ are non-informative. The multivariate normal prior $f(\underline{\mu})$ can be easily made non-informative by adopting large variances. However, it is challenging to make the inverse-Wishart prior $f(\mathbf{C})$ 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 $f(\mathbf{C})$ roughly non-informative.

Ching and Phoon (2018) showed that it is possible to draw ($\underline{\mu}$, $\mathbf{C}$) samples from $f(\underline{\mu},\mathbf{C}|\mathbf{X}_o)$ in an analytical manner by adopting the Gibbs sampler (GS) (Geman and Geman 1984; Gilks et al. 1996) in conjunction with the above 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, ($\underline{\mu}$, $\mathbf{C}$, $\mathbf{X}_u$), and the GS is adopted to consecutively sample them from the following conditional PDFs:

$$\underline{\mu}\sim f\left(\underline{\mu}\,|\,\mathbf{C},\mathbf{X}_u,\mathbf{X}_o\right)\qquad \mathbf{C}\sim f\left(\mathbf{C}\,|\,\underline{\mu},\mathbf{X}_u,\mathbf{X}_o\right)\qquad \mathbf{X}_u\sim f\left(\mathbf{X}_u\,|\,\underline{\mu},\mathbf{C},\mathbf{X}_o\right) \tag{4.}$$

Due to the assumed conjugate prior PDFs, the first two posterior PDFs are still with the same PDF forms: $f(\underline{\mu}|\mathbf{C},\mathbf{X}_u,\mathbf{X}_o)$ is still multivariate normal, and $f(\mathbf{C}|\underline{\mu},\mathbf{X}_u,\mathbf{X}_o)$ is still inverse-Wishart (Ching and Phoon 2018). Moreover, $f(\mathbf{X}_u|\underline{\mu},\mathbf{C},\mathbf{X}_o)$ is also multivariate normal (Ching and Phoon 2018) due to the assumed multivariate normality for $\underline{X}$. As a result, the GS algorithm can be executed conveniently because all the posterior PDFs in Eq. (3) can be sampled analytically. Let us denote the samples obtained using the GS by ($\underline{\mu}_t$, $\mathbf{C}_t$, $\mathbf{X}_{u,t}$). The GS starts with an initial sample of ($\underline{\mu}_0$, $\mathbf{C}_0$, $\mathbf{X}_{u,0}$) (time step t = 0), then it consecutively draws samples ($\underline{\mu}_t$, $\mathbf{C}_t$, $\mathbf{X}_{u,t}$) (t = 1, 2, …, T) from the conditional PDFs in Eq. (4) based on the latest parameter values. The ($\underline{\mu}_t$, $\mathbf{C}_t$) samples after the burn-in period are collected. It can be shown that these samples are distributed as $f(\underline{\mu},\mathbf{C}|\mathbf{X}_o)$. It is noteworthy that the scatter of the ($\underline{\mu}_t$, $\mathbf{C}_t$) samples quantifies the site-specific statistical uncertainty.

It is of practical interest to simulate the properties at a new depth (x_{new}) that does not appear in the training data in Table 1. By assuming $\underline{x}_{new}$ to be from the same population as $\mathbf{X}_o$, the multivariate PDF for $\underline{x}_{new}$ is also multivariate normal with mean = $\underline{\mu}$ and covariance matrix = $\mathbf{C}$.

However, ($\underline{\mu}$, $\mathbf{C}$) are uncertain and their conditional samples have been obtained by GS. Based on the total probability theorem, the conditional multivariate PDF $f(\underline{x}_{new}|\mathbf{X}_o)$ is a mixture of multivariate normal PDFs:

$$f\left(\underline{x}_{new} \mid \mathbf{X}_o\right) = \int f\left(\underline{x}_{new} \mid \underline{\mu}, \mathbf{C}\right) \cdot f\left(\underline{\mu}, \mathbf{C} \mid \mathbf{X}_o\right) \cdot d\underline{\mu} \cdot d\mathbf{C} \approx \frac{1}{T - t_b}\left[\sum_{t=t_b+1}^{T} N\left(\underline{x}_{new} \mid \underline{\mu}_t, \mathbf{C}_t\right)\right] \quad (5.)$$

where t_b is the end of the burning-period. Samples for $\underline{x}_{new}$ can be readily sampled using the following steps:

(i) Sample the t index randomly among the indices (t_b+1, t_b+2, …, T).

(ii) Sample $\underline{x}_{new} \sim N(\underline{x}_{new}|\underline{\mu}_t, \mathbf{C}_t)$, where t is the sampled t in Step (i).

These $\underline{x}_{new}$ samples have incorporated site-specific training data $\mathbf{X}_o$. Moreover, the statistical uncertainty due to sparse $\mathbf{X}_o$ is also characterized by the samples. These $\underline{x}_{new}$ samples can be further converted to the physical soil parameters $\underline{y}_{new}$ through the inverse CDF transform:

$$Y_i = F_i^{-1}\left[\Phi\left(X_i\right)\right] \quad (6.)$$

These $\underline{x}_{new}$ samples have incorporated site-specific training data $\mathbf{X}_o$. Moreover, the statistical uncertainty due to sparse $\mathbf{X}_o$ is also characterized by the samples. These $\underline{x}_{new}$ samples can be further converted to the physical soil parameters $\underline{y}_{new}$ through the inverse CDF transform:

3 Case Study

Now consider the silty clay layer for the Taipei site (Table 1). For the GS, the total sampling step size is taken to be T = 20,000, and the end of burn-in period is determined to be t_b = 1000. The behaviors of the conditional PDF $f(\underline{x}_{new}|\mathbf{X}_o)$ will be presented and compared with the measured data $\mathbf{Y}_o$.

To demonstrate the behaviors of the conditional PDF $f(\underline{x}_{new}|\mathbf{X}_o)$, consider a new depth in the same clay layer at the Taipei site with (transformed) property = $\underline{x}_{new}$ (LL, PI, LI, σ'_v/P_a, σ'_p/P_a, s_u/σ'_v). The GS samples for $\underline{x}_{new}$ can be readily obtained and converted to $\underline{y}_{new}$ samples through Eq. (6.). The total sampling step size is taken to be T = 20,000, and the end of burn-in period is determined to be t_b = 1000. Figure 1 shows the marginal cumulative density functions (CDFs) for the resulting $\underline{y}_{new}$ samples. In the figure, the marginal CDFs based on $f(\underline{x}_{new}|\mathbf{X}_o)$ are plotted as dashed lines, whereas the empirical CDFs based on the training data $\mathbf{Y}_o$ are plotted as solid lines. It is clear that the marginal PDFs for $\underline{y}_{new}$ are similar to the empirical CDFs.

Figure 1 only shows the marginal distributions for $\underline{y}_{new}$ samples. Figures **2**a and **2**b show the correlation behaviors among some $\underline{y}_{new}$ sample pairs, including the LI-σ'_p/P_a (Y_3 versus Y_5) sample pair and the σ'_v/P_a-s_u/σ'_v (Y_4 versus Y_6) sample pair. The LI-σ'_p/P_a samples spread widely because the LI-σ'_p/P_a data in $\mathbf{Y}_o$ is sparse, so the statistical uncertainty in $f(\underline{x}_{new}|\mathbf{X}_o)$ is significant. During the GS, the site-specific covariance matrix $\mathbf{C}$ is sampled. From each $\mathbf{C}$ sample, a sample for the site-specific correlation coefficient between (X_3, X_5) and (X_4, X_6), denoted by δ_{35} and δ_{46}, can be extracted. Figures **2**c and **2**d show the histograms for δ_{35} and δ_{46}. Again, the histogram for δ_{35} spread more widely because the LI-σ'_p/P_a data in $\mathbf{Y}_o$ is sparse. In general, the correlation behaviors for $\underline{y}_{new}$ are similar to those in the actual data $\mathbf{Y}_o$.

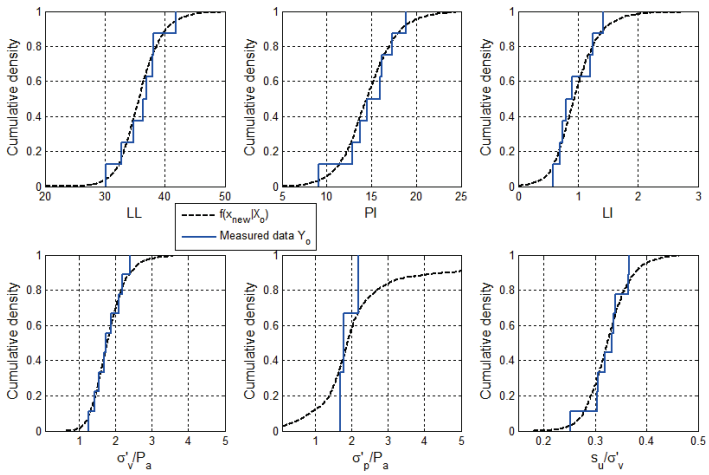

Figure 1 Marginal CDFs for the $\underline{y}_{new}$ samples (Taipei site).

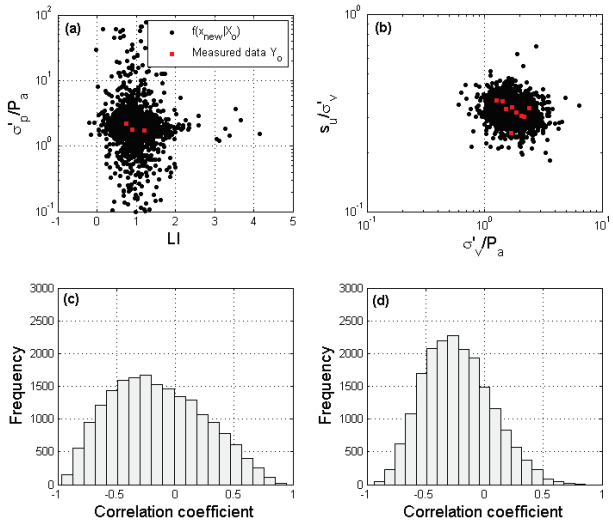

Figure 2 Correlation behaviors among some $\underline{y}_{new}$ sample pairs: (a) LI-σ'_p/P_a correlation plot; (b) σ'_v/P_a-s_u/σ'_v correlation plot; (c) histogram of δ_{35}; (d) histogram of δ_{46}.

4 Conclusion

This study proposes a novel method of constructing site-specific multivariate probability distribution for soil properties. The proposed method allows incomplete multivariate inputs. It can rigorously quantify statistical uncertainties. A real case study is used to demonstrate the usefulness of the proposed method. Analysis results show that the proposed method can effectively capture the marginal and correlation behaviors in site-specific data.

References

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

Ching, J., Lin, G.H., Phoon, K.K., and Chen, J.R., Correlations among Some Parameters of Coarse-grained Soils – the Multivariate Probability Distribution Model, *Canadian Geotechnical Journal*, 54(9), 1203-1220, 2017.

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

Ching, J., Phoon, K.K., Li, K.H., and Weng, M.C., Multivariate Probability Distribution for Some Intact Rock Properties, *Canadian Geotechnical Journal* (in review), 2018.

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.

Liu, S., Zou, H. Cai, G., Bheemasetti, B. V., Puppala, A. J., and Lin J., Multivariate Correlation among Resilient Modulus and Cone Penetration Test Parameters of Cohesive Subgrade Soils, *Engineering Geology*, 209, 128-142, 2016.

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.

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