Inferring spatial variation of soil classifications using clustered Bayesian analysis by both CPT and borehole data

Hassan Kamyab Farahbakhsh

Ph.D. student, National Taiwan University, department of Civil Engineering,
No. 1, section 4, Roosevelt Rd, Da'an District, Taipei City, Taiwan
[#]Corresponding author: hasan.kamyab@gmail.com

ABSTRACT

Soil boundary delineation is an important task in geotechnical site characterization. It can be achieved by either extracting borehole samples, conducting laboratory tests, and classifying them according to a soil classification system such as the Unified Soil Classification System (USCS) or utilizing multiple cone penetration test (CPT) soundings, and identifying soil boundaries at the soundings from the I_c (soil behavior type index) profiles. However, most soil-layer delineation methods can only take a single type of test result as the input. For instance, the well-known Markov random field (MRF) method can only take soil-type data such as sand, silt, or clay at boreholes as the input. Recognizing that soil classifications and soil properties are correlated, this paper proposes a novel coupled MRF-Bayesian framework to infer the spatial variation of USCS classifications (e.g., sand, silt, and clay) as well as soil properties by integrating both CPT and borehole data. This integrated approach leverages both CPT and borehole data to address some main challenges e.g., uncertainties and multivariate soil data input in underground stratification problems by simultaneous sampling of soil properties and soil types. The new unified framework can accommodate multivariate data, hence the new framework is compatible with the geotechnical engineering practice. The uncertainties for the spatial variation of USCS classification at sounding locations are quantified through a "layer-specific" Bayesian updating i.e., updating posterior cross-correlation behaviors for different layers (such as sand, silt, and clay), independently. In this Bayesian updating, soil-type data can provide some information about the soil properties according to the unified soil classification system. Further, the soil boundaries can be identified across the entire domain by the realization of conditional random fields of soil properties once the spatial variation of USCS classification is inferred at sounding locations, followed by a 3-dimensional Markov random field process.

Keywords: site characterization; clustered Bayesian analysis; soil boundary delineation; Markov random field; CPT data; borehole data.

1. Introduction

In geotechnical practices, soil boundary delineation stands as an important task, given the sparse and spatially variable soil types/properties. Various approaches have been developed for the soil boundary delineation problem (e.g., Phoon et al. 2003; Houlsby and Houlsby 2013; Wang et al. 2014; Ching et al. 2015; Depina et al.2016; Xiao et al. 2017; Hu and Wang 2020; Wu et al. 2021; Wei and Wang 2022). However, most existing soil delineating methods can take only one type of soil data such as soil types or soil properties/indices as the input. For instance, the well-known Markov random field (MRF) can only take soil-type data as the input. This is incompatible with typical site investigation programs where multi-type soil data, including soil properties (e.g., Atterberg limits, CPT data, etc.) and soil-type data in borehole logs, are provided. Consequently, a significant challenge in soil delineation lies in integrating all available site-specific data into the underground stratification process. This research aims to address this challenge by proposing a novel Bayesian approach capable of accommodating multivariate soil data for a comprehensive underground stratification, leveraging the entirety of available site-specific soil data for the soil boundary delineation problem.

Inferring soil properties and soil types are not two separate problems, nevertheless, they are correlated. Building on this correlation, Kamyab Farahbakhsh and Ching (2023) suggested transforming the soil boundary delineating problem into the delineation of USCS-classifications (Unified Soil Classification System) boundaries. This involves distinguishing between sand, silt, and clay based on borehole data (e.g., liquid limit (LL), plasticity index (PI), fines content (FC)) through a probabilistic framework. However, borehole data in a site are sparse i.e., statistical uncertainty. Other soil properties can also produce relevant information about soil types. In this paper, the I_c index (Robertson 2009), which are more abundant in space, is adopted as the 4^{th} soil property to assist borehole data.

The proposed MUSIC-3X framework (Ching et al. 2021; Ching et al. 2022) is employed in this research for a Bayesian updating process of integrated soil data in a site. The adopted MUSIC-3X framework assumes (LL, PI, FC, I_c) follow a single multivariate normal

distribution with a specific mean and covariance matrix for all layers. On the contrary, it is more likely that the (LL, PI, FC, and I_c) distribution for a multi-layered site cannot be represented by a single multivariate normal distribution i.e., the means and covariance matrices for different layers may vary significantly.

One possible solution for this issue is to classify observed data at i^{th} location into the one of independent clusters e.g., sand, silt, clay based on the soil type at that location and analyze them in parallel through the MUSIC-3X framework. This clustered MUSIC-3X framework is a more realistic approach compared to the original one, where all clusters are assumed to share same statistical properties.

The clustered MUSIC-3X framework raises another challenging issue i.e., transformation models (or cross-correlation) for different layers. Namely, the soil-properties simulation must be done in a "layer-specific" manner. That is, layer-specific transformation models are required to transform non-borehole data, e.g., I_c index in this research, at i^{th} depth to (LL, PI, FC) according to the soil type at that depth. Although layer-specific transformation models are precise, they require abundant layer-specific I_c vs. (LL, PI, FC) data, which are usually not available in a typical site investigation program. One possible option is to adopt a generic cross-correlation model as a prior for all target soil layers, and let this prior is further updated by the sparse layer-specific data into a posterior model through the clustered Bayesian process. The outcome is a layer-specific cross-correlation model i.e., posterior means and covariance matrices for different layers.

For this purpose, a soil database for (LL, PI, FC, I_c) is required to learn the intra-site and inter-site variability of the cross-correlation behaviors among these 4 parameters in different layers. The hierarchical Bayesian model (HBM) developed by Ching et al. (2021) is employed in this paper to learn the cross-correlation behaviors of these 4 soil indices in the soil database. The learned model, already absorbed the intra-site and inter-site cross-correlation information for different layers in the soil database, can serve as the prior model for the subsequent clustered Bayesian updating.

Once the posterior layer-specific cross-correlation models are identified, namely the "inference stage", it is still challenging to simulate conditional random fields (CRFs) of soil properties, the "CRF stage", at unexplored locations. In other words, it is unclear which layer-specific cross-correlation parameters should be adopted to simulate soil properties at an unexplored location without the knowledge of soil type at that location. The missing link between the inference and CRF stages in HBM-MUSIC-3X framework is a 3-dimensional MRF analysis proposed by Wei and Wang (2022). Given the inferred USCS-classifications at sounding locations, the soil-type samples can be simulated at unexplored locations through an MRF analysis first. Further, the layer-mannered CRFs of soil properties can be simulated at these locations.

This coupled MRF-HBM-MUSIC-3X framework not only can accommodate multivariate soil data e.g., soil type (to be elaborated later) and soil properties through a simultaneous sampling of soil types and soil properties, but also can probably address key sources of uncertainty associated with soil delineation problems, including statistical, transformation, and lithological uncertainties.

It is noteworthy to mention that soil types are sampled at unobserved borehole locations through independent 1D-MRFs in the Bayesian updating, whereas at CPT locations soil types are sampled based on the simulated soil properties (LL, PI, FC).

2. Learning of HBM

The previously compiled generic database (Kamyab Farahbakhsh and Ching 2023) consisting of four soil indices ($Y_1 = \log(LL)$, $Y_2 = \log(PI)$, $Y_3 = FC$, $Y_4 = I_c$) has been expanded to incorporate 188 sites, as depicted in Fig. 1. In this figure, site-specific data are represented with different colors and markers.

Given this extended database (Y^o), a Johnson family transformation (Johnson 1949; Ching and Phoon 2014) is

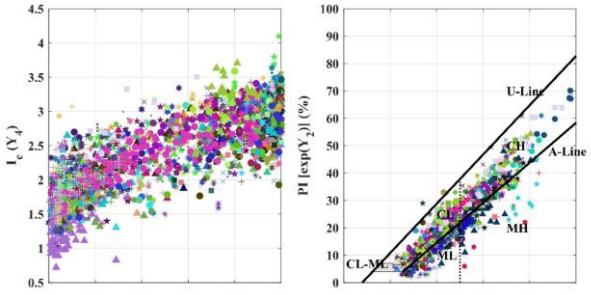

Figure 1. Generic database of 4 soil indices (LL, PI, FC, I_c)

first adopted to convert the database into the standard normal space (X^o), and then the hierarchical Bayesian model (HBM) analysis is performed to learn the intra-site and inter-site cross-correlation behaviors of these parameters. For this purpose, the HBM assumes that the i^{th} site in the database follows its own site-specific model characterized by ($\underline{\mu}_i$, C_i), where $\underline{\mu}_i \in \mathbf{R}^{4\times1}$ and $C_i \in \mathbf{R}^{4\times4}$ denote the mean vector and covariance matrix of the i^{th} site, respectively. The technical details for the HBM can be found in Ching et al. (2021).

To validate the HBM model, one can consider a hypothetical future site for which the learned HBM model can be employed to simulate its site-specific ($\underline{\mu}_i$, C_i). This site-specific ($\underline{\mu}_i$, C_i) can be further used to simulate site-specific soil properties (Y_1, …, Y_4) as shown in Fig. 2. The simulated data-points are depicted with different colors representing sand, silt, and clay

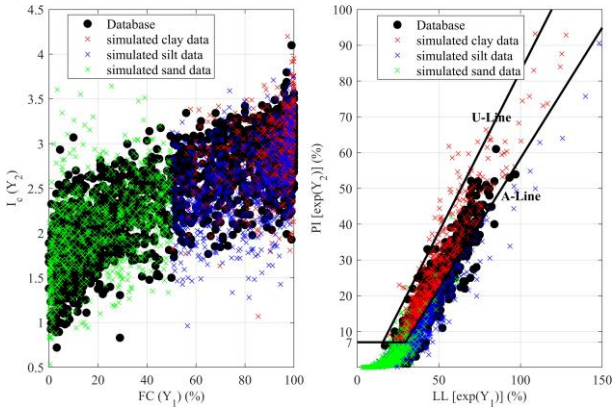

Figure 2. Simulated (LL, PI, FC, I_c) vs generic database

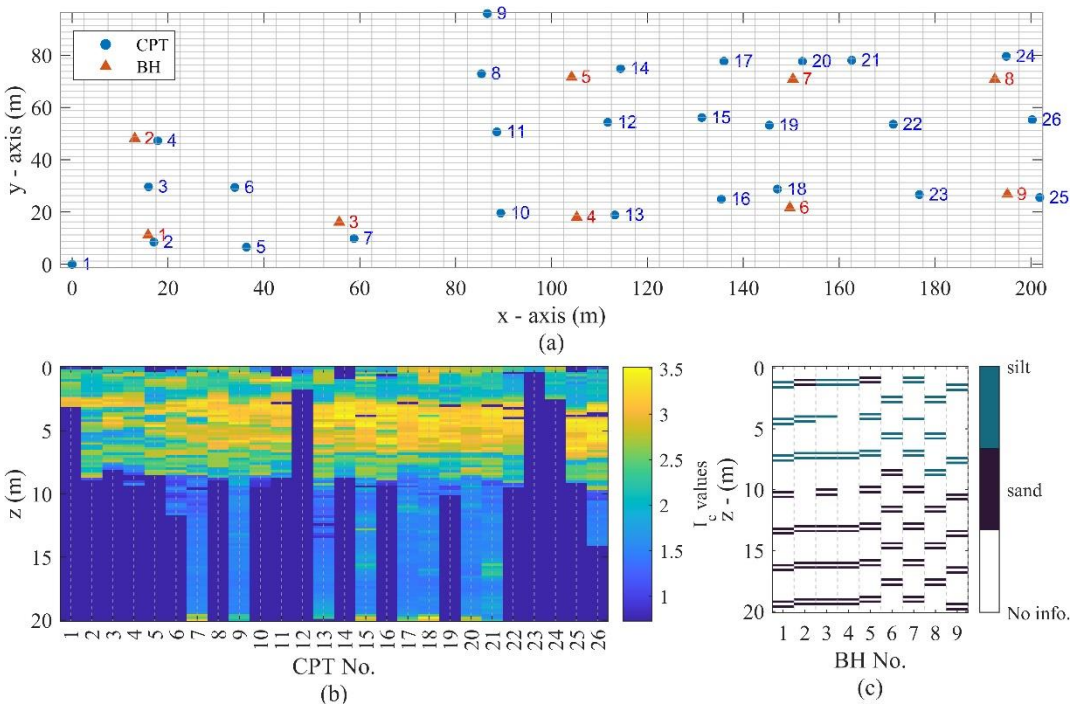

Figure 3. Target-site data in Christchurch, New Zealand: (a) site plan; (b) I_c-heat map at CPT locations; and (c) observed USCS-classifications at borehole locations

cases. Although some sand data-points with relatively large I_c values (e.g., exceeding 3.0) are generated, simulated sand data-points are more likely to have lower I_c values compared to silt/clay data-points. Moreover, there is a region with low LL and PI in Fig. 2 in which there are lots of hypothetical data (e.g., the majority of them are simulated sand data-points and some are simulated silt data-points) with PI values less than 7.0, but real data are rare. This may be because (LL, PI) values for either real sandy or non-plastic silty soil cases are usually not reported, but (LL, PI) for a hypothetical sandy/silty soil case are still simulated.

3. Clustered Bayesian updating by site-specific data

3.1. A target site

A target site (885 Colombo Street project, project No: PNZ2032) in Christchurch, New Zealand, has been selected to demonstrate the implementation of the proposed method in this paper. The target-site data are extracted from the website of the New Zealand Geotechnical Database (NZGD 2023). The plan view of in-situ test locations is presented in Fig. 3(a). The dataset comprises 26 CPT–sounding logs with different penetration depths and 9 borehole logs. Figs. 3(b) & 3(c) provide a perspective on the subsurface decomposition, probably indicating that the upper half and lower half of the subsurface are predominantly composed of silt and sand, respectively. While (LL, PI, FC) values are not documented in borehole logs, USCS-classifications at certain depths are available.

3.2. Auto-correlation function

It is noteworthy to mention that the MUSIC-3X framework necessitates the identification of the parameters for the auto-correlation function (ACF). In this study, the two-parameter Whitle-Matérn (WM) model (Stein 1999; Guttorp and Gneiting 2006) is employed as the ACF. For the target-site data (i.e., I_c values at CPT soundings), the vertical ACF parameters $\theta_z = (\delta_z, \nu_z)$ are identified as (0.76 m, 1.16) through the Gaussian process regression proposed by Ching et al. (2023). Here, "δ" and "ν" denote the scale of fluctuation and smoothness parameters, respectively. Although the horizontal ACF parameters $\theta_h = (\delta_h, \nu_h)$ are not identifiable due to the considerable distance between CPT locations, these parameters are assumed to be θ_h = (5.2m, 0.32) for demonstration.

3.3. Clustered Bayesian updating

The clustered Bayesian framework proposed in this research operates in a layer-specific manner i.e., it assumes different layers may have different cross-correlation parameter values. Compared to the original MUSIC-3X framework, this is a more realistic approach where the statistical parameters e.g., mean, standard deviation, and coefficient of correlation are allowed to vary layer by layer, independently. More specifically, the HBM model serves as the prior cross-correlation model for different layers, and it is subsequently updated to a posterior model by sparse layer-specific data of the target site. Fig. 4 illustrates the posterior layer-specific statistical parameters (e.g., the mean, standard deviation, and the coefficient of correlation) of 4 soil indices in the X-space. Figs. 4(a) & (d) depict the layer-specific posterior samples of the mean for fines content (X_3) vs I_c

(X_4) and liquid limit (X_1) vs plasticity index(X_2) in the standard normal space for different clusters, respectively.

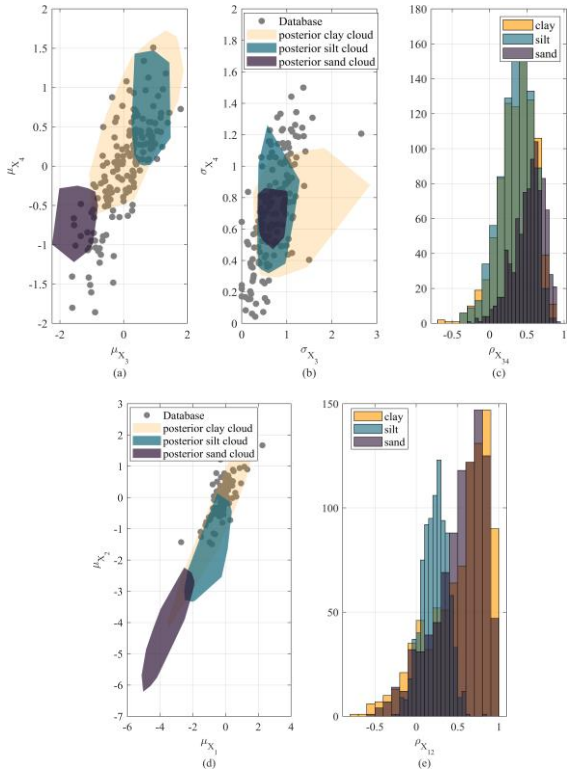

Figure 4. Clustered Bayesian updating: (a) the means of X_3 and X_4; (b) the standard deviations of X_3 and X_4; (c) the coefficients of correlation between X_3 and X_4 ; (d) the means of X_1 and X_2; and (e) the coefficients of correlation between X_1 and X_2

By comparison, the posterior silt clouds of (μ_{X3}, μ_{X4}) and (μ_{X1}, μ_{X2}) are relatively positioned in the top right of the sand clouds. This observation aligns with expectations, considering that I_c values for silts are typically anticipated to be larger than those for sands (see Fig. 1 for reference), and fines content values for silts should be larger than those for sands. In addition, plasticity index values for clays are usually larger than silt and sand. Figs. 5 & 6 illustrate the USCS-classifications probability and 95% confidence intervals (95%CIs) of soil properties variation with depth at CPT #20 and BH#2 locations, respectively.

In Figs. 5(a) & 5(b), it is evident that silt is the most probable USCS-classification between depths 3m to 7m, where the I_c values at CPT#20 for these depths are relatively large e.g., it exceeds 3.0 at some depths. Similarly, the most probable classification from a depth of 3m to 10m at BH#2 is silt (Figs. 6(a) & (b)), where 4 observed silt depths are available. At these depths, the probabilities of observed USCS-classifications are 100%. Fig. 6(f) shows the median profile of simulated I_c values at these depths are relatively large, compared to other depths.

Table 1. USCS-wise constraints for (LL, PI, FC)

Soil type	LL	PI	FC
Sand	-	-	FC<50
Silt	-	PI ≤ A-line for L≥29.6% PI ≤ 7% for LL < 29.6%	FC≥50
Clay	-	A-line < PI ≤ U-line PI > 7%	FC≥50

It has been highlighted that the proposed framework in this research can accommodate multivariate soil data including the soil-type data. This type of information is employed in the Bayesian updating process by incorporating USCS-wise constraints, as outlined in Table 1, on simulated LL, PI, and FC at observed borehole depths. Fig.7 illustrates the simulated (LL, PI, FC, I_c) at depths of 1m and 4m, where sand and silt classifications, respectively, are observed at BH#2.

4. Markov random field

Given the layer-specific posterior cross-correlation parameters samples of soil properties, MUSIC-3X framework is unable to simulate the CRFs of soil properties at unexplored locations, where soil types are unidentified in advance i.e., it is unclear which layer-specific cross-correlation parameters are deployed to simulate the soil properties. An MRF analysis can fill this gap i.e., before the CRF stage, an MRF analysis can be conducted to simulate the USCS-classifications at unexplored locations.

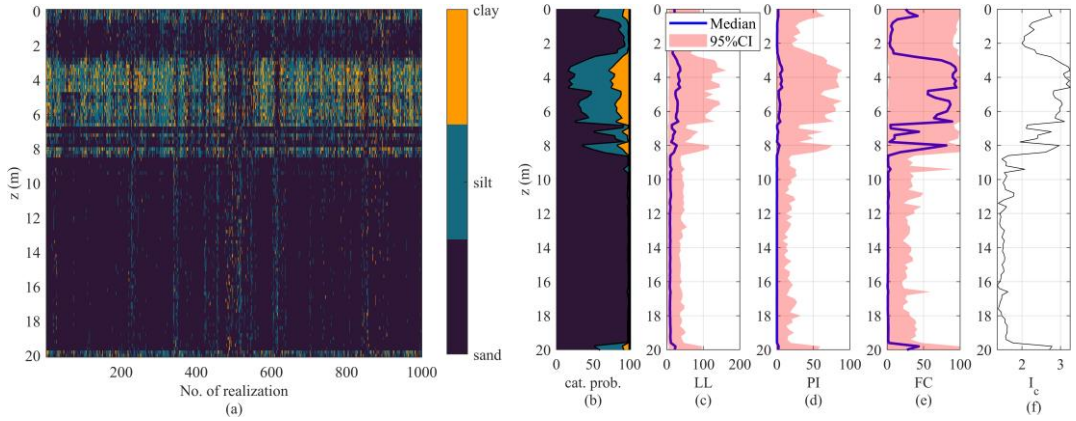

Figure 5. Probability of USCS-classifications and 95% CIs at CPT#20: (a) realization of USCS-classifications; (b) probability of USCS-classifications; (c) 95%CI for LL; (d) 95%CI for PI; (e) 95%CI for FC; and (f) Actual profile of I_c.

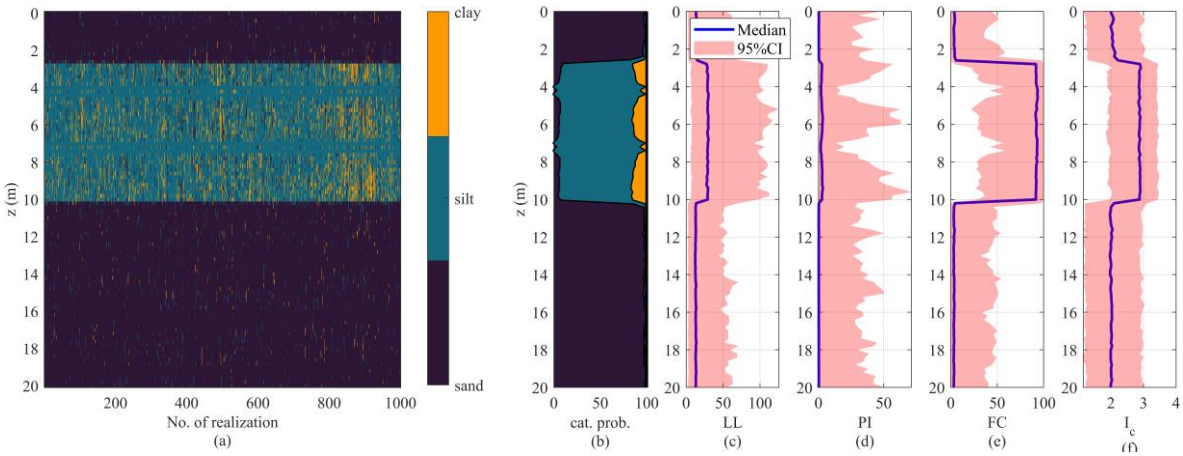

Figure 6. Probability of USCS-classifications and 95% CIs at BH#2: (a) realization of USCS-classifications; (b) probability of USCS-classifications; (c) 95%CI for LL; (d) 95%CI for PI; (e) 95%CI for FC; and (f) 95%CI for I_c.

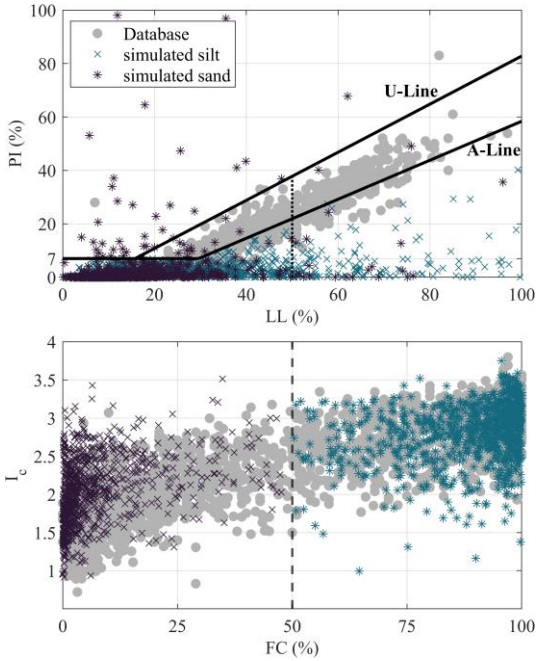

Figure 7. Simulated soil properties at observed borehole locations

A necessary step for MRF analyses is a "pre-test", where an informative prior (a multivariate normal PDF) of granularity coefficients β_i (i=1, 2, …, 7) is constructed. Fig. 8 illustrates the effective directions of different β_i in 3-dimension.

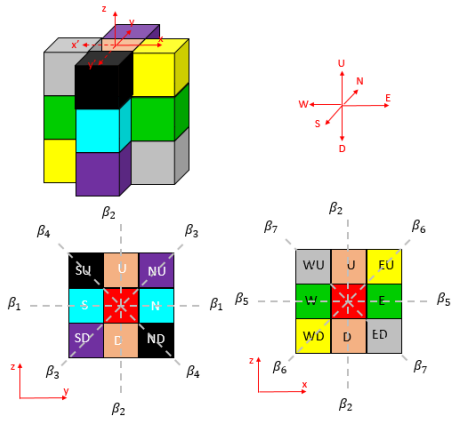

Figure 8. The most probable USCS-classification profiles at sounding locations

To reduce the computational cost, the pre-test can be performed once instead of N_{infer}-times (N_{infer} =1000 denotes the number of inferences stage samples). For this purpose, the most probable profiles at sounding locations are considered as shown in Fig. 9. The mean vector ($\mu_{\beta i}$) and covariance matrix ($\Sigma_{\beta i}$) for the prior PDF of β~MVN ($\mu_{\beta i}$, $\Sigma_{\beta i}$) are set to $\mu_{\beta i}$ = [0.08, 3.16, 0.00, -0.07, 3.86, -1.29, -1.72] and $\Sigma_{\beta i}$ =diag(0.10, 0.41, 0.08, 0.29, 0.18, 0.43). Further, N_{infer} samples of USCS-classification can be inferred at unexplored locations. The details have been elaborated in Wei and Wang (2022).

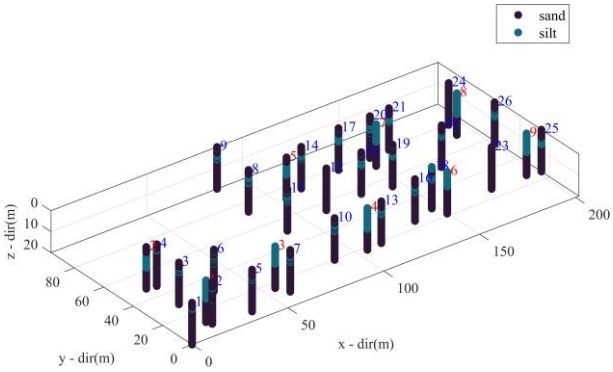

Figure 9. Most probable profile at soundings locations

5. Conditional random fields

Given the USCS-classification samples at unexplored locations, the CRFs of soil properties can subsequently be realized. Fig. 10 illustrates the silt probability at a depth of 8.2m. Borehole locations are marked by red lines in this figure, in addition to 3 CPT locations. Fig. 10 indicates that the silt probability distribution is sharp (e.g., displaying either a peak or minimum) near borehole locations, probably due to 2 reasons: (a) at borehole locations, there are observed USCS-classifications i.e., no uncertainties, and (b) USCS-classifications at unobserved borehole locations are simulated based on the independent 1D-MRF during the inference stage.

Fig. 11 illustrates the sand probability distribution through a section that crosses at y=50m. It is noteworthy to mention that to construct the soil-type probability distribution in this figure, the N_{infer} USCS-classification

samples are obtained based on the simulated CRFs of (LL, PI, FC). Moreover, the I_c profiles of close-by CPTs are overlaid in this figure for comparison. Some consistency can be observed in Fig. 11. For instance, the probability of sand decreases where the I_c values increase at CPT locations.

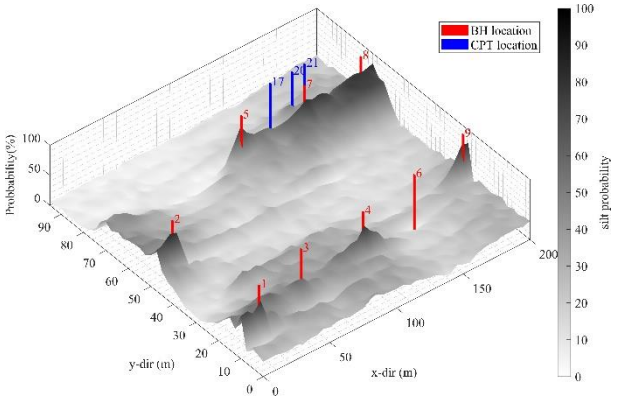

Figure 10. Silt probability surface at z = 8.2m

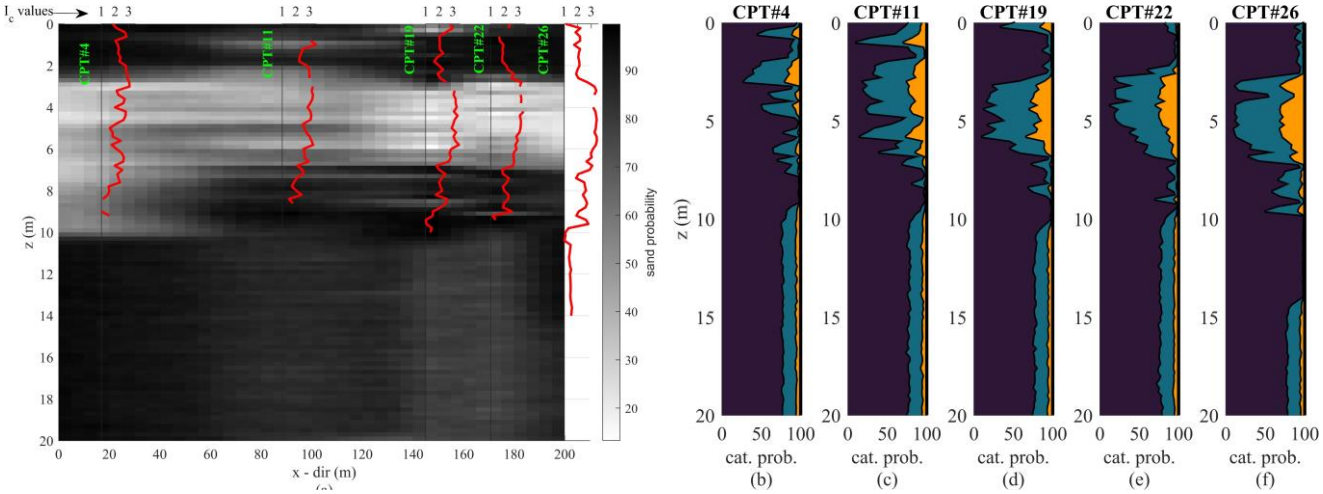

Figure 11. sand probability distribution at y = 50m plane
and probabilistic USCS-classifications profiles near close-by CPT soundings

Moreover, I_c values near CPT#4 between a depth of 6m to 10m are relatively low, anticipating the probability of sand is high. On the contrary, Fig. 11 indicates that the sand probability is relatively low. This inconsistency probably can be explained by the nearby BH#2 (see Fig. 6 for the reference) whereas the silt is highly probable at these depths.

As has been highlighted earlier, site-specific data e.g., I_c profiles and borehole logs suggest that the lower half of the underground is predominated by sand. The illustrated sand probability in Fig.11 aligns with this engineering perception. One realization of soil properties has been illustrated for the cross-section at y = 50m in Fig. 12. It is clear that the realized FC and I_c values in shallow depths are relatively larger than those in deeper depths.

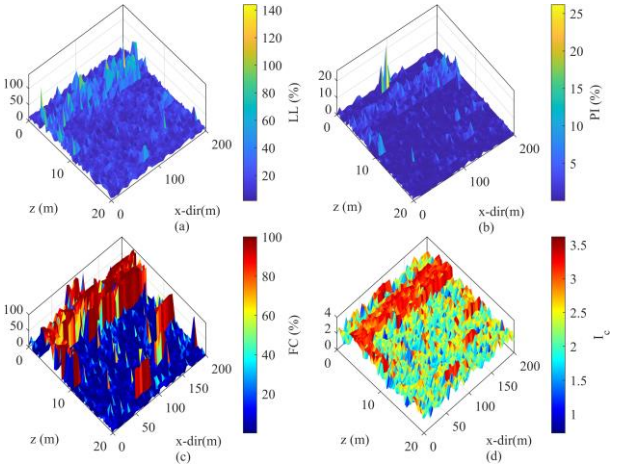

Figure 12. One realization of soil properties
at the section y = 50m

6. Conclusions

Soil properties and soil types simulation problems are correlated. Based on this correlation, the soil delineating problem can be converted to the USCS-classification boundaries problem. In the current paper, a novel probabilistic framework has been introduced to address one important challenge e.g., considering taking the advantage of all available soil data contribution for soil boundary delineating problems. For this purpose, a supporting generic database of liquid limit, plasticity index, fines content, and Robertson soil behavior type index has been compiled, and the HBM is adopted to learn the inter-site and intra-site cross-correlation behaviors of these parameters for different layers e.g., sand, silt, and clay. The learned HBM model, which has absorbed the intra-site and inter-site cross-correlation information in the soil database, further serves as the prior model for target layers in the subsequent Bayesian updating. Conditioning on the layer-specific data, this prior model is updated into a layer-specific posterior model i.e., discernible layer-specific cross-correlation parameters for different layers through the clustered MUSIC-3X Bayesian updating framework proposed by Ching et al. (2022).

Moreover, the HBM-MUSIC-3X method, as the core engine for layer-manner simulating of soil properties, is coupled with the MRF framework, proposed by Wei and Wang (2022), to simulate soil classifications and soil

properties simultaneously at unexplored locations. Further, the CRFs of soil properties (LL, PI, FC, I_c) are sampled based on the inferred 3D-MRF results and posterior layer-specific cross-correlation parameters.

Essentially, these coupled frameworks are implemented to infer the 3-dimensional spatial variation of USCS classifications (e.g., sand, silt, and clay) as well as soil properties for a real site in Christchurch, NZ. The analysis outcome seems reasonable e.g., there are some consistency between the outcome of analysis and the input data. For instance, probability of sand increases with the increase of I_c values at CPT locations.

Acknowledgments

The author would like to thank Prof. Jianye Ching, my advisor, for his great contribution for this paper. Also, the author thanks Dr. Szu-Wei Lee for his efforts in compiling the generic database and also Dr. Xingxing Wei for developing 3D-MRF codes used in this study.

References

Ching, J., Yoshida, I., & Phoon, K. K. (2023). "Comparison of trend models for geotechnical spatial variability: Sparse Bayesian learning vs. Gaussian process regression." Gondwana Research, 123, 174-183.

Ching, J., Phoon, K. K., Yang, Z., and Stuedlein, A. W. (2022). "Quasi-site-specific multivariate probability distribution model for sparse, incomplete, and three-dimensional spatially varying soil data." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), 53-76.

Ching, J., Wu, S., and Phoon, K. K. (2021). "Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model." J. Eng. Mech., 147(10), 04021069.

Ching, J., Wang, J. S., Juang, C. H., and Ku, C. S. (2015). "Cone Penetration Test (CPT)-based stratigraphic profiling using the wavelet transform modulus maxima method." Can. Geotech. J., 52(12), 1993-2007.

Ching, J. and Phoon, K. K. (2014). "Correlations among some clay parameters – the multivariate distribution." Can. Geotech. J., 51(6), 686-704.

Depina, I., Le, T. M. H., Eiksund, G., & Strøm, P. (2016). "Cone penetration data classification with Bayesian Mixture Analysis." Georisk: Assessment and management of risk for engineered systems and geohazards, 10(1), 27-41.

Guttorp, P. and Gneiting, T. (2006). "Studies in the history of probability and statistics XLIX on the Matérn correlation family." Biometrika, 93(4), 989-995.

Houlsby, N. M. T. and Houslby, G. T. (2013). "Statistical fitting of undrained shear strength data." Géotechnique, 63(14), 1253-1263.

Hu, Y. and Wang, Y. (2020). "Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation." Computers and Geotechnics, 124, 103634.

Kamyab Farahbakhsh, H. and Ching, J. (2023). "Inferring Spatial Variation of Soil Classification by Both CPT and Borehole Data". In *Geo-Risk 2023*, 142-151.Washington D.C., USA.

Johnson, N. L. (1949). "Systems of frequency curves generated by methods of translation." Biometrika, 36(1/2), 149-176.

NZGD (2022). New Zealand Geotechnical Database: https://www.nzgd.org.nz/.

Phoon, K. K., Quek, S. T., and An, P. (2003). "Identification of statistically homogeneous soil layers using modified Bartlett statistics." J. Geotech. Geoenviron. Eng., 129(7), 649-659.

Robertson, P. K. (2009). "Interpretation of cone penetration tests - a unified approach." Can. Geotech. J., 46(11), 1337-1355.

Wang, Y., Huang, K., and Cao, Z. (2014). "Bayesian identification of soil strata in London clay." Géotechnique, 64(3), 239-246.

Wei, X., & Wang, H. (2022). "Stochastic stratigraphic modeling using Bayesian machine learning." Engineering Geology, 307, 106789.

Wu, S., Zhang, J. M., and Wang, R. (2021). "Machine learning method for CPTu based 3D stratification of New Zealand geotechnical database sites." Advanced Eng. Informatics, 50, 101397.

Xiao, T., Zhang, L. M., Li, X. Y., and Li, D. Q. (2017). "Probabilistic stratification modeling in geotechnical site characterization." ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A: Civil Eng., 3(4), 04017019.