A case study for data-driven soil-layer delineation

Jianye Ching[1#], *Hassan* Kamyab Farahbakhsh[1], *Giovanna* Vessia[2], and *Diego* Di Curzio[3]

[1]*National Taiwan University, Department of Civil Engineering, Taipei, Taiwan*
[2]*University "G. d'Annunzio" of Chieti-Pescara, Department of Engineering and Geology, Chieti, Italy*
[3]*Delft University of Technology, Department of Water Management, Delft, the Netherlands*
[#]*Corresponding author: jyching@gmail.com*

ABSTRACT

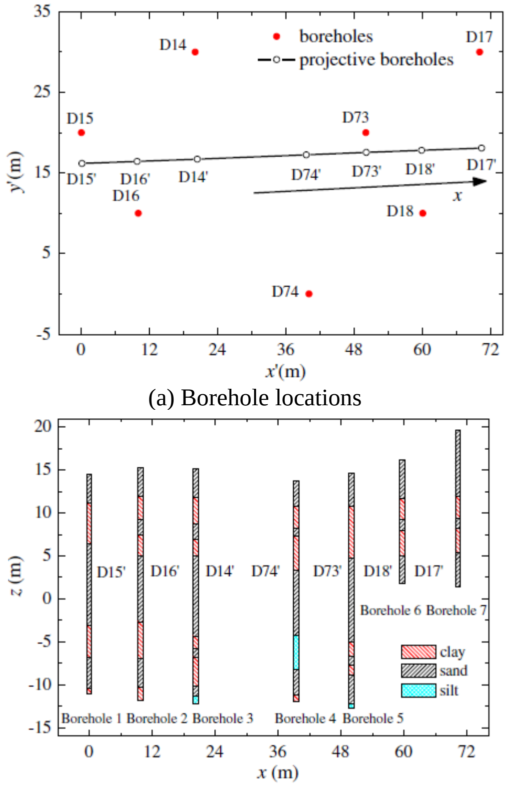
In the past, soil-layer delineation methods can usually only take a single type of input data, e.g., soil-type data at boreholes. However, this does not fit in the geotechnical engineering practice where multiple types of data are usually available during site investigation (e.g., borehole data and cone penetration test data are both available). This paper adopts a novel data-driven method for soil-layer delineation that accommodates multiple types of site investigation data. The basic idea is to include liquid limit (LL), plasticity index (PI), and fines content (FC) into the soil parameters of analysis. According to the Unified Soil Classification System (USCS), the information of (LL, PI, FC) can be used to determine whether the soil is sand, silt, or clay. As a result, the conditional random field simulation results for (LL, PI, FC) can be used to delineate sand, silt, and clay layers. If extra soil parameters (such as cone penetration test results) are incorporated, the novel method can accommodate multiple types of site investigation data. A real example of the Fucino Basin in Italy is adopted to demonstrate the application of the novel data-driven soil-delineation method.

Keywords: soil-layer delineation; site characterization; spatial variability; conditional random field.

1. Introduction

In geotechnical engineering, there are two main tasks in site characterization. One is to delineate soil layers based on site-specific data. The other is to determine the spatial variation of soil parameters based on the data within each delineated soil layer. The first task can be achieved by soil-layer delineation methods. In the past, various soil-layer delineation methods have been proposed, such as the coupled Markov chain (CMC) methods (e.g., Qi et al. 2016; Li et al. 2019; Varkey et al. 2023a), Markov random field (MRF) methods (e.g., Li et al. 2016a; Zhao et al. 2021; Wei and Wang 2022), methods based on a training image (e.g., Caers and Zhang 2004; Hu and Chugunova 2008; Shi and Wang 2021a,b), CPT-based SBT methods (CPT stands for cone penetration test, and SBT stands for soil behavior type) (e.g., Li et al. 2016b; Wang et al. 2020; Varkey et al. 2023b), etc.

A common feature of the aforementioned soil-layer delineation methods is that they only take a single type of input data. For example, the CMC, MRF, and training-image methods only take the soil-type data at boreholes as input. Figure 1 shows the locations and soil-type data of the boreholes at a site in Perth city, Australia. The site in Figure 1 was analyzed by Qi et al. (2016) using the CMC method (Elfeki and Dekking 2001). The CMC method takes the soil-type data at the boreholes as the input to simulate the soil types at unexplored locations using the Markov chain theory. The task of soil-layer delineation is done once the soil types at unexplored locations are simulated (e.g., Figure 2). In contrast, the CPT-based SBT methods only take the CPT data as input to simulate conditional random fields of CPT parameters at unexplored locations. The SBTs at unexplored locations can be determined based on the simulated CPT parameters according to the Robertson's SBT chart (Robertson 2009). The task of soil-layer delineation is done once the SBTs at unexplored locations are simulated.



(a) Borehole locations

(b) Soil-type data at boreholes

Figure 1. Soil-type data of the boreholes at a site in Perth city, Australia (source: Qi et al. 2016).

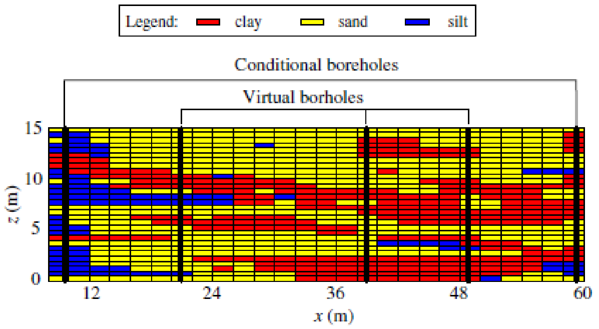


Figure 2. Realization of soil types at unexplored locations using the CMC method (source: Qi et al. 2016).

However, a routine geotechnical site investigation program usually consists of multiple types of site-specific data, e.g., borehole data such as soil type, Atterberg limits, water content, fines content, SPT N value, undrained shear strength, preconsolidation stress, etc. and CPT data such as cone tip resistance, sleeve friction, CPT pore water pressure, etc. The CMC, MRF, and training-image methods can only take the soil-type data at boreholes as input, whereas the CPT-based SBT methods only take the CPT data as input. The limitation of these soil-layer delineation methods is evident. It is not clear how to conduct soil-layer delineation using multiple types of site-specific data (e.g., take both borehole soil-type and CPT data as inputs) with these methods. It is also not clear how to incorporate other soil parameters (such as SPT N value) that may be correlated to the soil type into the analysis. There is a need to develop a new soil-type delineation method that can take multiple types of site-specific data to fit in the geotechnical engineering practice.

Recently, Kamyab Farahbakhsh and Ching (2024) developed a new soil-layer delineation method that can take multiple types of site-specific data. This method adopts the HBM-MUSIC-3X method (Ching et al. 2022) previously developed by the second author as the main analysis engine. The HBM-MUSIC-3X method has the following features:

- It can model the cross-correlation between the multivariate soil parameters in site investigation (e.g., Atterberg limits, water content, SPT N value, CPT parameters, etc.). If there are sufficient pairwise site-specific data, the cross-correlation parameters (such as the covariance matrix) can be estimated by the site-specific data.
- It can model the spatial-correlation (or auto-correlation) of the soil parameters. If there are sufficient CPTs, the auto-correlation parameters such as the scale of fluctuation can be estimated by the CPT data).
- In the case that there are insufficient pairwise site-specific data (which is usually the case), it can learn the cross-correlation behaviors from (generic) sites in a soil database using the hierarchical Bayesian model (HBM) (Ching et al. 2021). The HBM learning outcome can be transferred to the target site to reduce the uncertainty in the cross-correlation.

If the cross-correlation and auto-correlation parameters of the target site are known (or estimated), the HBM-MUSIC-3X method can further simulate the conditional random fields of the soil parameters by conditioning on the site-specific borehole and CPT data.

The basic idea proposed by Kamyab Farahbakhsh and Ching (2024) of implementing HBM-MUSIC-3X to soil-layer delineation is simple. If liquid limit (LL), plasticity index (PI), and fines content (FC) are considered in HBM-MUSIC-3X, it can therefore simulate the conditional random fields of (LL, PI, FC) by conditioning on the site-specific borehole and CPT data. Because the USCS main soil type (e.g., sand, silt, and clay) can be determined based on (LL, PI, FC), the conditional random fields of (LL, PI, FC) can be converted to the conditional soil-type field. The task of soil-layer delineation is done once the conditional soil-type field at unexplored locations is simulated. Moreover, if extra soil parameters (e.g., I_c and SPT N; I_c is the CPT SBT index proposed by Robertson 2009) are included in the analysis, the HBM-MUSIC-3X method can consider the cross-correlation among (LL, PI, FC, I_c, SPT N). By doing so, the new method can take multiple types of site-specific data into the analysis and fuse all available information to simulate the conditional random fields of (LL, PI, FC). This circumvents the main limitation of the past soil-layer delineation methods that they can only take a single type of input data. Moreover, the new method can simulate the conditional random fields of (I_c, SPT N) as well, so the second task of site characterization (simulate the spatial variation of soil parameters) is also achieved in the meantime.

There are two technical gaps in the new method that cannot be addressed by the original HBM-MUSIC-3X method. First, the ground is categorized into sand, silt, and clay. Each soil type has its own cross-correlation parameters, so some clustering analysis is needed in the new method. Kamyab Farahbakhsh and Ching (2024) developed a clustered-HBM-MUSIC-3X method to fill this gap. Second, the original HBM-MUSIC-3X method does not model the soil-layer transition behavior (e.g., the transition probability matrix in the CMC method). Some probabilistic soil-layer transition model is needed in the new method. Kamyab Farahbakhsh and Ching (2024) adopted the Markov random field (MRF) model to fill this gap. These technical details are not presented in the current paper. Interested readers are referred to Kamyab Farahbakhsh and Ching (2024) for these details. The main purpose of the current paper is to present the analysis results for the real case study of the Fucino Basin in Italy (Abruzzo, L'AQ).

2. Real case study

In the real case study, the investigated area is the Fucino Basin, located in Abruzzo Region, central Italy. It is a tectonic basin filled with hundreds of meters of soft lacustrine deposits. For further geological details refer to Boncio et al. (2018). Over an investigation region of roughly 8000 m × 8000 m (see Figure 3), 165 boreholes and 15 cone penetration tests (CPTs) are conducted. At each borehole, only quantitative soil-type data (gravel, sand, silt, and clay) are available at certain depths. Figure 4 shows the soil-type data at the boreholes. At each CPT, the SBT index (I_c) data are available. The boreholes and

CPTs are sparse: only a small fraction of the total area of 8000 m × 8000 m is investigated. Given the sparse data, the soil types at unexplored locations are highly uncertain. With the significant uncertainty in soil types, it is challenging to assess the liquefaction risk of the Fùcino Basin because the liquefaction potential of soil is closely related to its soil type. The main purpose of the case study is to simulate the soil types of unexplored locations based on the sparse regional investigation data. To simplify the illustration, only the simulation of the soil types along the A-A section in Figure 3 is demonstrated in the current paper. Figure 5 shows a zoom-in plot around the A-A section. There are three CPTs nearby the A-A section. Their I_c profiles are shown in Figure 6. Because the site investigation includes boreholes and CPTs, we consider the following four soil parameters: (LL, PI, FC, I_c). Note that (LL, PI, FC) are necessary for our soil-layer delineation method. To consider the CPT data in the investigated region, the parameter I_c is also included.

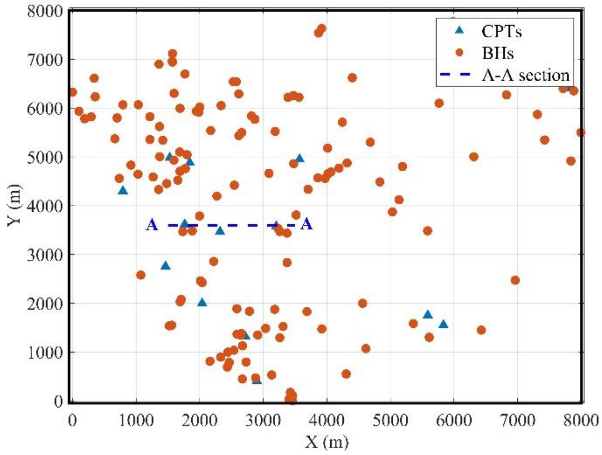


Figure 3. Plan view of the investigated region in the Fucino Basin, Italy.

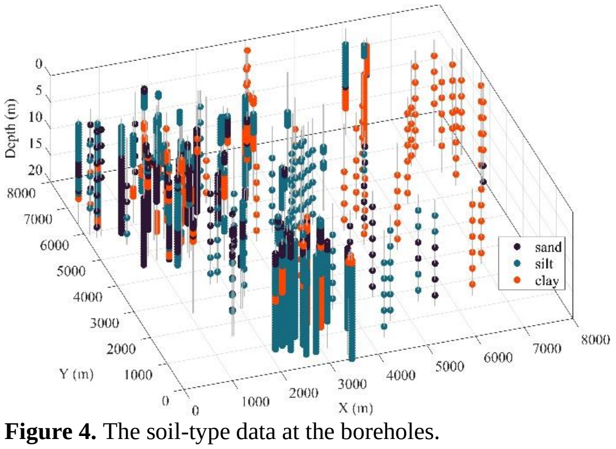


Figure 4. The soil-type data at the boreholes.

3. Soil database and HBM

For this particular case study, there is insufficient pairwise data to estimate the site-specific cross-correlation among the four soil parameters (LL, PI, FC, I_c). This is because (a) (LL, PI, FC) data are not available at boreholes (only soil-type data are available); (b) there are very limited nearby CPT-borehole pairs. As a result, the uncertainty in the cross-correlation is significant. To reduce this uncertainty, the HBM is adopted to learn the cross-correlation behaviors of generic sites in a soil database of (LL, PI, FC, I_c). Figure 7a and Figure 8a show the database of (LL, PI, FC, I_c) from 188 generic sites compiled by Kamyab Farahbakhsh and Ching (2023), where Figure 7a shows the database in the (LL, PI) space, and Figure 8a shows the database in the (FC, I_c) space. Data points from different sites are shown as different colors. The HBM can learn the cross-correlation behaviors of the 188 generic sites. To illustrate the HBM learning outcome, Figure 7b and Figure 8b show the cross-correlation behaviors of the "hypothetical sites" generated by the trained HBM. For instance, each (skewed) ellipse in Figure 7b represents the cross-correlation of (LL, PI) of a hypothetical site. These cross-correlation behaviors are transferred to the target site (Fucino Basin) through the trained HBM and serve as the "prior (cross-correlation) model" for the target site.

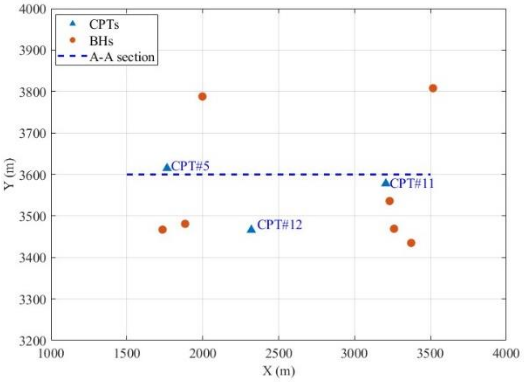


Figure 5. The zoom-in plot for the A-A section.

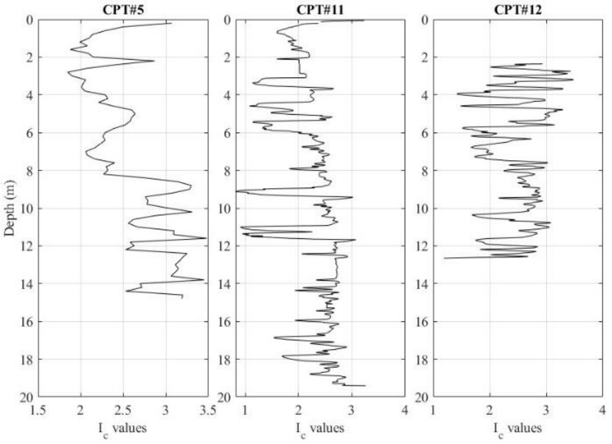


Figure 6. I_c profiles of the CPTs nearby the A-A section.

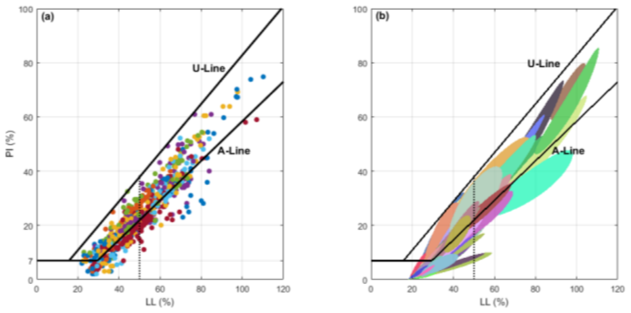


Figure 7. (a) (LL, PI) data; (b) cross-correlations of the hypothetical sites generated by the trained HBM.

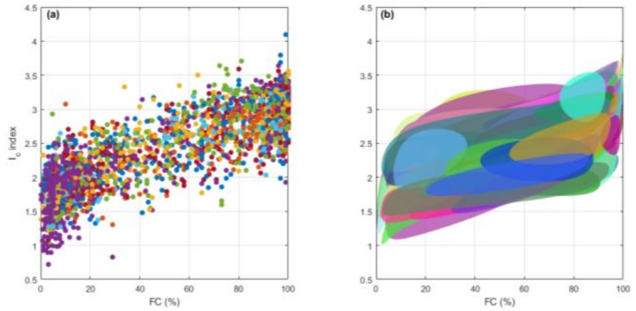


Figure 8. (a) (FC, I_c) data; (b) cross-correlations of the hypothetical sites generated by the trained HBM.

4. Analysis results

The simulation of soil types at unexplored locations of the Fucino Basin is demonstrated in this section. This is done by first updating the prior model into the posterior model by Bayesian analysis based on the HBM-MUSIC-3X method (Section 4.1) and then followed by simulating the conditional random fields (LL, PI, FC) at unexplored locations (Section 4.2). Then, the simulated (LL, PI, FC) are converted to the soil type (sand, silt, clay) based on the USCS criteria.

4.1. Bayesian analysis based on HBM-MUSIC-3X method

The Bayesian analysis of the HBM-MUSIC-3X method requires the knowledge of the following three items:

- (Item #1: prior cross-correlation model) There are insufficient pairwise site-specific data at the target site (Fucino Basin) to estimate these parameters. The cross-correlation parameters learned from the soil database are transferred to the target site using the HBM to serve as the prior cross-correlation model.
- (Item #2: auto-correlation model) The Whittle-Matérn (WM) auto-correlation model (Guttorp and Gneiting 2006; Liu et al. 2017; Ching and Phoon 2018) is adopted to model the spatial correlation. There are two kinds of auto-correlation parameters for the WM model: the scale of fluctuation (δ) and smoothness (ν). The vertical scale of fluctuation (δ_z) and vertical smoothness (ν_z) are identified from the CPT I_c profiles: $\delta_z \approx 0.32$ m and $\nu_z \approx 1.35$. However, it is not feasible to identify the horizontal scale of fluctuation (δ_h) and horizontal smoothness (ν_h) because the horizontal spacings among the CPTs are large. Instead, their values are assumed to be $\delta_h \approx 50$ m and $\nu_h \approx 1.35$ for the purpose of demonstration.
- (Item #3: site-specific data) The site-specific data include the soil-type data at the boreholes (i.e., Figure 4) and the I_c data at all CPTs. Note that for this particular case study, (LL, PI, FC) information is not available at the boreholes. Only the soil-type data (sand, silt, clay, etc.) are available.

In the essence of HBM-MUSIC-3X, the HBM trained by the soil database (item #1, e,g,, Figures 7 and 8) serves as the "prior cross-correlation model" of the Fucino Basin site. The likelihood function specifies the cross-correlation and spatial-correlation (item #2) in the site-specific data. The prior cross-correlation model is then updated by the site-specific data (item #3) into the "posterior cross-correlation model" of the Fucino Basin site through the Bayesian analysis. There is no analytical solution for this Bayesian analysis. The Gibbs sampler (GS) algorithm (Geman and Geman 1984; Gilk et al. 1996) is adopted to solve the Bayesian problem numerically by drawing samples from the posterior model. During the GS algorithm, a "truncation sampling" method is used to deal with the situation that only soil-type data are available at boreholes but (LL, PI, FC) are not: the (LL, PI, FC) samples are drawn from a truncated distribution, i.e., the probability density of (LL, PI, FC) inconsistent with the observed soil type is set to zero.

Figure 9 illustrates the behaviors of the posterior cross-correlation model. Figure 9a (posterior FC-I_c) can be compared with Figure 8b (prior FC-I_c), whereas Figure 9b (posterior LL-PI) can be compared with Figure 7b (prior LL-PI). As mentioned earlier, a clustered-HBM-MUSIC-3X method is adopted in this study, so the LL-PI & FC-Ic behaviors for the sand, silt, and clay clusters are separately shown in Figure 9.

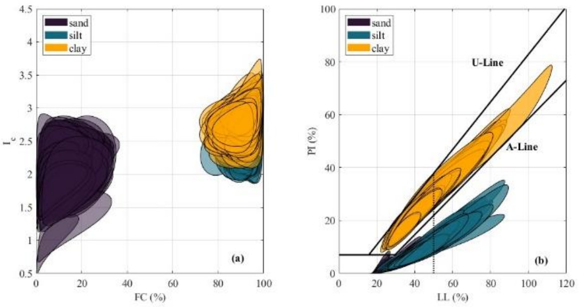


Figure 9. (a) posterior FC-I_c behaviors; (b) posterior LL-PI behaviors.

4.2. Conditional random field simulation results of soil types at unexplored locations

With the posterior cross-correlation, auto-correlation model, and the site-specific data, the conditional random fields of (LL, PI, FC, I_c) can be simulated at unexplored locations. The conditional random field simulation results over two unexplored locations are demonstrated in this section:

- (Case 1) The AA-section in Figure 3. Its Y coordinate is fixed at Y = 3600 m, whereas its X coordinate ranges from 1500 to 3500 m, and its z coordinate ranges from 0 to 20 m.
- (Case 2) The horizontal plane with z = 6 m. Its (X, Y) coordinate cover the full range: its X coordinate ranges from 0 to 8000 m, and its Y coordinate ranges from 0 to 8000 m.

Figure 10 shows one realization of the conditional random fields for Case 1. The (LL, PI, FC) results in Figures 10a, b, c can be converted into one realization of the USCS results shown in Figure 11a. One hundred realizations are simulated, and Figure 11b shows the most probable USCS result over the AA section. Note that CPT#11 is close to the AA section (see Figure 5; the distance from CPT#11 to the AA section is about 22 m). For comparison, Figure 12 shows the most probable soil-type profile and the sample medians and 95% confidence intervals of (LL, PI, FC, I_c) calculated based on the 100 realizations at the location closest to CPT#11. Consistency between the simulated (LL, PI, FC, I_c) and the I_c profile of CPT#11 is evident. Figure 13 shows the one realization of the conditional random field for Case 2. Again, one hundred realizations are simulated. One realization of the USCS results shown in Figure 14a, whereas Figure 14b shows the most probable USCS result.

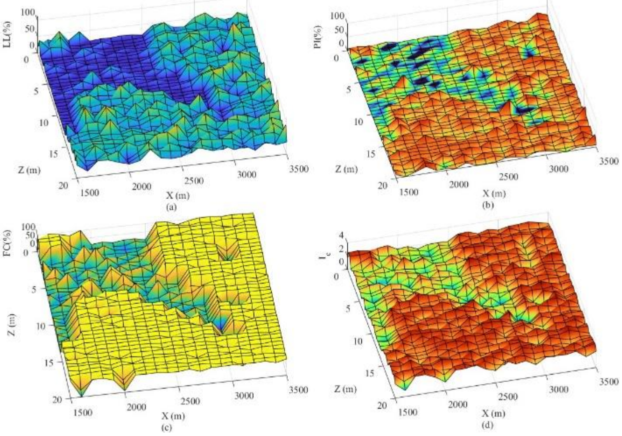


Figure 10. One realization of the conditional random fields for Case 1: (a) LL; (b) PI; (c) FC; (d) I_c.

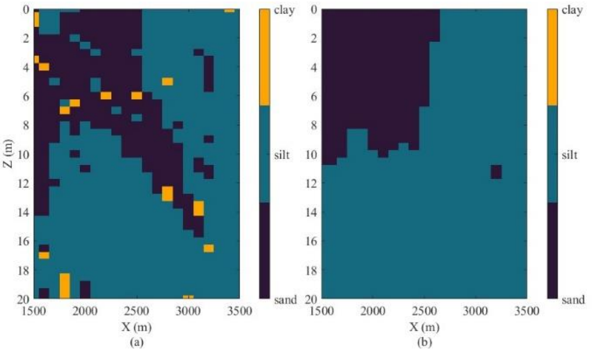


Figure 11. (a) One realization of USCS results for Case 1; (b) most probable USCS result.

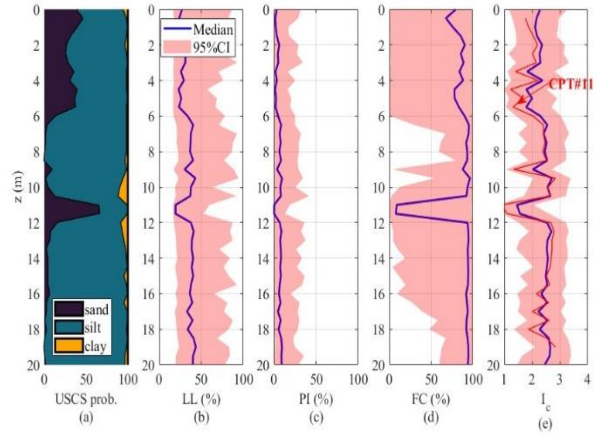


Figure 12. Conditional random field results (median & 95% confidence interval) at the location closest to CPT#11: (a) probability profile of USCS; (b) LL; (c) PI; (d) FC; (e) I_c.

5. Conclusion

In this paper, a real example of the Fucino Basin in Italy is adopted to demonstrate the application of a novel data-driven soil-delineation method. The method can handle multiple types of inputs, including soil-type data at boreholes and other soil parameters such as CPT results. The method can also take advantage of a soil database to reduce uncertainty in cross-correlation. The technical details for this novel method can be found elsewhere (Kamyab Farahbakhsh and Ching 2024). The purpose of the current paper is to demonstrate the analysis results of the real example.

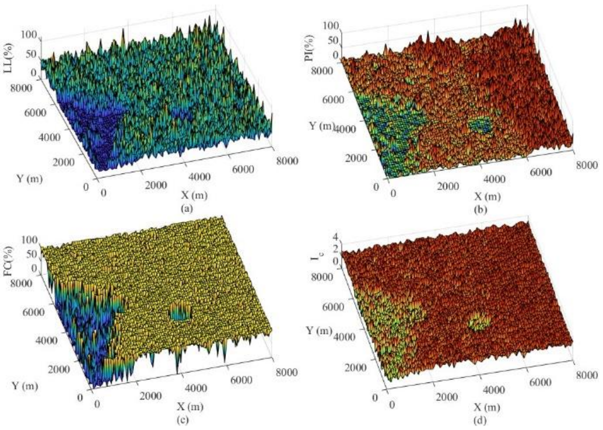


Figure 13. One realization of the conditional random fields for Case 2: (a) LL; (b) PI; (c) FC; (d) I_c.

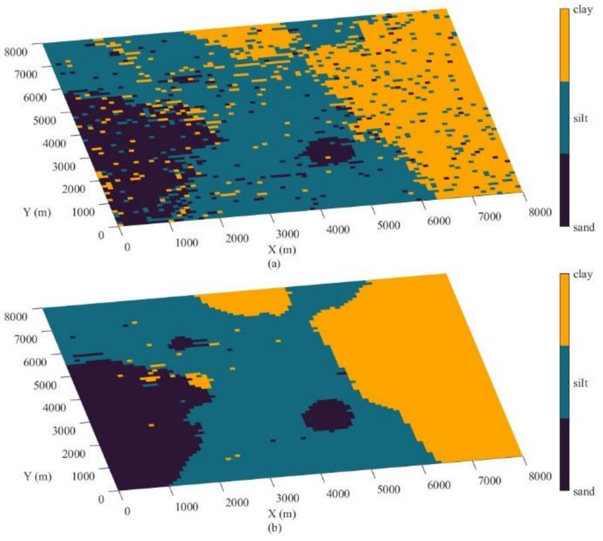


Figure 14. (a) One realization of USCS results for Case 2; (b) most probable USCS result.

Acknowledgements

The first two authors are grateful for the financial support provided by the National Council of Science and Technology of Taiwan (grant 112-2221-E-002-074-MY3).

References

Boncio, P., S. Amoroso, G. Vessia, M. Francescone, M. Nardone, P. Monaco, D. Famiani, D. Di Naccio, A. Mercuri, M.R. Manuel, F. Galadini, and G. Milana. 2018. "Evaluation of Liquefaction Potential in an Intermountain Quaternary

Lacustrine Basin (Fucino Basin, Central Italy): Implications for Seismic Microzonation Mapping." Bulletin of Earthquake Engineering 16(1): 91-111.

Caers, J. and T.F. Zhang. 2004. "Multiple-point Geostatistics: A Quantitative Vehicle for Integrating Geologic Analogs into Multiple Reservoir Models." In *Integration of Outcrop and Modern Analogs in Reservoir Modeling* (eds: Grammar, G.M., P.M. Harris, and G.P. Eberli), American Association of Petroleum Geologists, Memoirs, 383- 394.

Ching, J. and K.K. Phoon. 2018. "Impact of Autocorrelation Function Model on the Probability of Failure." Journal of Engineering Mechanics 145(1): 04018123.

Ching, J., K.K. Phoon, Z.Y. Yang, and A.W. Stuedlein. 2022. "Quasi-site-specific Multivariate Probability Distribution Model for Sparse, Incomplete, and Three-dimensional Spatially Varying Soil Data." Georisk 16(1): 53-76.

Ching, J., S. Wu, and K.K. Phoon. 2021. "Constructing Quasi-site-specific Multivariate Probability Distribution Using Hierarchical Bayesian Model." ASCE Journal of Engineering Mechanics 147(10): 04021069.

Elfeki, A. and M. Dekking. 2001. "A Markov Chain Model for Subsurface Characterization: Theory and Applications." Mathematical Geology 33(5): 569-89.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images." IEEE Transactions on Pattern Analysis and Machine Intelligence 6: 721-741.

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

Guttorp, P. and T. Gneiting. 2006. "Studies in the History of Probability and Statistics XLIX on the Matérn Correlation Family." Biometrika 93(4): 989-995.

Hu, L.Y. and T. Chugunova. 2008. "Multiple-point Geostatistics for Modeling Subsurface Heterogeneity: A Comprehensive Review." Water Resources Research 44: W11413.

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

Kamyab Farahbakhsh, H. and J. Ching. 2024. "Data-driven Soil-layer Delineation and Conditional Random Field Simulation - A Unified Approach." In preparation.

Li, J., M.J. Cassidy, J. Huang, L.M. Zhang, and R. Kelly. 2016b. "Probabilistic Identification of Soil Stratification." Géotechnique 66(1): 16-26.

Li, J., Y.M. Cai, X.Y. Li, and L.M. Zhang. 2019. "Simulating Realistic Geological Stratigraphy Using Direction-dependent Coupled Markov Chain Model." Computers and Geotechnics 115: 103147.

Li, Z., X.R. Wang, H. Wang, and R.Y. Liang. 2016a. "Quantifying Stratigraphic Uncertainties by Stochastic Simulation Techniques Based on Markov Random Field." Engineering Geology 201: 106-122.

Liu, W.F., Y.F. Leung, and M.K. Lo. 2017. "Integrated Framework for Characterization of Spatial Variability of Geological Profiles." Canadian Geotechnical Journal 54(1): 47-58.

Qi, X.H., D.Q. Li, K.K. Phoon, Z. Cao, and X.S. Tang. 2016. "Simulation of Geologic Uncertainty Using Coupled Markov Chain." Engineering Geology 207: 129-140.

Robertson, P.K. 2009. "Interpretation of Cone Penetration Tests – A Unified Approach." Canadian Geotechnical Journal 46(11): 1337-1355.

Shi, C. and Y. Wang. 2021a. "Non-parametric and Data-driven Interpolation of Subsurface Soil Stratigraphy from Limited Data Using Multiple Point Statistics." Canadian Geotechnical Journal 58(2): 261-280.

Shi, C. and Y. Wang. 2021b. "Development of Subsurface Geological Cross-section from Limited Site-specific Boreholes and Prior Geological Knowledge Using Iterative Convolution XGBoost." Journal of Geotechnical and Geoenvironmental Engineering 147(9): 04021082.

Varkey, D., A.P. van den Eijnden, and M.A. Hicks. 2023a. "Predicting Subsurface Stratigraphy using an Improved Coupled Markov Chain Method." Proceedings of the 14th International Conference on Applications of Statistics and Probability in Civil Engineering, Dublin, Ireland, July, 9-13.

Varkey, D., M.A. Hicks, and A.P. van den Eijnden. 2023b. "Predicting Subsurface Classification in 2D from Cone Penetration Test Data." Transportation Geotechnics 43: 101128.

Wang, Y., Y. Hu, and T. Zhao. 2020. "Cone Penetration Test (CPT)-based Subsurface Soil Classification and Zonation in Two-dimensional Vertical Cross Section Using Bayesian Compressive Sampling." Canadian Geotechnical Journal 57(7): 947-958.

Wei, X.X. and H. Wang. 2022. "Stochastic Stratigraphic Modeling Using Bayesian Machine Learning." Engineering Geology 307: 106789.

Zhao, C., W.P. Gong, T.Z. Li, C.H. Juang, H.M. Tang, and H. Wang. 2021. "Probabilistic Characterization of Subsurface Stratigraphic Configuration with Modified Random Field Approach." Engineering Geology 288: 106138.