Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5182-7_00-002-cd

Uncertainty Quantification in Data-Driven Geotechnical Stratigraphic Modeling

Hui Wang[1]

[1]Department of Civil and Environmental Engineering, University of Dayton, Dayton, OH 45469, USA.
E-mail: hwang12@udayton.edu

Abstract: Industry 4.0 will bring about a new Geotech digital era dominated by data science. In alignment with this trend, data-driven geotechnics is an emerging research field that contributes to this digital transformation. On the other hand, the Annex D in the latest edition of the international standard "General Principles on Reliability for Structures" (ISO2394:2015) recognizes that geotechnical reliability and risk analysis are sound only when the source uncertainty – the interpretation of observed data– is well quantified. Yet at present, it still heavily depends on engineers' subjective experience and may result in a less- or over-conservative design/decision. The challenges are two-fold: 1) how to better interpret geotechnical data that are multivariate, sparse, incomplete, and potentially corrupted in an algorithmic and smart manner at probed points, and 2) conditional on observed data, how to infer and model geotechnical information at vast unobserved locations accurately with quantified uncertainty. In this paper, the above two challenges are addressed to a certain extent using an in-house developed Bayesian approach for 2D soil stratigraphic interpretation. This new approach builds upon the author's previously developed one-dimensional hidden Markov random field (HMRF) model and 2D anisotropic Markov random field (MRF) simulation algorithm. Bayesian machine learning is implemented to jointly perform parameter estimation and stochastic simulation of soil stratigraphic profiles. The advantages of the developed approach are 1) inferring stratigraphic profile and associated uncertainty in an automatic manner and 2) both aleatoric and epistemic uncertainties are taken into consideration. This paper contributes to the techniques leveraging limited site investigation data for informed decision-making in geo-risk management.

Keywords: stratigraphic modeling; Bayesian machine learning; uncertainty quantification; Markov random field.

1 Introduction

The Annex D in the latest edition of the international standard "General Principles on Reliability for Structures" ISO2394 (ISO 2015) recognizes that geotechnical reliability-based design should place site investigation and the interpretation of site conditions/data as the cornerstone of the methodology (Phoon et al. 2016). Despite this notable advance in honoring the value of site investigation data in geotechnical design, it is reasonable to say that site investigation data still plays a supporting rather than a leading role in decision-making (Phoon 2020). A major reason is data scarcity at a specific site. Another reason is that soils mass is formed from complex geological processes (i.e., weathering, erosion, transportation, and deposition), and hence are spatially correlated and heterogeneous. The two aspects render significant uncertainties, which inevitably propagate into downstream design and construction as the driving forces of potential risks (Gong et al. 2019; Juang et al. 2018; Lu et al. 2005; Wang et al. 2016; Wang et al. 2017).

In the context of geotechnical risk management, geotechnical stratigraphic modeling using in-situ data with quantified uncertainty is still an open problem (Juang et al. 2018; Phoon 2020; Phoon et al. 2019). The challenge is to "squeeze" ambiguous information of soil heterogeneity from in-situ data having attributes like multivariate, uncertain, sparse, incomplete, and potentially corrupted/missing/incomplete with "X" denoting the spatial/temporal dimension, aka. MUSIC-X (Phoon et al. 2019). This paper proposes a potential means to address certain MUSIC-X challenges.

Among the customarily available in-situ testing techniques, cone penetration testing (CPT) is considered as an effective tool to assess soil stratigraphic configuration due to its convenience, repeatability, and economic efficiency. Based on the one-dimensional (1D) vertical sounding (e.g., in-situ cone resistance qc, sleeve friction fs, and pore water pressure u), soil type classification (Robertson 1990; Robertson 2009) and stratification (Ching et al. 2015; Wang et al. 2018; Wang et al. 2013) can be performed locally at a per-sounding basis. The interpretation quality and consistency are affected by the above mentioned MUSIC-X attributes as well as the local soil vertical correlation (Ching et al. 2015; Wang et al. 2018), and the uncertainty of the classification criteria (e.g., the Robertson soil behavior type (SBT)) (Robertson 2009; Wang et al. 2013). These issues have been addressed to some extent via the state-of-the-art data-driven techniques such as supervised classification (Ching et al. 2015) and Bayesian unsupervised clustering (Wang et al. 2018). Nevertheless, beyond the local independent 1D stratigraphic interpretation at sparse isolated points, a complete image of soil stratigraphic profile at a site scale still cannot be revealed in a rigorous data-driven manner with quantified uncertainty.

Recently, joint stratigraphic interpretation of boreholes and CPT soundings (Wang et al. 2019), or multiple CPT soundings (Wang et al. 2018) can be found in literature. Compared with previous interpretation at a per sounding basis, the extracted statistical pattern of the sounding data has been greatly enhanced by pooling

observations from multiple probed spots at a site and hence the interpretation consistency across different soundings is improved. However, the analysis is still contained in the 1D space along vertical depth. Wang et al. (2019) proposed a Bayesian compressive sampling-based method for identifying two-dimensional (2D) soil stratification from limited test soundings (Wang et al. 2019) and later improved it by using fuzzy soil classification boundaries and Monte Carlo simulation for probabilistic interpretation (Hu and Wang 2020). This work is aiming at contributing new knowledge along this track.

In this study, a novel Bayesian approach is proposed for 2D horizontal-vertical soil stratigraphic interpretation using multiple CPT soundings along a cross-section. This new approach builds upon the author's previously developed 1D vertical hidden Markov random field (HMRF) model (Wang et al. 2018). A hybrid 2D HMRF model is developed in order to accommodate missing/incomplete data and to characterize 2D spatial constraints for stratigraphic interpretation. Bayesian machine learning is adopted to jointly perform parameter estimation and stochastic simulation of soil profiles.

2 Hybrid HMRF framework

The two-layer model framework (i.e., the Markov random field (MRF)-Gaussian mixture model (GMM) proposed by Wang et al. (2018) is adopted as a basis. The novelty in the present work is accommodating missing/incomplete data and establishing 2D spatial constraint for interpreting soil types at unobserved locations given probed geotechnical site characterization data (CPT in this context) at probed spots along a cross-section. Some theoretical basics are presented below.

An MRF is an undirected graph and is defined on a regular lattice $S = \{(i,j)|\, i \in \{1,2,\dots,n_{column}\}, j \in \{1,2,\dots,n_{row}\}\}$, which is named as *physical space* for modeling the soil profile. Every node (i,j) in this lattice is a soil element and has its own *state type label* $x_{i,j} \in L$, where $L = \{1,2,\dots,K\}$ is a set of soil state code with no semantic meaning. A complete soil state configuration is denoted as $\boldsymbol{x} = \{x_{i,j}|\ (i,j) \in S\}$. The neighborhood system (Figure 1) of an element (i,j) is defined as the nearest eight elements $\partial_{i,j} = \{(i-1,j),(i+1,j),(i,j-1),(i,j+1),(i-1,j-1),(i+1,j+1),(i+1,j-1),(i-1,j+1)\}$. Any two neighboring elements have mutual influence in terms of *spatial contextual constraint* (in short, *spatial constraint*) measured by *potential* (Song et al. 2016; Zhang et al. 2001). The *granularity coefficients* $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4]$ are used to characterize the potential in the four directions respectively. Details regarding the *granularity coefficients* can be found in (Pereyra et al. 2013).

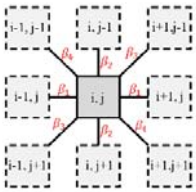

Figure 1. 2D neighborhood system of an MRF.

For extracting the statistical features from multiple CPT soundings, A GMM is defined in the $\log_{10}(F_r)$-$\log_{10}(Q_t)$ *feature space* (i.e., the space framed by the logarithm of normalized friction ratio and that of the tip resistance (Robertson 1990)). Probed soil elements S_{probed} along the depth at sounding locations having observed *features* $\boldsymbol{y} = \{\boldsymbol{y}_{i,j} = [\log_{10}(F_{r\_i,j}),\ \log_{10}(Q_{t\_i,j})]|(i,j) \in S_{probed}\}$ show up in feature space as scatter points. Similar probed soil elements are clustered together in feature space and assigned with a same soil state label $l \in L$, and hence they are assumed to share the same statistical characteristics (i.e., mean $\boldsymbol{\mu}_l \in \{\boldsymbol{\mu}_k|k \in L\}$ and covariance $\boldsymbol{\Sigma}_l \in \{\boldsymbol{\Sigma}_k|k \in L\}$) in feature space.

The soil state configuration in the physical space is assumed to be generated from a combination of an MRF model (i.e., only undirected graph defined at un-probed elements) and an HMRF model (i.e., undirected graph with derived conditional independent observations following a GMM at probed elements). Hence, inference of soil states can be categorized into two scenarios:

1) States $\boldsymbol{x}_{probed}$ assigned to probed soil elements S_{probed} are determined by both its neighborhood system in physical space and the GMM likelihood in feature space. These elements are referred to as HMRF elements;

2) Labels $\boldsymbol{x}_{unprobed}$ assigned to un-probed soil elements $S_{unprobed} = S - S_{probed}$ are determined solely by the label configuration of its neighbors in physical space, hence these elements are MRF elements.

Under the first scenario $(i,j) \in S_{probed}$, $\boldsymbol{x}_{probed}$ and $\boldsymbol{y}$ form a bijection. Given a set of probed soil features $\boldsymbol{y}$ at sounding locations S_{probed}, $P(\boldsymbol{x}_{probed}|\boldsymbol{y})$ is a Gibbs distribution (Wang et al. 2016). The local conditional probability can be expressed as below:

$$P\left(x_{i,j}\middle|\boldsymbol{x}_{\partial_{i,j}},\boldsymbol{y}_{i,j};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}\right)=\frac{\exp\left[-U_{i.j}\left(x_{i,j},\boldsymbol{x}_{\partial_{i,j}}|\boldsymbol{y}_{i,j};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}\right)\right]}{\sum_{k\in L}\exp\left[-U_{i.j}\left(k,\boldsymbol{x}_{\partial_{i,j}}|\boldsymbol{y}_{i,j};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}\right)\right]} \tag{1}$$

where the local energy function $U_{i.j}(.)$ can be calculated using the equation:

$$U_{i.j}\left(x_{i,j},\boldsymbol{x}_{\partial_{i,j}}|\boldsymbol{y}_{i,j};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}\right)=U_{i,j}^{MRF}\left(x_{i,j},\boldsymbol{x}_{\partial_{i,j}};\boldsymbol{\beta}\right)+U_{i,j}^{LH}\left(\boldsymbol{y}_{i,j};\boldsymbol{\mu}_{x_{i,j}}\boldsymbol{\Sigma}_{x_{i,j}}\right) \tag{2}$$

The MRF energy $U_{i,j}^{MRF}$ has the expression

$$U_{i,j}^{MRF}\left(x_{i,j},\boldsymbol{x}_{\partial_{i,j}};\boldsymbol{\beta}\right)=\sum_{(i',j')\in\partial_{i,j}}V(x_{i,j},x_{i',j'}) \tag{3}$$

where the potential function $V(x_{i,j},x_{i',j'})$ is defined as below:

$$V\left(x_{i,j},x_{i',j'}\right)=\begin{cases}0, & if\ x_{i,j}=x_{i',j'}\\ \beta_1, & if\ x_{i,j}\neq x_{i',j'}\ and\ (i',j')\in\{(i,j-1),(i,j+1)\}\\ \beta_2, & if\ x_{i,j}\neq x_{i',j'}\ and\ (i',j')\in\{(i-1,j),(i+1,j)\}\\ \beta_3, & if\ x_{i,j}\neq x_{i',j'}\ and\ (i',j')\in\{(i-1,j+1),(i+1,j-1)\}\\ \beta_4, & if\ x_{i,j}\neq x_{i',j'}\ and\ (i',j')\in\{(i-1,j-1),(i+1,j+1)\}\end{cases} \tag{4}$$

where $\beta_d, d\in\{1,2,3,4\}$ can be either positive or negative. A positive β_d encourages neighboring elements in this direction to have the same label (i.e., attraction effect), while a negative β_d encourages neighboring elements in this direction to have different labels (i.e., repulsion effect). Therefore, $\boldsymbol{\beta}=[\beta_1,\beta_2,\beta_3,\beta_4]$ characterize the anisotropy of the label field. The likelihood energy $U_{i,j}^{LH}$ has the expression

$$U_{i,j}^{LH}\left(\boldsymbol{y}_{i,j};\mu_{x_{i,j}}\boldsymbol{\Sigma}_{x_{i,j}}\right)=\frac{1}{2}\left(\boldsymbol{y}_{i,j}-\boldsymbol{\mu}_{x_{i,j}}\right)^T\boldsymbol{\Sigma}_{x_{i,j}}{}^{-1}\left(\boldsymbol{y}_{i,j}-\boldsymbol{\mu}_{x_{i,j}}\right)+\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{x_{i,j}}\right| \tag{5}$$

Note that the likelihood energy consists of 1) the Mahalanobis distance (De Maesschalck et al. 2000) between the observed feature $\boldsymbol{y}_{i,j}$ and the GMM component density function $f_{x_{i,j}}$ corresponding to label $x_{i,j}$, and 2) the volume measure (i.e., determinant of the covariance matrix) of $f_{x_{i,j}}$. An intuitive explanation of the likelihood energy is that a state may be preferred if its corresponding GMM component density function has a narrower shape and is closer (in terms of Mahalanobis distance) to the observed feature point $\boldsymbol{y}_{i,j}$ as both means can result in a higher probability density of $\boldsymbol{y}_{i,j}$.

Under the second scenario $(i,j)\in S_{unprobed}$, $P(\boldsymbol{x}_{S_{unprobed}})$ is a regular Gibbs distribution. The local conditional probability formed only by the MRF energy.

$$P\left(x_{i,j}\middle|\boldsymbol{x}_{\partial_{i,j}};\boldsymbol{\beta}\right)=\frac{\exp\left[-U_{i,j}^{MRF}\left(x_{i,j},\boldsymbol{x}_{\partial_{i,j}};\boldsymbol{\beta}\right)\right]}{\sum_{x'_{i,j}\in L}\exp\left[-U_{i,j}^{MRF}\left(x'_{i,j},\boldsymbol{x}_{\partial_{i,j}};\boldsymbol{\beta}\right)\right]} \tag{6}$$

Eq. (1) and Eq. (6) indicate that a preferred local soil state should minimize the local energy of the random field system. For an HMRF element, a preferred label should minimize the summation of the MRF energy and the likelihood energy. Lowering the MRF energy will result in an inclination of having the same label as its neighbors in case of attraction effect ($\beta_d>0$), or of having a different label to its neighbors in case of repulsion effect ($\beta_d<0$). The strength of the effect is controlled by the corresponding granularity coefficients. Lowering the likelihood energy will prefer a soil state representing the cluster with the highest probability density of the observed features. In other words, the soil state of an HMRF element is jointly controlled by both MRF and GMM models, while for an MRF element, a preferred soil state will only minimized the MRF energy and hence only honor the stratigraphic anisotropy controlled by $\boldsymbol{\beta}$. Numerically, given a set of model parameters $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ and an initial guess $\boldsymbol{x}_0$ of the label field, realizations of the hybrid Gibbs field consisting of $P\left(\boldsymbol{x}_{probed}\middle|\boldsymbol{y}\right)$ and $P(\boldsymbol{x}_{unprobed})$ can be sampled iteratively via a Gibbs sampler (Geman and Geman 1984) using Eq. (1) and Eq. (6) for two types of elements respectively.

On the other hand, based on the principle of mean field-like approximation (Celeux et al. 2003; Forbes and Peyrard 2003), approximating the random field $P(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ can be achieved by assuming that the soil state $\boldsymbol{x}_{i,j}$ of a probed element $(i,j)\in S_{probed}$ does not depend on the label settings of its neighbors in the current field during the sampling process, but is generated independently by borrowing the neighboring label settings in a mean field-like approximation $\tilde{\boldsymbol{x}}$ of the original unknown label configuration (Celeux et al. 2003). The approximated probability is denoted as $P_{\tilde{\boldsymbol{x}}}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ and the mathematical expressions are shown below.

$$P(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})\approx P_{\tilde{\boldsymbol{x}}}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})=\prod_{(i,j)\in S_{probed}}P\left(\boldsymbol{y}_{i,j}|\tilde{\boldsymbol{x}}_{\partial_{i,j}},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \tag{7}$$

$$P\left(\boldsymbol{y}_{i,j}|\tilde{\boldsymbol{x}}_{\partial_{i,j}},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\sum_{k\in L}P\left(k|\tilde{\boldsymbol{x}}_{\partial_{i,j}};\boldsymbol{\beta}\right)f_k\left(\boldsymbol{y}_{i,j};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right) \tag{8}$$

where the mean field-like approximation $\tilde{\boldsymbol{x}}$ of the original unknown label configuration is the key for this calculation. The mixture coefficients $P\left(k\left|\tilde{\boldsymbol{x}}_{\partial_{i,j}};\boldsymbol{\beta}\right.\right)$ of the GMM at the probed elements $(i,j)\in S_{probed}$ are derived from the local MRF contextual information of $\tilde{\boldsymbol{x}}$.

$$P\left(k\left|\tilde{\boldsymbol{x}}_{\partial_{i,j}};\boldsymbol{\beta}\right.\right)=\frac{\exp\left[-U_{i,j}^{MRF}\left(k,\tilde{\boldsymbol{x}}_{\partial_{i,j}};\boldsymbol{\beta}\right)\right]}{\sum_{k'\in L}\exp\left[-U_{i,j}^{MRF}\left(k',\tilde{\boldsymbol{x}}_{\partial_{i,j}};\boldsymbol{\beta}\right)\right]} \tag{9}$$

Finding a good $\tilde{\boldsymbol{x}}$ is not a trivial task. Fortunately, a simple and effective *simulated field algorithm* has been developed and rigorously validated in the seminal work (Celeux et al. 2003). To be more concrete, $\tilde{\boldsymbol{x}}$ is a realization of the hybrid Gibbs field consisting of $P\left(\boldsymbol{x}_{probed}\left|\boldsymbol{y}\right.\right)$ and $P(\boldsymbol{x}_{unprobed})$, and in the current context, generating $\tilde{\boldsymbol{x}}$ can be achieved by running a Gibbs sampler iteratively using Eq. (1) and Eq. (6) given a set of estimated model parameters and an inferred label field, both of which can be updated iteratively. Based on Eqs. (7-9), given a simulated realization $\tilde{\boldsymbol{x}}$ and prior distributions of $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$, posterior distributions of $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ can be derived in a Bayesian manner.

Eqs. (1-9) together form the novel hybrid HMRF model. As can be noticed, generating realizations $\tilde{\boldsymbol{x}}$ of the hybrid Gibbs field consisting of $P\left(\boldsymbol{x}_{probed}\left|\boldsymbol{y}\right.\right)$ and $P(\boldsymbol{x}_{unprobed})$ and estimating the model parameters $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ are coupled with each other. An iterative Bayesian machine learning algorithm is developed for inferring both two parts given $\boldsymbol{y}$ only at S_{probed}.

3 Bayesian unsupervised learning

It is assumed that both $\boldsymbol{x}=\left\{\boldsymbol{x}_{probed},\boldsymbol{x}_{unprobed}\right\}$ and $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ are unknown *in priori*. Based on the established learning algorithm in Wang et al. (2018), a modified approach is employed to simulate the inferred soil state field realizations $\tilde{\boldsymbol{x}}$ and posterior samples of $\{\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ iteratively. The simulation steps are shown below.

Step 0: Estimating the initial $\tilde{\boldsymbol{x}}^0$, $\boldsymbol{\beta}^0$, $\boldsymbol{\mu}^0$, $\boldsymbol{\Sigma}^0$

a) Identify the number K of soil states using Bayesian information criterion (BIC) (Schwarz 1978) by fitting $\boldsymbol{y}$ to a family of GMMs with different number of clusters using EM algorithm (McLachlan and Peel 2004) and calculate corresponding BIC values. The optimal model corresponds to the lowest BIC value;

b) Initialize $\tilde{\boldsymbol{x}}_{probed}^0$, $\boldsymbol{\mu}^0$, $\boldsymbol{\Sigma}^0$with the output from the optimal GMM model. Initialize $\boldsymbol{\beta}^0=\boldsymbol{\mu}_{\boldsymbol{\beta}}$ (i.e., prior mean of $\boldsymbol{\beta}$, more details for setting $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ is shown in Step 2).

c) Initialize $\tilde{\boldsymbol{x}}_{unprobed}^0$ in a row-wise manner on the lattice S. For each unprobed element, a soil state is randomly generated by following a uniformly distributed probability mass function (PMF) corresponding to a set of available labels appearing only in the current row. The uniform PMF reflects the ignorance of the label preference at unprobed elements as no feature observation is available. The set of available states only in the current row honors the information from the preliminary inference of $\tilde{\boldsymbol{x}}_{probed}^0$ along the horizontal direction as soil stratigraphic profiles are usually horizontally dominated.

Step 1: Generate $\tilde{\boldsymbol{x}}$

Recall that $P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a hybrid Gibbs field consisting of $P\left(\boldsymbol{x}_{probed}\left|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma}\right.\right)$ and $P(\boldsymbol{x}_{unprobed}|\boldsymbol{\beta})$. After a closer comparison between Eqs. (1-2) and Eq. (6), we can notice that the only difference between the energy function of an HMRF element and that of an MRF element is whether having the likelihood energy portion. To unify the mathematical energy expression form of the two types of elements, we assign "0" as the corresponding "virtual" likelihood energy of each possible label to all MRF elements as a placeholder since, in concept, an MRF element does not have likelihood energy. Then, $P(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gibbs field with a unified energy function expression. This setting simplifies the algorithm implementation because there is no need to perform element distinction during the sampling step. Detailed implementations are shown below:

Given $\tilde{\boldsymbol{x}}^{t-1}$, $\boldsymbol{\beta}^{t-1}$, $\boldsymbol{\mu}^{t-1}$, $\boldsymbol{\Sigma}^{t-1}$ from the t-1th iteration,

a) Calculate $U_{i,j}^{MRF}\left(k,\tilde{\boldsymbol{x}}_{\partial_{i,j}}^{t-1};\boldsymbol{\beta}^{t-1}\right)$ corresponding to each and every label $k\in L$ for all elements $(i,j)\in S$ using Eqs. (3-4);

b) Initialize $U_{i,j}^{LH}\left(\boldsymbol{y}_{i,j};\boldsymbol{\mu}_k^{t-1},\boldsymbol{\Sigma}_k^{t-1}\right)=0$ corresponding to each and every label $k\in L$ for all elements $(i,j)\in$ S. Note that $\boldsymbol{y}_{i,j}$ is not available for MRF elements $(i,j)\in S_{unprobed}$, however, as described above, "0" is assigned as a placeholder. Calculate $U_{i,j}^{LH}\left(\boldsymbol{y}_{i,j};\boldsymbol{\mu}_k^{t-1},\boldsymbol{\Sigma}_k^{t-1}\right)$ corresponding to each label $k\in L$ using Eq. (5) and overwrite the initial "0"s for elements $(i,j)\in S_{probed}$;

c) Calculate the local conditional probability $P\left(k\middle|\tilde{\boldsymbol{x}}_{\partial_{i,j}}^{t-1}, \boldsymbol{y}_{i,j}; \boldsymbol{\beta}^{t-1}, \boldsymbol{\mu}_k^{t-1}, \boldsymbol{\Sigma}_k^{t-1}\right)$ corresponding to each label $k \in L$ using Eq. (1-2) for all elements $(i,j) \in \mathrm{S}$. Note that $\boldsymbol{y}_{i,j}$ is not available for MRF elements $(i,j) \in S_{unprobed}$, the placeholder "0" is used as the likelihood energy in calculation;

d) Simulate $\tilde{\boldsymbol{x}}^t$ according to $P\left(k\middle|\tilde{\boldsymbol{x}}_{\partial_{i,j}}^{t-1}, \boldsymbol{y}_{i,j}; \boldsymbol{\beta}^{t-1}, \boldsymbol{\mu}_k^{t-1}, \boldsymbol{\Sigma}_k^{t-1}\right)$ via the Gibbs sampler proposed by (Geman and Geman 1984) without any distinction between MRF and HMRF elements. Alternatively, based on steps (a-c), a parallel strategy named chromatic sampler can be used to boost up the computational efficiency. More details in this regard can be found in (Wang et al. 2016).

Step 2: Generate posterior samples of $\{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

Based on the Bayesian theorem, the posterior distribution of the three model parameters are formulated as

$$post(\boldsymbol{\beta}|\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto piror(\boldsymbol{\beta}) Likelihood(\boldsymbol{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{10}$$

$$post(\boldsymbol{\mu}|\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto piror(\boldsymbol{\mu}) Likelihood(\boldsymbol{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{11}$$

$$post(\boldsymbol{\Sigma}|\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\mu}) \propto piror(\boldsymbol{\Sigma}) Likelihood(\boldsymbol{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

where $\tilde{\boldsymbol{x}}$ is the simulated realization from Step 1. The likelihood function $Likelihood(\boldsymbol{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be approximated by $P_{\tilde{\boldsymbol{x}}}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and calculated using Eqs. (7-9).

Prior knowledge of $\boldsymbol{\beta}$ is required for imposing regularization due to the ill-posedness and directional unbalancedness of the 2D stratigraphic interpretation problem. More specific, a four-dimensional multivariate normal distribution with a suitable center $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is used as the $piror(\boldsymbol{\beta})$. In this study, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ are determined in a heuristic or empirical manner by assuming the latent field is horizontally dominated, which is consistent with the initial soil state simulation strategy mentioned above. To be more specific, regarding $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, β_1 is set to be substantially greater than other three granularity coefficients. A rule of thumb is that β_1 could be in the interval between 3 and 10, while $\beta_{d\in\{2,3,4\}}$ are positive and less than 0.5. $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is set to be a diagonal matrix with variances less than 1 for each $\beta_{d\in\{1,2,3,4\}}$. Preliminary testing results show that the developed algorithm is not sensitive to $piror(\boldsymbol{\beta})$ with different parameter values if the above rule of thumb is followed. However, using non-informative or less-informative $piror(\boldsymbol{\beta})$ with no regularization may result in realizations that are mathematically MRFs while not compliant with prior geological knowledge or engineering judgment. Hence, some tuning work is needed to honor engineering experiences in different contexts. How to rigorously choose them (or automatically choose them via an algorithm) is beyond the research scope of the present work, yet it is currently under another thorough investigation.

Less informative priors of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are generally sufficient, while specifically defined priors are recommended if available. We use the same prior settings as developed in (Wang et al. 2018). In brief, for all $k \in L$, flat multivariate Gaussian priors are used for $\boldsymbol{\mu}_k$ with arbitrarily chosen center $\boldsymbol{\eta}_k$ and large covariance $\boldsymbol{\Xi}_k$; the separation strategy (Barnard et al. 2000) is used to construct the prior distribution of $\boldsymbol{\Sigma}_k$ with three parameters $\{\nu, \boldsymbol{b}_k, \xi\}$. More details are referred to (Wang et al. 2018).

The most widely used Metropolis—Hastings (M-H) algorithm is employed in this step. To be more specific: Given $\boldsymbol{\beta}^{t-1}$, $\boldsymbol{\mu}^{t-1}, \boldsymbol{\Sigma}^{t-1}$ from the previous iteration, and $\tilde{\boldsymbol{x}}^t$ simulated in Step 1 of the current iteration, $\boldsymbol{\beta}^t$, $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$ are sampled from the posterior distributions by implementing Algorithm 2 developed in (Wang et al. 2018) while plugging in the priors introduced above and using Eq. (10-12) as target functions.

4 Uncertainty quantification

A significant difference between the present investigation and the author's previous work (Wang et al. 2018) is that the initial guess of $\tilde{\boldsymbol{x}}_{unprobed}$ cannot be inferred in Step 0 since no soil feature is available at those locations. As a result, the random sampling strategy described in Step 0 (c) applies very weak spatial constraint on drawing $\tilde{\boldsymbol{x}}^0$, which is the starting point of the Markov chain. Note that $\tilde{\boldsymbol{x}}$ is a realization from a hybrid Gibbs field defined in the soil state configuration space $\boldsymbol{L^S}$. The probability distribution of the hybrid Gibbs field could be extremely complicated with numerous local optimums that can trap the MCMC sampling process. Once a chain process $\{\tilde{\boldsymbol{x}}^t\}$ is trapped around a specific local optimum, the realization can only reflect limited uncertainty around this local optimal label configuration (i.e., a certain soil stratigraphic setting). Therefore, without a reasonable guess of $\tilde{\boldsymbol{x}}^0$, simulating a single Markov chain may result in a significantly underestimated uncertainty of the soil label configuration. Alternatively, generating multiple parallel chains with different random starting configurations $\tilde{\boldsymbol{x}}^0$ seems to be a more robust approach.

During the sampling process, the uncertainty of falling into any local optimum is referred to as "*global uncertainty*", while the randomness around a specific local optimum is referred to as "*local uncertainty*". Note that the "global" or "local" concept is defined in the context of soil state configuration space $\boldsymbol{L^S}$ (s-dimensional) and hence should not be interpreted from a two-dimensional spatial perspective. The uncertainty quantification is conducted at both two levels. To be more specific, a total number of N_{mc} chains of $\{\boldsymbol{\beta}^t, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \tilde{\boldsymbol{x}}^t\}_n, n \in$

$\{1,2,\dots,N_{mc}\}, t \in \{1,2,\dots,T_{iter}\}$ are simulated with T_{iter} iterations for each chain. Both global and local uncertainty of the soil labels are characterized using information entropy (Wellmann and Regenauer-Lieb 2012) at a per-pixel basis following the expression

$$H(i,j) = -\sum_{k \in L} P_k(i,j) \log P_k(i,j) \tag{13}$$

where $H(i,j)$ is the information entropy at pixel (i,j), and $P_k(i,j)$ is the membership probability of soil state k (i.e., the probability of assigning soil state k to pixel (i,j)). High $H(i,j)$ indicates a high uncertainty level at pixel (i,j).

For local uncertainty quantification, the membership probability of each soil state can be calculated by using their frequencies in a single chain divided by the chain length T_{iter}. The calculated information entropy at each pixel is referred to as *local information entropy*. The means and standard deviations of $\{\boldsymbol{\beta}^t, \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\}$ corresponding to each chain are used as parameter estimators and the measure of parameters' local uncertainty respectively.

For global uncertainty quantification, the following two steps are designed. 1) At a per-chain basis, the *maximum a posteriori* (MAP) estimate of the soil state at each element is achieved by taking the one with the highest membership probability and the result of each chain is referred to as *local MAP soil state field*. 2) The local MAP state fields of all chains form an ensemble and the membership probability of each soil state at all elements are then calculated using the frequencies of each soil state divided by the number of chains N_{mc}. The calculated information entropy at each pixel is referred to as *global information entropy*. Accordingly, the MAP label field is referred to as *global MAP label field*. The mean and standard deviation of the model parameter estimators (e.g., mean) of different chains are reported as the global estimators and the measure of parameters' global uncertainty respectively.

From a conceptual point of view, the local information entropy is mainly controlled by the posterior of $\boldsymbol{\beta}$, which is jointly affected by the prior of $\boldsymbol{\beta}$ and the complexity/heterogeneity of the underlying soil label configuration. A prior of $\boldsymbol{\beta}$ with weaker regularization and/or a complex/heterogeneous underlying soil state configuration may result in a high information entropy level around boundaries between different soil states in a single Markov chain. On the other hand, the global information entropy is mainly dominated by the amount of known information (aka. anchors) from the probed locations. In the current context, it is straightforward to understand that the more CPT soundings we have the less global information entropy we can expect.

The global uncertainty level of model parameters is generally lower than the local uncertainty level per the definition given above. This is due to the fact that the parameter estimation is mainly based on the probed soil features. As the amount of known information has no change from one chain to another, the local uncertainty levels and the derived parameter estimators are also similar across different chains. The variation of the estimators (e.g., the mean) across multiple chains is lower than that of the model parameter samples along a single chain. Better intuitive understanding can be gained from several synthetic and real-world examples demonstrated in the following sections.

5 Validation using synthetic examples

A validation using synthetic data is demonstrated in this section. The "observed" geotechnical soil properties (i.e., $\log_{10}(F_r)$ and $\log_{10}(Q_t)$) are simulated in two different ways: 1) conditional independently sampled at a per-pixel basis; 2) conditional jointly simulated using matrix decomposition method equipped with the transverse anisotropy exponential correlation function (Zhu and Zhang 2013) in order to reflect spatial correlation. It is assumed that there are three different soil states and their corresponding observed properties follow multivariate Gaussian distributions if generated using the first method, or follow an anisotropic Gaussian random field if generated using the second method. The probability distribution and random field parameters used are shown in Table 1.

Table 1. Distribution and random field parameters for simulating synthetic observed features

Soil state	Multivariate Gaussian distribution parameters		Random field parameters	
	μ	Σ	θ_h (m)	θ_v (m)
1	[-0.0969, 2.0308]	$\begin{bmatrix} 0.01 & 0.0075 \\ 0.0075 & 0.0625 \end{bmatrix}$	20	2
2	[-0.301, 1.3046]	$\begin{bmatrix} 0.0049 & 0.0014 \\ 0.0014 & 0.01 \end{bmatrix}$	20	4
3	[-0.0458, 1.4862]	$\begin{bmatrix} 0.01 & 0.005 \\ 0.005 & 0.01 \end{bmatrix}$	20	2

Note: 1) $\boldsymbol{\mu}$: mean; $\boldsymbol{\Sigma}$: covariance matrix; θ_h: scale of fluctuation of the transverse anisotropy exponential correlation function along the horizontal direction; θ_v: scale of fluctuation of the transverse anisotropy exponential correlation function along the vertical direction. 2) For independent sampling method, only Multivariate Gaussian distribution parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ are used; for matrix decomposition method, all four parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta_h, \theta_v\}$ are used.

Example 1 is designed to demonstrate that the approach performs well if both the latent field and the associated virtual soil property field follow the basic model assumptions (i.e., MRF for the latent field and conditional independency for the soil property field). The 100×100 latent field is simulated using the classical Gibbs sampler (Geman and Geman 1984) with $\boldsymbol{\beta}_{\text{true}} = [4.5, 0.15, 0.15, 0.15]$. The two soil property fields (i.e., $\log_{10}(F_r)$ and $\log_{10}(Q_t)$) are then simulated conditionally on the soil states using parameters as described in Table 1. 15% of the columns are randomly selected as virtual "known CPT soundings". Note that the known soundings do not need to be equally spaced as there is no such assumption or requirement during the model development. The synthetic data is shown in Figure 2. The original full image corresponding to $\log_{10}(F_r)$ and $\log_{10}(Q_t)$ are shown in Figure 2(a-b). The complete latent field is shown in Figure 2(c). After removing the "unobserved" portion, the "probed" features are shown in Figure 2(d-e) and their corresponding projected scatter points in the feature space (i.e., the Robertson chart (Robertson 1990)) are shown in Figure 2(f).

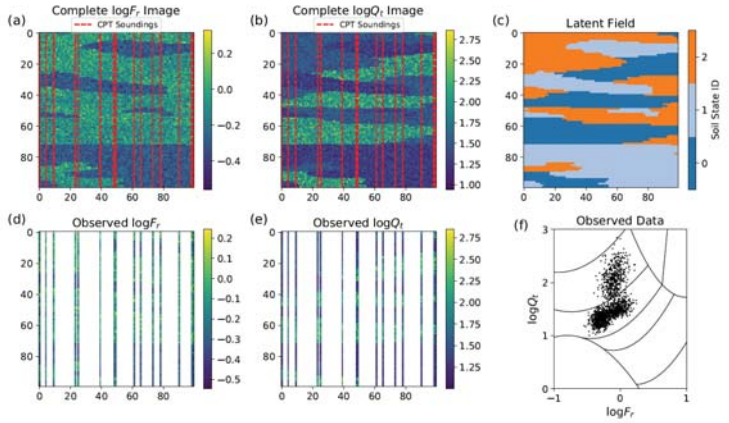

Figure 2. Simulated synthetic data of Example 1.

After performing model selection using BIC, the underlying three soil states can be successfully identified (Figure 3(a)). A single MCMC simulation with 1000 iterations is then performed following the steps described in Section 3. The local MAP latent field (Figure 3(b)) inferred from only 15% information achieves an overall accuracy of 97.01%. Only layer boundary pixels that are far away from known columns have difficulties to be correctly inferred and result in higher information entropy as illustrated in Figure 3(c-d). The estimated statistical pattern in the feature space clearly shows a three-cluster pattern (Figure 3(e)). A more detailed quantitative comparison between the true model parameters and the estimated parameters is listed in Table 2. It is shown that all parameters can be inferred with reasonable accuracy and precision.

Table 2: Estimated parameters of Example 1 using samples from a single Markov chain

Soil State	Parameter	True value	Mean	STD
1	μ_1	[-0.0969, 2.0308]	[-0.0953, 2.0554]	[0.0039, 0.0108]
	Σ_1	$\begin{bmatrix} 0.01 & 0.0075 \\ 0.0075 & 0.0625 \end{bmatrix}$	$\begin{bmatrix} 0.0101 & 0.0087 \\ 0.0087 & 0.0648 \end{bmatrix}$	$\begin{bmatrix} 0.0007 & 0.0013 \\ 0.0013 & 0.0039 \end{bmatrix}$
2	μ_2	[-0.301, 1.3046]	[-0.3035, 1.3016]	[0.0029, 0.0074]
	Σ_2	$\begin{bmatrix} 0.0049 & 0.0014 \\ 0.0014 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 0.0052 & 0.0014 \\ 0.0014 & 0.0105 \end{bmatrix}$	$\begin{bmatrix} 0.0003 & 0.0002 \\ 0.0002 & 0.0006 \end{bmatrix}$
3	μ_3	[-0.0458, 1.4862]	[-0.0497, 1.4861]	[0.0077, 0.0062]
	Σ_3	$\begin{bmatrix} 0.01 & 0.005 \\ 0.005 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 0.0104 & 0.0048 \\ 0.0048 & 0.0096 \end{bmatrix}$	$\begin{bmatrix} 0.0007 & 0.0006 \\ 0.0006 & 0.0006 \end{bmatrix}$
	β	[4.5,0.15,0.15,0.15]	[4.68,0.21,0.09,0.14]	[0.68,0.32,0.35,0.28]

Then, $N_{mc} = 100$ parallel Markov chains are simulated for a robust uncertainty quantification and the analyzing results are provided in Figure 4. The global MAP label field achieves an overall accuracy of 97.22% (Figure 4(a)). For comparison purpose, the quality of the best and worst cases are also reported in Figure 4(b-c). In addition, a per-pixel evaluation of accuracy and global uncertainty is shown in Figure 4(d-e). The empirical histogram of the overall accuracy from 100 chains is shown in Figure 4(f) indicating small variation of the performance across different runs. The results show that the proposed approach is fairly robust under the current problem setting.

Example 2 is designed for illustrating that even the underlying conditional independent assumption cannot fully hold, the proposed approach still can provide satisfactory results. To demonstrate so, both horizontal and

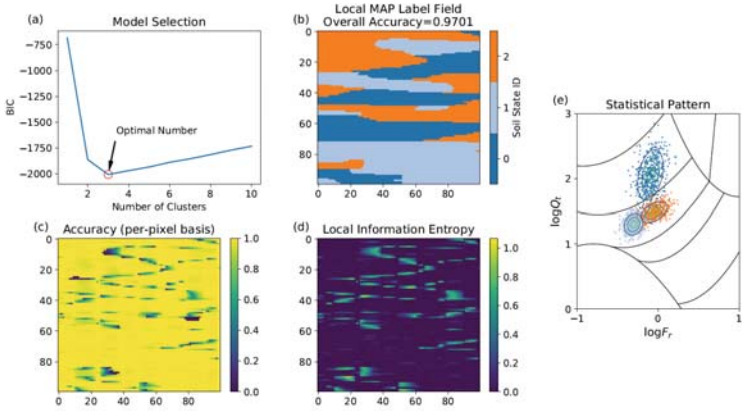

Figure 3. Estimation results from a single simulation of Example 1.

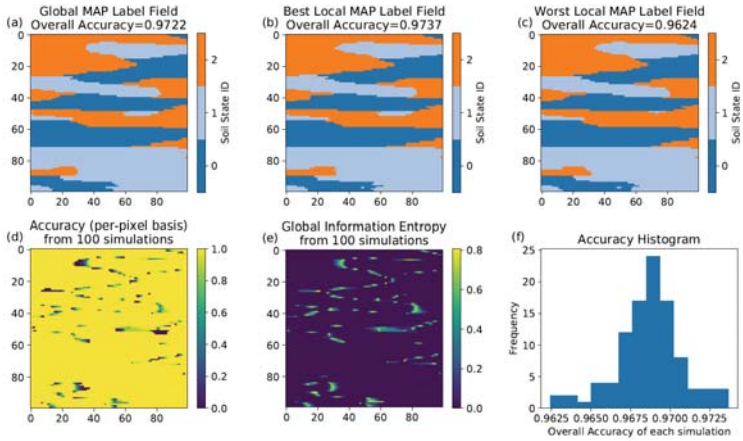

Figure 4. Estimation results using 100 chains of Example 1.

vertical correlations of the observed field are introduced as a more realistic scenario and the simulated observed fields are shown in Figure 5(a-b) using parameters in Table 1 and conditional on the same true latent field in Example 1 (Figure 5(c)). Same as Example 1, 15% of the columns are randomly selected as "known" CPT soundings (Figure 5(d-f)).

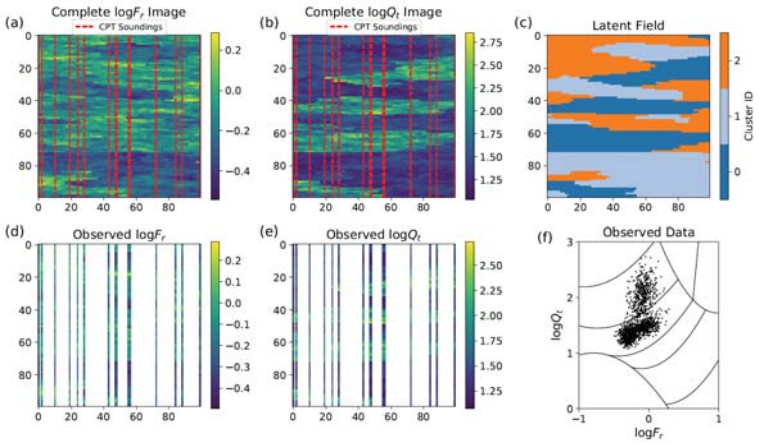

Figure 5. Simulated synthetic data of Example 2.

Given these spatially correlated observed pixels, the optimal number of clusters still can be correctly identified as three via BIC (Figure 6(a)) and, then, a single simulation of 1000 iterations is performed. The inferred local MAP label field (Figure 6(b)) achieved a decent overall accuracy while some small patches of Cluster 1 are misclassified as Cluster 2. A possible reason could be that these two clusters are partially overlapped in feature space (Figure 6(e)), and the spatial correlation may cause a small group of data points close to each other in both feature space and physical space, which may result in a bulk misclassification. Generally, the local information entropy level (Figure 6(d)) is similar compared with Example 1. Then, $N_{mc} = 100$ parallel Markov chains are simulated for Example 2 and the results are shown in Figure 7. The global MAP label field has an overall accuracy of 95.32%, which is slightly lower than that in Example 1 while the global information entropy is noticeably higher than that of Example 1 due to the higher likelihood of bulk misclassification. Accordingly, the variability of the

local MAP accuracy (Figure 7(f)) is also more significant compared to that of Example 1. The results indicate that the spatial correlation of the observed fields, though does not substantially compromise the performance of the proposed approach, still can introduce certain difficulties and reduce the local variability yet increase the global variability of the inferred label field. This observation indicates that the performance of the proposed approach may be compromised when applied to sites having strong spatial correlation. Both qualitative and quantitative assessments in this regard will be performed to have a better understanding of this limitation. A critical spatial correlation length or scale of fluctuation is expected for defining the application scope of the proposed approach.

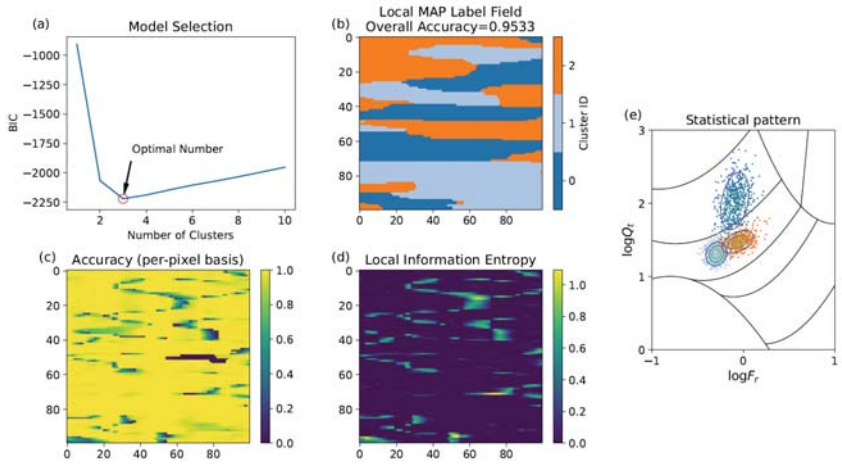

Figure 6. Estimation results from a single simulation of Example 2.

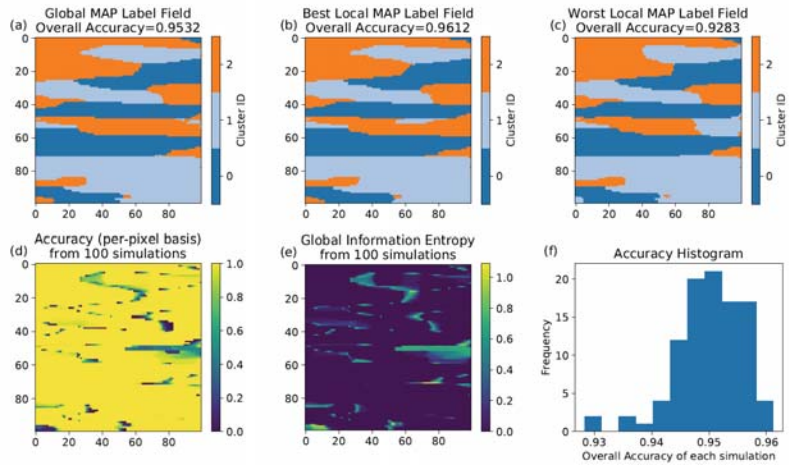

Figure 7. Estimation results using 100 chains of Example 2.

6 Conclusions

This paper contributes to the techniques leveraging valuable data asset for decision-making in geotechnical risk management. A hybrid HMRF model is developed for modeling the stratigraphic heterogeneity. CPT soundings are used as data source, from which both spatial and statistical patterns are extracted by the in-house developed Bayesian unsupervised learning algorithm. The stratigraphic uncertainty is quantified using simulated realizations. The proposed approach aims at improving the interpretation performance of the data-driven and uncertainty aware techniques for geotechnical site characterization regarding accuracy and quantified uncertainty, especially in the situation that only sparse CPT soundings are available. The preliminary results from synthetic examples demonstrate that the model parameters of the proposed hybrid random field can be initially defined in terms of prior distributions based on prior geological knowledge if available or leave as default. Later these parameters are further updated and optimized with constraints from the site exploration results through Bayesian machine learning. Specific attention has been paid on the conditional independency assumption adopted in the proposed model via conducting numerical experiments. The results indicate that the spatial correlation of soil properties, though does not substantially compromise the performance of the proposed approach, still cause certain difficulties in distinguishing similar soil layers in both physical and feature space. Currently, finding a critical spatial correlation length or scale of fluctuation is on-going for a better definition of the application scope. More detailed results on this track will be reported in future submissions.

Acknowledgments

This work is partially supported by the Ohio Department of Transportation under the Agreement Number 31795 Subtask 6, and partially supported by the STEM Catalyst grant from the University of Dayton. The financial supports are gratefully acknowledged.

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 1281-1311.

Celeux, G., Forbes, F., and Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern recognition*, 36(1), 131-144.

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. *Canadian Geotechnical Journal*, 52(12), 1993-2007.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1), 1-18.

Forbes, F., and Peyrard, N. (2003). Hidden Markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), 1089-1101.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721-741.

Gong, W., Tang, H., Wang, H., Wang, X., and Juang, C. H. (2019). Probabilistic analysis and design of stabilizing piles in slope considering stratigraphic uncertainty. *Engineering Geology*, 259, 105162.

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.

ISO (2015). ISO 2394: General principles on reliability for structures. ISO Geneva.

Juang, C. H., Zhang, J., Shen, M., and Hu, J. (2018). Probabilistic methods for unified treatment of geotechnical and geological uncertainties in a geotechnical analysis. *Engineering geology*, 249, 148-161.

Lu, X., Li, P., Chen, B., and Chen, Y. (2005). Computer simulation of the dynamic layered soil pile structure interaction system. *Canadian geotechnical journal*, 42(3), 742-751.

McLachlan, G., and Peel, D. (2004). *Finite mixture models*, John Wiley & Sons, Hoboken, N.J.

Pereyra, M., Dobigeon, N., Batatia, H., and Tourneret, J.-Y. (2013). Estimating the granularity coefficient of a Potts-Markov random field within a Markov Chain Monte Carlo algorithm. *Image Processing, IEEE Transactions on*, 22(6), 2385-2397.

Phoon, K.-K. (2020). The story of statistics in geotechnical engineering. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 14(1), 3-25.

Phoon, K.-K., Ching, J., and Wang, Y. Managing risk in geotechnical engineering—From data to digitalization. *Proc., Proceedings, 7th International Symposium on Geotechnical Safety and Risk (ISGSR 2019), Taipei, Taiwan*, 13-34.

Phoon, K.-K., Retief, J. V., Ching, J., Dithinde, M., Schweckendiek, T., Wang, Y., and Zhang, L. (2016). Some observations on ISO2394: 2015 Annex D (reliability of geotechnical structures). *Structural Safety*, 62, 24-33.

Robertson, P. (1990). Soil classification using the cone penetration test. *Canadian Geotechnical Journal*, 27(1), 151-158.

Robertson, P. (2009). Interpretation of cone penetration tests—a unified approach. *Canadian geotechnical journal*, 46(11), 1337-1355.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464.

Song, S., Si, B., Herrmann, J. M., and Feng, X. (2016). Local Autoencoding for Parameter Estimation in a Hidden Potts-Markov Random Field. *IEEE Trans Image Process*, 25(5), 2324-2336.

Wang, H., Wang, X., Wellmann, F., and Liang, R. Y. (2018). A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data. *Canadian Geotechnical Journal*, 56(8), 1184-1205.

Wang, H., Wellmann, J. F., Li, Z., Wang, X., and Liang, R. Y. (2016). A Segmentation Approach for Stochastic Geological Modeling Using Hidden Markov Random Fields. *Mathematical Geosciences*, 49(2), 145-177.

Wang, X., Li, Z., Wang, H., Rong, Q., and Liang, R. Y. (2016). Probabilistic analysis of shield-driven tunnel in multiple strata considering stratigraphic uncertainty. *Structural safety*, 62, 88-100.

Wang, X., Wang, H., and Liang, R. Y. (2017). A method for slope stability analysis considering subsurface stratigraphic uncertainty. *Landslides*, 1-12.

Wang, X., Wang, H., Liang, R. Y., and Liu, Y. (2019). A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data. *Engineering Geology*, 248, 102-116.

Wang, X., Wang, H., Liang, R. Y., Zhu, H., and Di, H. (2018). A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data. *Structural Safety*, 70, 128-138.

Wang, Y., Hu, Y., and Zhao, T. (2019). 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.

Wang, Y., Huang, K., and Cao, Z. (2013). Probabilistic identification of underground soil stratification using cone penetration tests. *Canadian Geotechnical Journal*, 50(7), 766-776.

Wellmann, J. F., and Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. *Tectonophysics*, 526, 207-216.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Medical Imaging*, 20(1), 45-57.

Zhu, H., and Zhang, L. (2013). Characterizing geotechnical anisotropic spatial variations using random field theory. *Canadian Geotechnical Journal*, 50(7), 723-734.