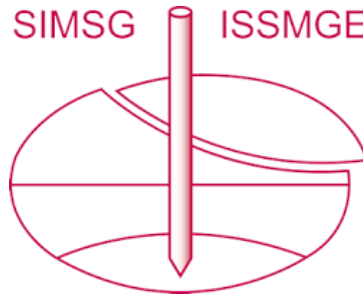


INTERNATIONAL SOCIETY FOR SOIL MECHANICS AND GEOTECHNICAL ENGINEERING



This paper was downloaded from the Online Library of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE). The library is available here:

<https://www.issmge.org/publications/online-library>

This is an open-access database that archives thousands of papers published under the Auspices of the ISSMGE and maintained by the Innovation and Development Committee of ISSMGE.

The paper was published in the proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering and was edited by Mizanur Rahman and Mark Jaksa. The conference was held from May 1st to May 5th 2022 in Sydney, Australia.

Stratigraphic uncertainty quantification using Bayesian machine learning and Markov random field

Quantification de l'incertitude stratigraphique à l'aide de l'apprentissage automatique bayésien et du champ aléatoire de Markov

Hui Wang

Department of Civil and Environmental Engineering, The University of Dayton, USA, hwang12@udayton.edu

Xingxing Wei

Department of Civil and Environmental Engineering, The University of Dayton, xwei02@udayton.edu

School of Civil Engineering, Central South University, China

ABSTRACT: In the field of geotechnical site characterization, the Markov random field (MRF) models, used to be widely applied in physics and image processing, recently have gained much attention, and have been adopted for modeling and simulating stratigraphic heterogeneity using limited/sparse boreholes, performing which by far is still a challenging task in engineering geology. The gap lies in the difficulty of developing an integration of subjective engineering judgment (e.g., geological knowledge) and the objective site exploration data (e.g., the borehole observations). Recently, Bayesian machine learning, with its advances in artificial intelligence and uncertainty quantification, has been widely applied in a variety of research fields. In this paper, we combine Bayesian machine learning with the MRF model. An effective MRF stochastic modeling framework is developed to characterize the stratigraphic uncertainty. The model parameters are initially defined in terms of prior distribution. These parameters are then further calibrated with additional constraints from the site exploration results using Bayesian machine learning. Throughout the learning process, the stratigraphic uncertainty (inherent) and the model uncertainty (imperfect knowledge) are taken into consideration. The effects of both prior knowledge and borehole observations in quantifying stratigraphic uncertainty are discussed. To demonstrate the effectiveness of the developed approach, both synthetic and real-world examples are demonstrated. We envision this approach can be further generalized in industry practices for improved risk control in geotechnical engineering.

RÉSUMÉ : Ci-joint les instructions pour la préparation de votre communication au 20ème congrès CIMSG de Sydney, 12-17 Septembre 2021. Les articles, écrits en Times New Roman 9 ne doivent pas dépasser **6 pages A4** et être fournis sous **format MS Word (.docx,)** et **PDF**. Les résumés ne doivent pas dépasser 10 lignes. Il n'est pas demandé de transfert de copyright mais seulement une autorisation de publication. Il n'y aura pas de volume post-congrès pour les communications en retard. Vous êtes invités à utiliser directement cette feuille canevas pour mettre en forme votre contribution. Pour un article en français, inverser la disposition des titre et résumés avec les mots clés en français et anglais (ajouter « Mots-clés »).

KEYWORDS: Markov random field; Bayesian machine learning; stratigraphic uncertainty

1 INTRODUCTION

Acquiring accurate site-specific soil layer information is a crucial and essential step for planning and design of any geotechnical project. However, subsurface soil layers are natural forming materials associated with inherent heterogeneity and randomness (Juang et al. 2018). Therefore, the design and construction of a geotechnical system need to take these randomness into account. Due to inadequate knowledge of the soil forming histories and/or other geological/human activities, the subsurface information at a project site can be difficult to infer (Gong et al. 2019; Wang et al. 2016).

Only sparse borehole logs are collected in a project, partially due to the limited budget and the tight project schedule. As a result, the geological and geotechnical information only can be probed at sparsely distributed locations; whereas, subsurface information at other locations may have to be inferred or simulated based on available sparse information either from archived borehole data or planned site investigation data. The incomplete knowledge of the formation process of the geological bodies, together with the insufficient number of borehole logs and in-situ test results, leads to significant uncertainty in the inferred subsurface profile. It is fair to say that the issues of the subsurface uncertainty and the influence of such uncertainty on the geotechnical design have long posed challenges to practitioners (Gong et al. 2019).

To acquire reasonable and confident estimates of subsurface uncertainty, Markov random field (MRF)-based stochastic simulation approaches have been adopted in engineering geology. The MRF models can provide a flexible and intuitive way to describe the Markovian contextual constraints, which enables the reflection and reproduction of the anisotropy and heterogeneity of subsurface geological structures. While the current MRF models (Gong et al. 2019; Wang et al. 2016; Wang et al. 2018; Wang et al. 2019; Wang et al. 2018) have some shortcomings, such as the model parameters, representing the spatial correlation of stratigraphic structures, need to be defined in priori by using subjective engineering judgments from local experience and the parameters cannot be updated during the inferential process once they are determined initially. This strategy is not robust as the simulated subsurface profile may deviate from reality due to the subjective guess of model parameters and the uncertainty could be underestimated.

In this study, a novel stratigraphic uncertainty quantification approach is developed by combining the MRF theory with Bayesian machine learning (Wang 2018). The new approach has a more advanced and flexible spatial correlation model so that it can better leverage the prior information of the model parameters and enhance uncertainty quantification regarding soil heterogeneity when only sparse borehole information is available. The model parameters are regularized in terms of prior distribution and further updated with additional constraints from the site exploration results using Bayesian machine learning. As

a result, the developed new approach not only relies on fewer subjective information, but also has better adaptivity compared with previous MRF-based methods. To demonstrate the effectiveness of the developed approach, both synthetic and real-world cases are studied and some preliminary results are presented in this paper.

2 METHOD

2.1 Markov Random Field

One of the practical approaches analyzing sparse geotechnical data is to discretize a subset (e.g., a two-dimensional section) of the physical space into pixels according to the measurement resolution. For categorical data (as discussed in this paper), each pixel is assigned with a label indicating the associated soil type.

A Markov random field model is a graphical description of the spatial pattern (e.g., the soil layers in the current context) in physical space. Pixels having the same label belong to the same soil type. The label field can be represented as a label *configuration* of all pixels $\mathbf{x} = (x_1, x_2, x_3, \dots, x_N)$, $x_i \in L$ where $L = \{1, 2, 3, \dots, m\}$ is a set of all possible labels indicating different soil types. A typical example of a graph model describing the spatial correlation of a Markov random field is a two-dimensional lattice with a second-order neighborhood system (Besag 1986). For pixel i , the neighbors ∂_i are defined as the nearest eight pixels around it. The local conditional probabilities of a specific label given the labels of all neighbors can be calculated in the following form (Besag 1986; Geman and Geman 1984):

$$P(x_j | \mathbf{x}_{\partial_j}) = \frac{P(x_j, \mathbf{x}_{\partial_j})}{\sum_{x_j' \in L} P(x_j', \mathbf{x}_{\partial_j})} = \frac{\exp[-U(x_j, \mathbf{x}_{\partial_j})]}{\sum_{x_j' \in L} \exp[-U(x_j', \mathbf{x}_{\partial_j})]} \quad (1)$$

where $U(\cdot)$ is the so-called local energy. We adopt the widely used Potts model (Koller and Friedman 2009) to characterize the local interaction of the pixels within a neighborhood system. The local energy has the following form:

$$U(x_j, \mathbf{x}_{\partial_j}) = \sum_{i \in \partial_j} V_{i,j}(x_i, x_j) \quad (2)$$

with the potential function

$$V_{i,j}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \beta_d & \text{if } x_i \neq x_j \end{cases} \quad (3)$$

where $\beta_d \in \{\beta_1, \beta_2, \beta_3, \beta_4\}$ indicates the spatial constraint corresponding to four independent directions $0, \pi/2, \pi/4, 3\pi/4$ referring to a two-dimensional lattice grid, and is referred to as the *granularity coefficients*. The local energy reflects the spatial correlation of categorical data: pixels close to each other tend to have the same soil type. The behavior of an MRF model is intimately related to the granularity coefficients $\{\beta_d\}$. For an anisotropic Potts model (as introduced in this work), positive values of these parameters cause attraction of neighboring pixels, or encourage clustering effects along a certain direction, while negative values result in repulsion, or prevent clustering (Cross and Jain 1983).

2.2 Bayesian machine learning

All pixels can be categorized into two types: a) pixels with known labels indicating sparse borehole information $\mathbf{x}_{BH}$ and b) pixels with unknown labels $\mathbf{x}_{unknown}$ elsewhere. Both $\mathbf{x}_{unknown}$ and the granularity coefficients β need to be inferred from $\mathbf{x}_{BH}$. A Markov chain Monte Carlo (MCMC) technique is employed to implement Bayesian machine learning and sample $(\mathbf{x}_{unknown}, \beta)$

iteratively via two conditional *a posteriori* distributions $P(\mathbf{x}_{unknown} | \mathbf{x}_{BH}, \beta)$ and $P(\beta | \mathbf{x}_{unknown}, \mathbf{x}_{BH})$ iteratively.

2.2.1 Simulation of conditional MRF $P(\mathbf{x}_{unknown} | \mathbf{x}_{BH}, \beta)$

Given a random initial field at unknown pixels $\mathbf{x}_{unknown}$, and conditional on a specific setting of granularity coefficients and pixels with known soil type at borehole locations $\mathbf{x}_{BH}$, $P(\mathbf{x}_{unknown} | \mathbf{x}_{BH}, \beta)$ is a Gibbs distribution with fixed soil labels only at the borehole locations. The local energy at unknown pixels can be calculated using Eq. (2) and the corresponding probability of choosing each label can be evaluated via Eq. (1). Realizations of the conditional random field $P(\mathbf{x}_{unknown} | \mathbf{x}_{BH}, \beta)$ can be simulated via a parallel algorithm named chromatic sampler (Wang et al. 2016).

2.2.2 Simulation of the model parameters $P(\beta | \mathbf{x}_{unknown}, \mathbf{x}_{BH})$

In this step, β is sampled following the conditional posterior distribution:

$$\text{Post}(\beta) \propto \text{Prior}(\beta) L(\mathbf{x}_{unknown}, \mathbf{x}_{BH} | \beta) \quad (4)$$

where $\text{Post}(\beta)$ is the posterior distribution of β ; $\text{Prior}(\beta)$ is the prior distribution of β ; $L(\mathbf{x}_{unknown}, \mathbf{x}_{BH} | \beta)$ is the likelihood function indicating the possibility of having the simulated soil configuration given the known borehole information and can be evaluated via the following equation.

$$L(\mathbf{x}_{unknown}, \mathbf{x}_{BH} | \beta) = \prod_{x_j \in \{\mathbf{x}_{unknown}, \mathbf{x}_{BH}\}} P(x_j | \mathbf{x}_{\partial_j}; \beta). \quad (5)$$

To incorporate prior knowledge, the prior distribution of β should be defined. In this work, a multivariate Gaussian distribution with a mean vector μ indicating the rough estimates of the granularity coefficients, and a diagonal covariate matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$, where σ_i is the standard deviation of the corresponding granularity coefficient.

The Metropolis-Hasting algorithm is employed to implement the conditional MCMC sampling process. The log(target) function can be expressed as

$$\log(\text{target}) = \log(\text{Prior}(\beta)) + \log(L(\mathbf{x}_{unknown}, \mathbf{x}_{BH} | \beta)). \quad (6)$$

The log(target) function measures the log scale of the joint probability of the simulated granularity coefficients and the simulated soil profile. The higher the log(target) is, the higher possibility that the simulated soil profile is plausible and the corresponding granularity coefficients is compatible with the simulated field. In other words, Eq. (6) is being optimized in a probabilistic sense during the Bayesian machine learning process through MCMC. Since no training information is needed during the sampling process, this method is fully unsupervised.

2.2.3 Maximum a posteriori (MAP) of soil profile and the most likely realization

After the burn-in period of the Markov chain, each realization of the simulated soil profile is stored and the marginal probability of choosing each label is calculated for each pixel. The MAP estimate of the entire soil profile can be derived based on the majority vote principle at each pixel (i.e., the MAP label will be the one with the highest sampling probability). The most likely realization is defined as the realization corresponding to the highest log(target) value.

2.3 Uncertainty-aware algorithm

The marginal probability of each soil label derived from the simulated realizations takes the uncertainty of granularity coefficients β into consideration, and hence unlike other non-

Bayesian methods with fixed model parameters, this algorithm automatically assesses possible model bias and incorporates it into the overall uncertainty quantification. This is one of the contributions of the current work.

3 SYNTHETIC CASE STUDY

The stratigraphic uncertainty estimated using an MRF model includes both *local uncertainty* and *global uncertainty* in *configuration space* when prior information of β and boreholes data are provided. The illustration of local and global uncertainties is shown in the Figure 1. Local uncertainty exists in a single MCMC simulation given a random initial field, and each realization in a single simulation is the result of the local uncertainty (see Figure 1 (a), (b), (c)). While the global uncertainty refers to the variation of the MAPs (or local optimums) derived from multiple Markov chains (see Figure 1 (d), (e), (f)) generated from different initial fields.

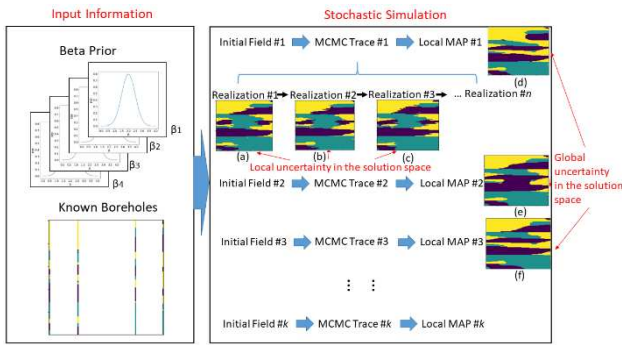


Figure 1. Illustration of simulated stratigraphic uncertainty.

In this section, we focus on discussing how to determine a suitable prior distribution of β when only sparse boreholes are available. A synthetic stratigraphic profile shown in Figure 2(a) is simulated using a Potts model with $\beta=[4, 0.1, 0.1, 0.1]$ and five virtual boreholes are extracted and shown in Figure 2(b) for inferring the “unknown” portion. The soil profile in Figure 2(a) is considered as the “ground truth” throughout the synthetic example.

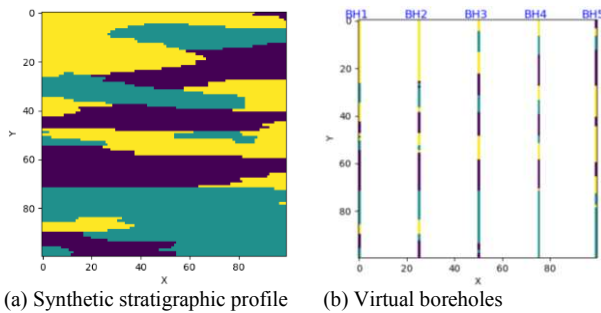


Figure 2. Synthetic stratigraphic profile and selected virtual boreholes.

3.1 Prior information of β

The spatial correlation of a soil profile tends to be horizontally dominated. Therefore, it is intuitive that β_1 should be much greater than β_2 , β_3 , and β_4 . For simplicity, β_2 , β_3 , and β_4 share the same prior distribution, while β_1 has its own prior distribution.

3.1.1 Suitable range of β

Different prior information of β can result in different local optimums in the configuration space. Yet only a small portion of local optimums controlled by reasonable β are compliant with geological knowledge, and can be considered as possible

candidates. In order to figure out a reasonable range of β , we first choose a significantly high β_1 , such as 100, and then get the MAPs (local optimums) by gradually increasing $\beta_{2,3,4}$. The realizations with different $\beta_{2,3,4}$ are shown in Figure 3. It can be seen that the realizations are less realistic when $\beta_{2,3,4} \geq 0.3$. Although a strong β_1 could result in a horizontally dominated layered pattern, while it is only under the condition that $\beta_{2,3,4}$ is small enough. This indicates that $\beta_{2,3,4}$ should well controlled. According to experiences, the reasonable range for $\beta_{2,3,4}$ could be 0~0.3.

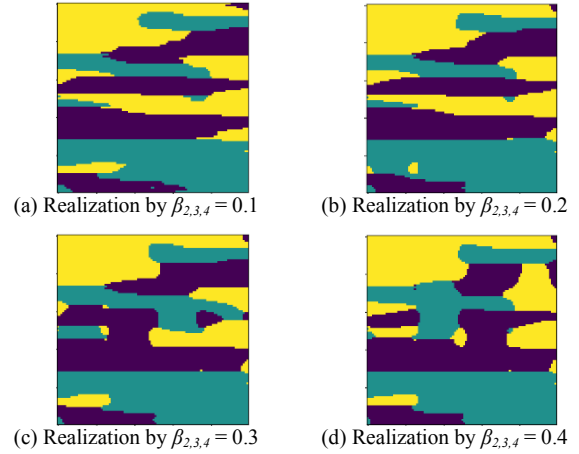


Figure 3. Realizations with different $\beta_{2,3,4}$ when $\beta_1 = 100$.

In the next experiment, $\beta_{2,3,4}$ is set to be 0.1, and β_1 gradually increases from 1. Typical realizations are shown in Figure 4. Obviously, the realizations with weak β_1 do not conform to the geological knowledge as the horizontal constraint cannot extend the information of neighboring boreholes to a long distance. On the other hand, the realizations using β_1 starting from 3 show reasonable layered patterns. Empirically, a suitable range of β_1 can be greater than 3.

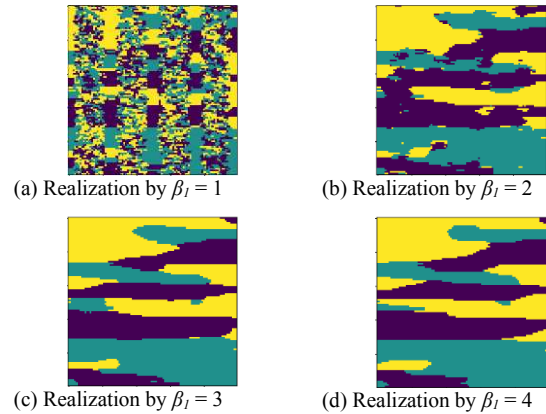
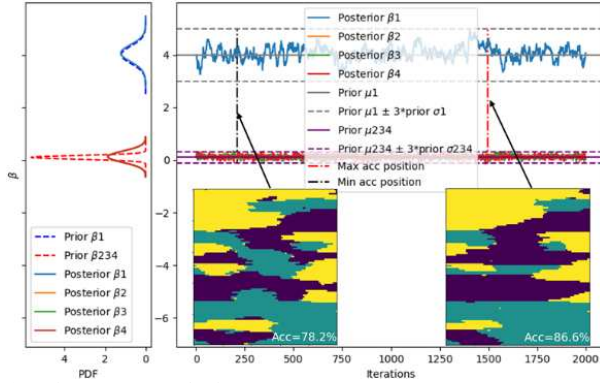


Figure 4. Realizations with different β_1 when $\beta_{234} = 0.1$.

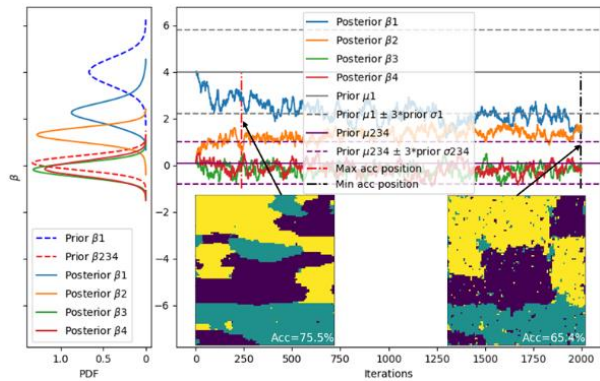
3.1.2 Prior setting of β

For $\beta_{2,3,4}$, a prior mean $\mu_{2,3,4}$ of 0.1 is chosen, then a prior standard deviation $\sigma_{2,3,4}$ of at least 0.07 is selected to guarantee the varying range $[\mu_{2,3,4}-3\sigma_{2,3,4}, \mu_{2,3,4}+3\sigma_{2,3,4}]$ covers 0~0.3. While σ_1 needs to be chosen according to $\mu_1=4$. To have a better illustration on the behavior, Figure 5 (a) and (b) shows the simulation results using a prior with small σ_1 , $\sigma_{2,3,4}$ and a prior with big σ_1 , $\sigma_{2,3,4}$, respectively. The images labeled as “Max Acc” and “Min Acc” are the realizations having highest and lowest accuracy compared with the ground truth. It is observed that both posterior β and estimated soil profile (see Max Acc image and Min Acc image) have good results using small σ_1 , $\sigma_{2,3,4}$, while posteriors of β_1 , β_2

tend to mix with each other and drift out of the original range $[\mu - 3\sigma, \mu + 3\sigma]$ seriously when big $\sigma_1, \sigma_{2,3,4}$ is adopted. As a result, the realizations are in poor quality, though, posteriors of β_3 and β_4 (controlling the direction of $\pi/4$ and $3\pi/4$) seem to perform well.



(a) Prior $\sigma_1=0.33$ and prior $\sigma_{2,3,4}=0.07$



(b) Prior $\sigma_1=0.5$ and prior $\sigma_{2,3,4}=0.3$

Figure 5. Bayesian estimations using prior $\mu_1=4, \mu_{2,3,4}=0.1$ and different standard deviation settings.

The reason is that the algorithm tries to minimize the total energy of the configuration during the parameter estimation process. The way to reduce the total energy is to reduce β_1 while increase β_2 under the current configuration (i.e., vertically dense while horizontally sparse). Consequently, the equilibrium configuration with similar β_1 and β_2 possess a lower total energy. To compensate for this issue, appropriate regularizations on the standard deviations $\sigma_{2,3,4}$ for controlling the variation of β_1 and β_2 is needed, which prevents them from mixing with each other.

To further demonstrate the above finding, Table 1 shows the estimation results of β_1 and β_2 with different $\sigma_{2,3,4}$. As $\sigma_{2,3,4}$ increases, posterior μ_1 and posterior μ_2 get closer gradually. It can be noticed that the smaller $\sigma_{2,3,4}$ is, the more favorable it is to the stochastic simulation process as the drifting effects can be well constricted via the regularization using small standard deviation.

Table 1. Effect of different prior $\sigma_{2,3,4}$ when prior $\sigma_1=0.34$

Prior σ_{234}	Posterior mean		β for Max Acc		
	μ_1	μ_2	Acc	β_1	β_2
0.07	4.07	0.13	87.4%	4.45	0.09
0.20	3.43	0.48	83.3%	3.95	0.31
0.30	3.40	0.65	82.5%	3.87	0.45
0.40	3.36	1.03	82.2%	3.83	0.85
0.50	3.22	1.38	76.3%	3.28	1.42

The estimation results of β_1 and β_2 with gradually increased σ_1 while keep $\sigma_{2,3,4}$ as a constant is shown in Table 2. Posterior

mean μ_1 has a small increase over a long range (i.e. 0.33~10), and posterior mean μ_2 is stable subject to a strong regularization. It can be concluded that the growth of prior σ_1 has almost no effect on the estimation process. A large value of 10 is used as the prior setting of σ_1 , which enable β_1 to explore in a large range in the process of stochastic simulation.

Table 2. Effect of prior σ_1 when prior $\sigma_{2,3,4}=0.07$

Prior σ_1	Posterior mean		β for Max Acc		
	μ_1	μ_2	Acc	β_1	β_2
0.34	4.07	0.13	87.4%	4.45	0.09
0.50	4.20	0.13	84.6%	4.02	0.17
0.60	4.33	0.12	85.5%	4.93	0.12
...					
1	4.36	0.13	84.4%	4.47	0.12
...					
10	4.63	0.13	87.3%	5.82	0.13

3.1.3 Suitable choice of μ_1

Different prior mean μ_1 is used to perform multiple synthetic experiments. Four boreholes (i.e. BH1, BH2, BH4, BH5 as shown in Figure 2(b)) are used for stochastic simulation, and BH3 is used as the validation borehole.

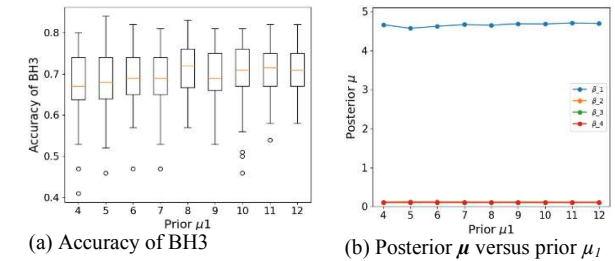
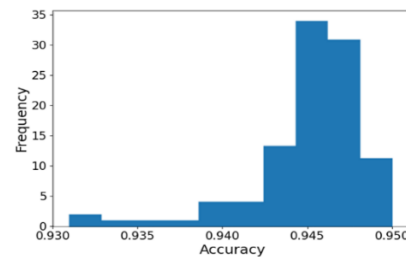


Figure 6. Estimation results with different prior mean μ_1 .

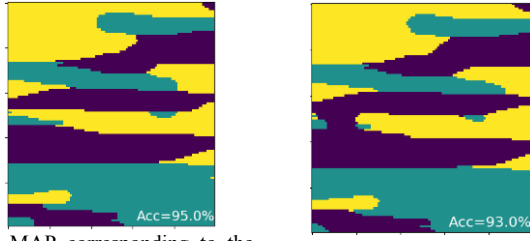
The best (regarding the accuracy at BH3) profiles using different prior mean μ_1 generally have a high similarity to each other via visual inspection (not shown here). Quantitative results are shown in Figure 6 (a,b). The medians of the accuracy corresponding to different prior mean μ_1 basically remain around 0.69. The results indicate that the posteriors of β always converge into a local optimal regardless the setting of prior mean μ_1 . This result demonstrate the robustness of the proposed approach.

3.2 Estimate the stratigraphic uncertainty

The prior parameter μ_1 and $\mu_{2,3,4}$ is set to be 4.60 and 0.10, respectively. A small σ for β (i.e. $\sigma_1=0.1, \sigma_{2,3,4}=0.07$) is chosen for applying the regularization. The histogram of accuracy from 100 simulations is shown in Figure 7. As can be noticed, most of the results fall in the range 93% - 95%, with a minimum accuracy of 93.0% and a maximum accuracy of 95.0%.

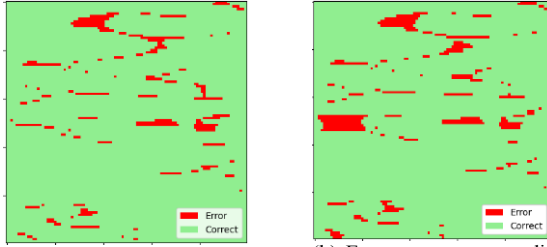


(a) Distribution of accuracy



(b) MAP corresponding to the Max Accuracy (c) MAP corresponding to the Min Accuracy

Figure 7. Validation results of the synthetic example.



(a) Error map corresponding to the Max Accuracy (b) Error map corresponding to the Min Accuracy

Figure 8. Error maps of the synthetic example.

The simulated profiles corresponding to the minimum accuracy and maximum accuracy are shown in Figure 7. The accuracy is computed from 2000 realizations by using Eq. (7),

$$Acc = \frac{\sum_{i=1}^n I(Z_R(x_i) = Z_T(x_i))}{n} \quad (7)$$

where $I(\cdot)$ is an indication function and equals to 1 when the i -th element in realizations $Z_R(x_i)$ has the same categorical value (e.g., soil type) as that from the ground truth (see Figure 2(a)) $Z_T(x_i)$, and n represents the total number of realizations (i.e. 2000). Figure 8 shows that the error pixels are mainly concentrated at the boundaries of different soil layers, which agrees with our intuition.

4 REAL-WORLD EXAMPLE STUDY

In this section, the developed algorithm is applied to a construction site in Hong Kong. The same case history has been studied by Li et al. 2019 use Coupled Markov Chain (CMC) modeling approach.

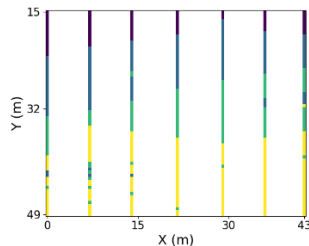


Figure 9. Known borehole data of the Hong Kong case.

The known boreholes are shown in Figure 9. BH1, BH2, BH5 and BH7 are used to estimate model parameters and BH3, BH4 and BH6 are used as validation boreholes. The estimation results for different $\beta_{2,3,4}$ when $\beta_1=100$ are shown in the Figure 10, from which we can notice that the geological information of the thin patches indicated by several boreholes have no horizontal extension due to the strong vertical constraints caused by β_2 starting from $\beta_{2,3,4}=0.2$. Therefore, the range of suitable $\beta_{2,3,4}$ is

0-0.2. The trials for choosing β_1 are shown in Figure 11. Obviously, the simulation results do not conform to the geological knowledge when β_1 is less than 2.5.

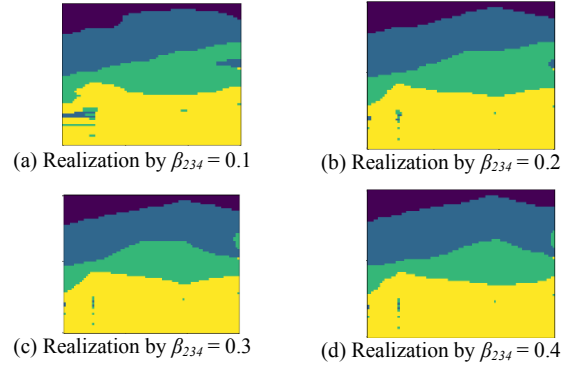


Figure 10. Realizations with different $\beta_{2,3,4}$ when $\beta_1 = 100$.

Since the larger $\sigma_{2,3,4}$ is, the more unfavorable the result will be, $\sigma_{2,3,4}$ takes the minimum value 0.07 when set $\mu_{2,3,4}=0.1$, with which the range $[\mu_{2,3,4}-3\sigma_{2,3,4}, \mu_{2,3,4}+3\sigma_{2,3,4}]$ covers 0-0.3. It is more preferred for σ_1 to have a larger value (i.e. 10).

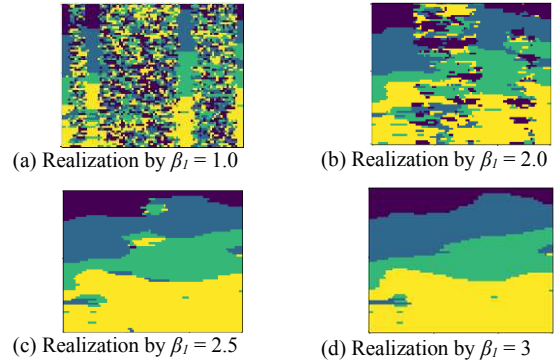
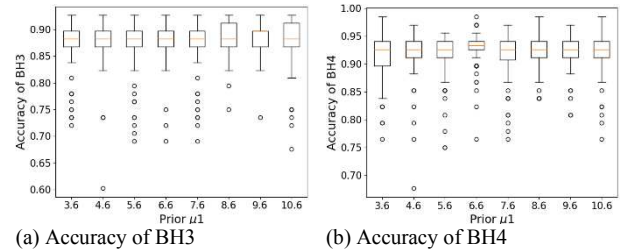


Figure 11. Realizations with different β_1 when $\beta_{2,3,4} = 0.1$.

We investigate the effect of prior knowledge by choosing different μ_1 . As can be seen from Figure 12(a), (b) and (c), the accuracy of BH3, BH4 and BH6 basically remain at 0.88, 0.93 and 0.93, respectively. During the stochastic simulation, the posterior distributions of β converge to a reasonable value as long as the prior mean $\mu_1 > 3$ is chosen, as can be seen in Figure 12(d) and the stable posterior mean is approximate 4.15 for β_1 and 0.10 for $\beta_{2,3,4}$. Again, it shows strong robustness. Then, prior parameter μ_1 is set to be 4.15 and $\mu_{2,3,4}$ is set to be 0.10, by using small $\sigma_1=0.1$ and $\sigma_{2,3,4}=0.07$, all known boreholes are used to infer the soil profile. The final results are shown in Figure 13.



(a) Accuracy of BH3 (b) Accuracy of BH4

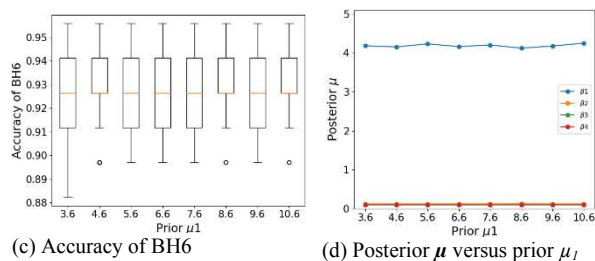


Figure 12. Estimation results with different prior μ_l .

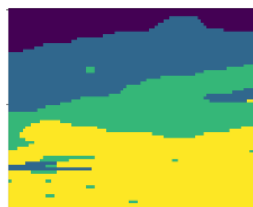


Figure 13. Estimated soil profile of the Hong Kong case.

5 CONCLUSIONS

In this paper, a Markov random field-based soil profile modeling approach is presented. Bayesian machine learning is integrated with MRF models to infer and simulate subsurface profiles. The model parameters are initially defined in terms of prior distribution, and these parameters are further calibrated with constraints from the site exploration results (e.g., boreholes) using Bayesian machine learning. Both synthetic and real-world examples are used to validate the method and demonstrate its robust performance. It can be seen that, from the simulation results, the performance of this approach does not rely on the spatial pattern of the stratigraphy and/or the soil types, and is independent of the stratigraphical setting or complexity.

Though the theoretical basis is the well-developed MRF, significant contributions are made to push forward this methodology toward real-world engineering applications with a higher technical readiness level. The excellent performance in both synthetic and real-world examples of the developed approach can demonstrate its effectiveness and robustness. To be more specific, the novelty can be summarized into three points: 1) A more flexible anisotropic Potts model is used for complex heterogeneity simulation; 2) a proper setting for parameter priors so that the model uncertainty/bias can be considered with less assumptions; and 3) both local and global uncertainties in the configurations space are taken into consideration for inferring the soil profiles.

Note that subsurface modeling is a routine process and essential step for geotechnical design and construction. For example, in tunnel projects, this new approach can be applied to analyze the effect of stratigraphic uncertainty on the supporting system performance of tunnels and obtain uncertainty-aware evaluations and reasonable design for tunnel construction in ground conditions with complex strata using sparsely located borehole information. The developed approach also can be employed to identify the uncertainty of foundation systems of wind turbines to be installed within a wind farm since predicting soil conditions away from boreholes with quantified uncertainty is important as often the wind farm layout changes during the development phase, and hence, information at some specific foundation locations may not be readily available. Furthermore, it can be applied for decision-making regarding general geotechnical risk management, namely whether additional budget should be spent on the site characterization (i.e., more boreholes, hence more accurate subsurface profile resulting in less budget for design and construction) or on structural design

and construction (i.e., less budget on boreholes while implement more conservative design).

The new approach still has certain limitation. More specific, it may generate unrealistic initial stratigraphic configurations without proper regularization at the beginning of the stochastic simulation. And hence this behavior may slow down the convergence rate and affect the computational efficiency. We are actively working on some promising solutions to mitigate this behavior and more results will be presented in another paper. As this new approach is more and more complete and polished, we envision this approach can be further promoted and applied in industry practices for improved risk control in geotechnical engineering.

6 ACKNOWLEDGEMENTS

This work is partially supported by the Ohio Department of Transportation under the Agreement Number 31795 Subtask 6: A Study of AI Based Methods for Characterization of Geotechnical Site Investigation Data, and partially supported by the STEM Catalyst grant from the University of Dayton. The financial supports are gratefully acknowledged.

7 REFERENCES

- Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3), 259-302.
- Cross, G. R., and Jain, A. K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(1), 25-39.
- 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.
- 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.
- Koller, D., and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*, MIT press.
- Li, J., Cai, Y., Li, X., & Zhang, L. (2019). Simulating realistic geological stratigraphy using direction-dependent coupled Markov chain model. *Computers and Geotechnics*, 115, 103147.
- Wang, H. (2018). Finding patterns in subsurface using Bayesian machine learning approach. *Underground Space*, 2018.
- 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., 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.
- 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.