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_01-003-cd

Probabilistic Characterization of 3D Non-Stationary Spatial Variability from Limited Boreholes Using Bayesian Supervised Learning

Yue Hu[1] and Yu Wang[2]

[1]Department of Civil and Environmental Engineering, National University of Singapore, Lower Kent Ridge Road, Singapore.
E-mail: yuehu@nus.edu.sg; yuehu47-c@my.cityu.edu.hk
[2]Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Hong Kong, China.
E-mail: yuwang@cityu.edu.hk

Abstract: Natural geomaterials are affected by many spatially varying factors during complex geological formation process and exhibit spatial variability and heterogeneity in three-dimensional (3D) space. Real geotechnical structures always interact with geomaterials with 3D spatial variability. However, in engineering practice, due to the limitation of investigation borehole number and the complex nature of geomaterials properties, it is challenging to fully characterize the 3D spatial variability. This challenge can be exacerbated when the layout of sparse boreholes has irregular spacing and the concerned spatial variability shows highly non-linear non-stationary features. Even the commonly used conventional geostatistical tools can hardly tackle this situation because the associated parameters, e.g., 3D non-stationary trend function, correlation structures along three different directions, are difficult to determine in the presence of limited borehole data. This paper aims to present a novel Bayesian supervised learning (BSL) method to resolve this challenge. Characterization of 3D spatial variability is formulated as a supervised learning problem and is solved under Bayesian framework. Interpretation uncertainty is quantified automatically. The BSL is non-parametric and data-driven. It bypasses the trend function determination and correlation structures estimation processes. The performance of BSL can evolve with the borehole number. The BSL method is illustrated using a highly non-linear non-stationary and anisotropic example. The results show the BSL can properly learn the spatial variability from limited borehole data and quantify associated uncertainty. By leveraging the sparsity-enhanced framework of BSL, the computational effort in 3D is feasible for personal computer.

Keywords: spatial variability; digital geological models; sparse data; machine learning; compressed sensing.

1 Introduction

Geomaterials are affected by many spatially varying factors during complex geological formation process and exhibit spatial variability and heterogeneity in three-dimensional (3D) space (e.g., Baecher and Christian 2003). Real geotechnical structures (e.g., engineered slopes, excavation, foundation structures) always interact with geomaterials with 3D spatial variability, although classical geotechnical analyses are often simplified to 1D or 2D. Accurate characterization of the 3D spatial variability is therefore required to promote realistic geotechnical analyses considering 3D spatial variability (e.g., Xiao et al. 2018; Zhao and Wang 2021). The recent trend of digital transformation in geotechnical engineering also encourages the development of digital twins of 3D spatial variability in terms of 3D geological models (e.g., Phoon et al. 2021). However, in practice, due to the limitation of investigation borehole number and the complex nature of geomaterials properties, it is often challenging to fully characterize the 3D spatial variability (e.g., Zhao and Wang 2020&2021). Classical geostatistical tools can hardly tackle this situation because the associated geostatistical parameters, e.g., 3D trend function and correlation structures along three different directions, are difficult to determine in the presence of limited boreholes data (e.g., Wang et al. 2019&2021). In addition, the limitation of borehole data further introduces significant statistical uncertainty. How to properly evaluate the statistical uncertainty of 3D spatial variability remains difficult. These challenges can be even exacerbated when the layout of sparse boreholes has irregular spacing and the concerned spatial variability shows highly non-linear and anisotropic features.

Recently, emerging machine learning (ML) algorithms have demonstrated great potential in modelling geotechnical spatial variability (e.g., Ching et al. 2020; Hu et al. 2020; Shi and Wang 2021). Generally, ML algorithms require extensive amount of measurement data either from boreholes database or high-resolution geological images to reliably train a model, which is difficult to implement at local specific site where data available are limited. Moreover, these models can often suffer from growing computational efforts in 3D. It is therefore imperative to develop an efficient model to cater site-specific 3D spatial variability modelling from limited data. To this end, this paper presents a novel ML algorithm called Bayesian supervised learning (BSL) to characterize 3D spatial variability from limited borehole data with consideration of statistical uncertainty. In BSL, characterization of 3D spatial variability is formulated as a supervised learning problem and is solved under Bayesian framework. Interpretation uncertainty is quantified automatically. The BSL is non-parametric

and data-driven. It bypasses the complex 3D trend function determination and 3D correlation structures estimation processes. The performance of BSL can evolve with the borehole number.

After this Introduction, the proposed BSL method including mathematical formulation is introduced in Section 2. Then a highly non-linear 3D non-stationary example is provided in Section 3 to illustrate the method. Section 4 discussed the effect of measurement data amount.

2 Bayesian supervised learning of 3D spatial variability

2.1 Formulation of 3D spatial variability

How to efficiently formulate the 3D spatial variability is a remarkable question for ML in geotechnical engineering. In this study, a sparsity-enhanced mathematical formulation of 3D spatial variability is proposed. 3D spatial variability is digitized as a 3D data array $\underline{\mathbf{F}}$ with a dimension of $N_1 \times N_2 \times N_3$ and expressed as a weighted summation of 3D basis functions with the same shape (e.g., Zhao and Wang 2020; Wang et al. 2021):

$$\underline{\mathbf{F}} = \sum_{t=1}^{N} \omega_t^{3D} \underline{\mathbf{B}}_t^{3D} \tag{1}$$

in which $N = N_1 \times N_2 \times N_3$; ω_t^{3D} is the weight coefficient for 3D basis function $\underline{\mathbf{B}}_t^{3D}$. When a set of appropriate basis functions is adopted (e.g., discrete cosine transform (DCT) function), the distribution of N weight coefficients will be sparse. In other words, most weight coefficients ω_t^{3D} are negligible except a few non-trivial ones with significant magnitude. Characterization of 3D spatial variability with a dimension of $N_1 \times N_2 \times N_3$ is reduced to identification and estimation of those limited non-trivial weight coefficients. To achieve this sparsity feature, the 3D basis function $\underline{\mathbf{B}}_t^{3D}$ can be constructed appropriately using three sets of DCT basis functions. Each $\underline{\mathbf{B}}_t^{3D}$ essentially is an outer product of three 1D DCT basis functions along three different directions (e.g., Zhao and Wang 2020&2021):

$$\underline{\mathbf{B}}_t^{3D} = \underline{\mathbf{B}}_{i,j,k}^{3D} = \boldsymbol{b}_i^1 \otimes \boldsymbol{b}_j^2 \otimes \boldsymbol{b}_k^3 \tag{2}$$

in which "t" is the simplified 1D index of 3D index (i, j, k); $\underline{\mathbf{B}}_{i,j,k}^{3D}$ is the 3D basis function with frequencies components "i", "j", and "k" along three directions, respectively. $\boldsymbol{b}_i^1$, $\boldsymbol{b}_j^2$, and $\boldsymbol{b}_k^3$ are corresponding 1D basis functions, with length N_1, N_2, and N_3, respectively. The 1D basis functions can be readily constructed using built-in function in commercial software, such as "dctmtx" in MATLAB, or "dct" function in Python "scipy" library. The above settings enable the high-resolution 3D spatial variability to be sparse in the space of DCT weight coefficients. Therefore, the 3D spatial variability $\underline{\mathbf{F}}$ can be approximated if those non-trivial weight coefficients can be approximated from limited investigation data. In addition, the complex 3D trend function is automatically incorporated in the framework of Eq. (1). Detrending operation of 3D spatial variability as required in classical geostatistical tools is therefore bypassed.

2.2 Bayesian supervised learning from limited data

To account for the uncertainty associated with the 3D spatial variability interpretation based on limited data, a Bayesian framework is used to solve the non-trivial weight coefficients in Eq. (1). The posterior distribution of approximated non-trivial coefficients vector $\hat{\boldsymbol{\omega}}^{3D}$ given limited data $\boldsymbol{y}^{3D}$ is expressed as (e.g., Wang et al. 2021):

$$p\left(\hat{\boldsymbol{\omega}}^{3D} \mid \boldsymbol{y}^{3D}\right) = \frac{p\left(\boldsymbol{y}^{3D} \mid \hat{\boldsymbol{\omega}}^{3D}\right) p\left(\hat{\boldsymbol{\omega}}^{3D}\right)}{p\left(\boldsymbol{y}^{3D}\right)} \tag{3}$$

in which the $\hat{\boldsymbol{\omega}}^{3D}$ is the approximated non-trivial coefficients vector and $\boldsymbol{y}^{3D}$ is the limited data vector with length M (e.g., M<<N); $p\left(\boldsymbol{y}^{3D} \mid \hat{\boldsymbol{\omega}}^{3D}\right)$ is a Gaussian likelihood function, i.e., the plausibility of observing $\boldsymbol{y}^{3D}$ with $\hat{\boldsymbol{\omega}}^{3D}$; $p\left(\hat{\boldsymbol{\omega}}^{3D}\right)$ is a Gaussian prior distribution of $\hat{\boldsymbol{\omega}}^{3D}$; $p\left(\boldsymbol{y}^{3D}\right)$ is a normalizing constant which ensures the integration of the posterior distribution to be unity. The relationship between $\hat{\boldsymbol{\omega}}^{3D}$ and $\boldsymbol{y}^{3D}$ is formulated in accordance with the residual between them:

$$\boldsymbol{y}^{3D} = \mathbf{A}^{3D} \hat{\boldsymbol{\omega}}^{3D} + \boldsymbol{\varepsilon} \tag{4}$$

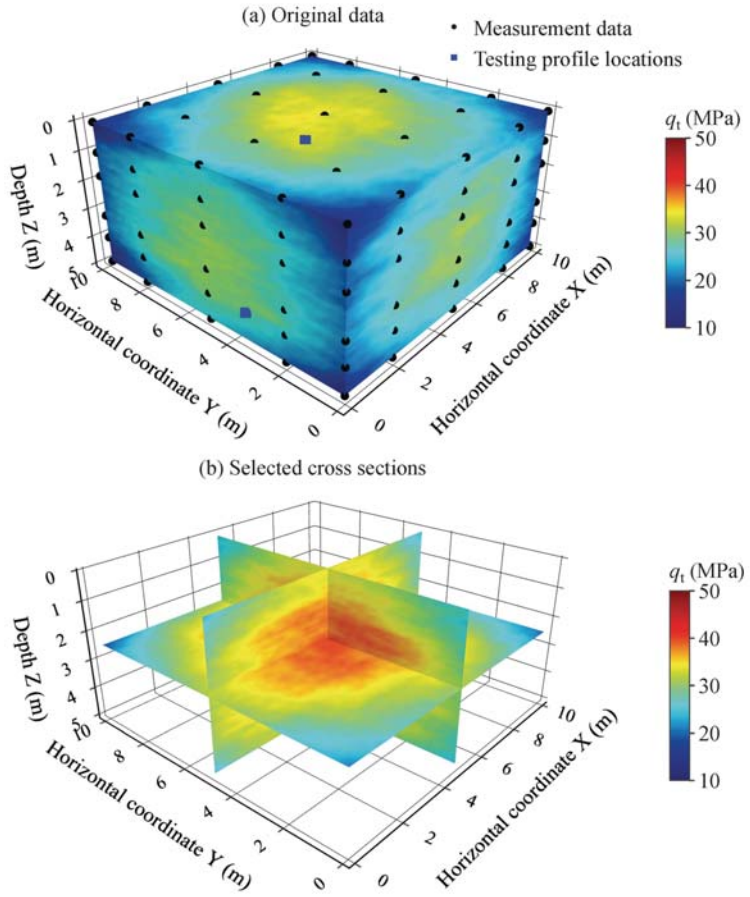

Figure 1. (a) Simulated 3D non-stationary spatial variability of q_t values; (b) selected cross sections

in which $\mathbf{A}^{3D}$ is a matrix with each column representing the reshaped sub-array of 3D basis function; $\boldsymbol{\varepsilon}$ is the error function. To learn the posterior distribution, Markov chain Monte Carlo simulation is adopted to approximate the posterior directly and efficiently. Extensive random samples of $\hat{\boldsymbol{\omega}}^{3D}$ can be simulated by iterations (e.g., Zhao and Wang 2020). After substituting the $\hat{\boldsymbol{\omega}}^{3D}$ samples to Eq. (1), random samples of the concerned 3D spatial variability can be reconstructed readily. The statistics (e.g., the average and standard deviation) of the approximated 3D spatial variability can then be calculated. The average of the samples is interpreted as the best estimate of the 3D spatial variability given limited data. The statistical uncertainty can be quantified by the standard deviation.

3 3D non-stationary example

To illustrate the application of the proposed 3D BSL framework, a set of simulated cone tip resistance data q_t of penetration test is provided in this Section. This example describes the 3D spatial variability of q_t for soils mass with a length of 10m, a width of 10m, and a thickness of 5m. The example is simulated by combining a zero-mean 3D stationary random field with a 3D trend function:

$$q_t(\mathrm{X},\mathrm{Y},\mathrm{Z}) = -0.4\times(\mathrm{X}-5)^2 - 0.4\times(\mathrm{Y}-5)^2 - (\mathrm{Z}-2.5)^2 + 40 + \xi(\mathrm{X},\mathrm{Y},\mathrm{Z})\ (\mathrm{MPa}) \tag{6}$$

in which "X" and "Y" are two horizontal coordinates both with ranges [0.1m, 10m]; "Z" is the vertical coordinate with range [0.1m, 5m]. The zero-mean 3D Gaussian random field $\xi(\mathrm{X},\mathrm{Y},\mathrm{Z})$ has unit variance and contains an exponential autocovariance structure with correlation lengths 2m, 2m, and 0.4m along three directions, respectively. The spatial resolution of this example is set to 0.1m along all directions, leading to a 100×100×50 3D array. One realization of this example is color-coded and shown in the Figure 1(a) and will be used for illustration. It shows that the q_t values are smaller around all eight corners than those in the center. This pattern can be observed in three color-coded cross-sections shown in Figure 1(b). To mimic the limited data situations, grid sampling with 6 data points along Z direction and 5 data points along both X and Y directions is implemented. In total 5×5×6=150 (e.g., a sampling ratio of 150/500000=0.03%) sampled data points, as shown

by black dots in Figure 1(a), are treated as input to BSL to learn the 3D spatial variability of q_t. Note that the sampling intervals are larger at the middle part along all three dimensions than those around the corners.

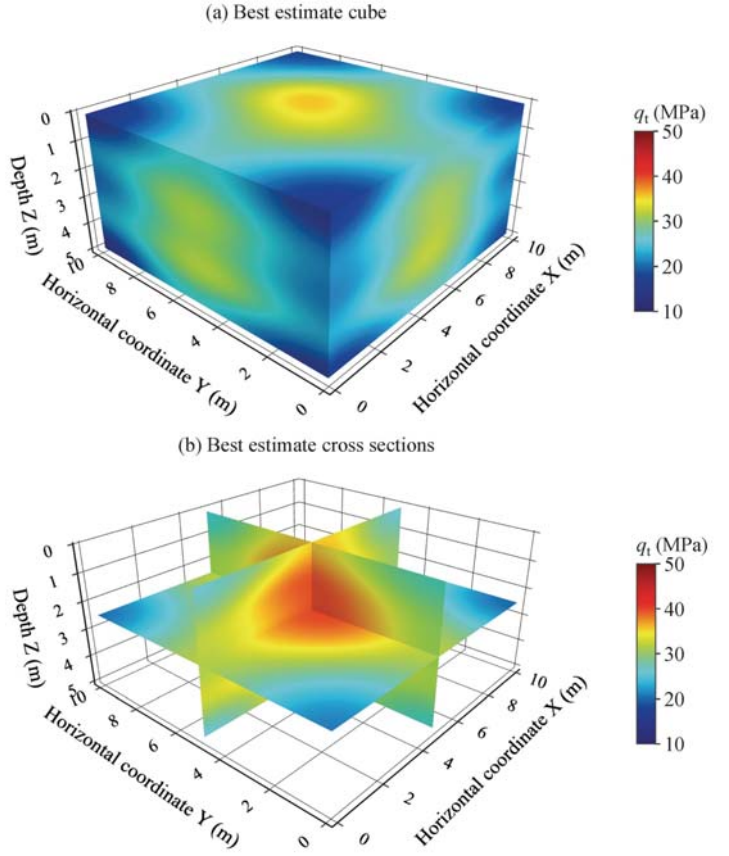

Figure 2. (a) BSL best estimate of 3D cube; (b) BSL best estimate cross sections

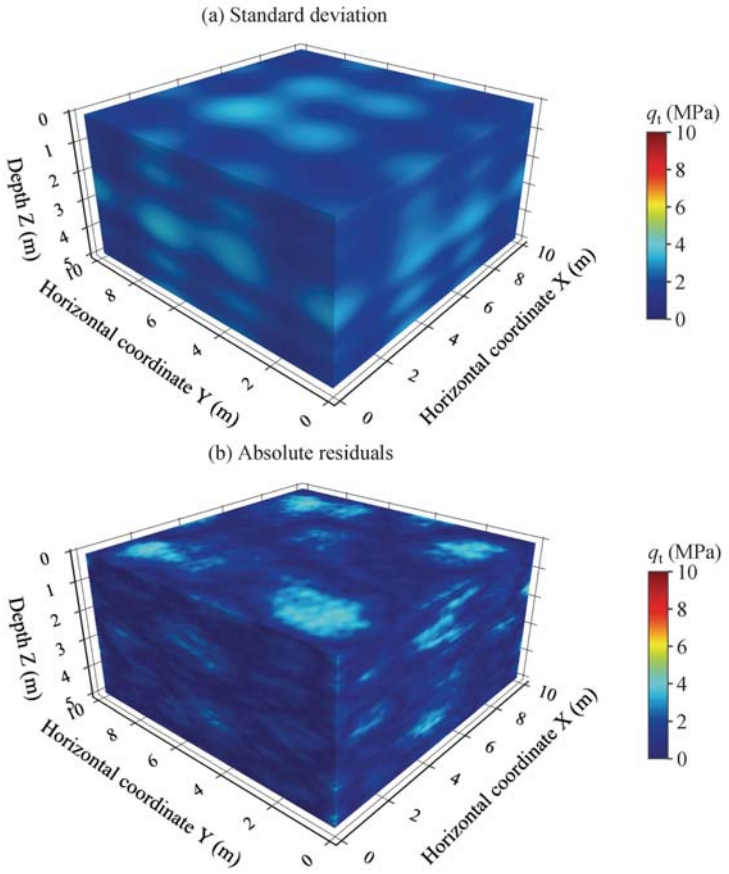

Figure 3. (a) BSL standard deviation; (b) Absolute residuals between BSL best estimate and original data

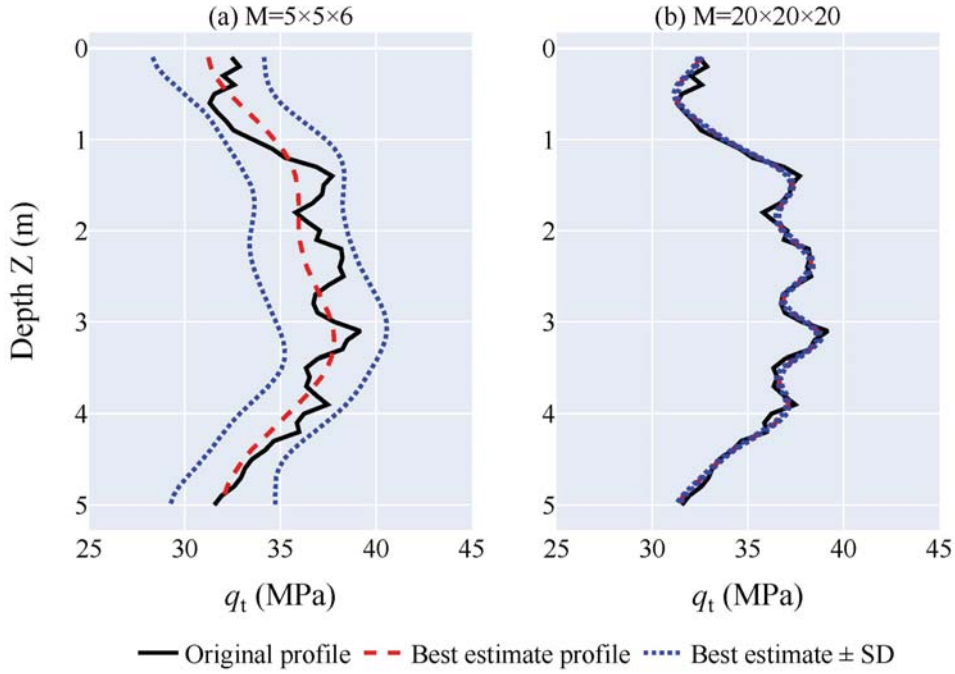

Figure 4. (a) BSL performance at selected vertical profiles (X=3m, Y=4m): (a) *M*=5×5×6; (b) *M*=20×20×20

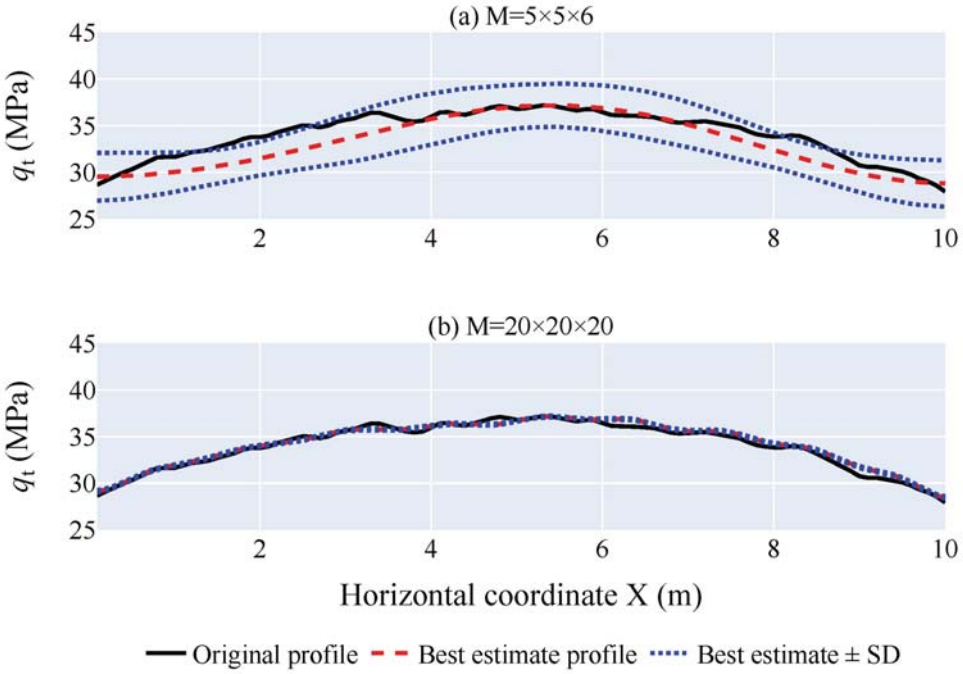

Figure 5. (a) BSL performance at selected horizontal profiles (Y=3.5m, Z=4m): (a) *M*=5×5×6; (b) *M*=20×20×20

The best estimate results are illustrated in Figure 2. Figures 2(a) and 2(b) are the best estimate cube and best estimate cross sections, respectively. Note that the best estimate result has a pattern similar to the original data in Figure 1 with the complex 3D trend function well preserved. Figures 3(a) and 3(b) compares the standard deviation cube and the absolute residual cube. Absolute residuals are generally smaller than standard deviation, suggesting the statistical uncertainty is properly quantified. To clearly examine the method, performances at two 1D testing profiles are extracted. The locations of these two unsampled profiles are shown in Figure 1(a) by blue squares. Profile at (X=3m, Y=4m) is a vertical profile, and profile at (Y=3.5m, Z=4m) is a horizontal profile. The learning results of these two profiles are shown in Figures 4(a) and 5(a), respectively. The original profile and the BSL best estimate are denoted by black solid lines, and red dashed lines, respectively. Two blue dotted lines depict the interval of best estimate ± one standard deviation. The profile comparison also suggests the non-stationary trend is well characterized without imposing an assumed trend function. The statistical uncertainty is properly quantified. The performance of the proposed 3D BSL is reasonable, even a very low sampling ratio is

adopted. The learning was performed using a PC with Intel Core i7-4790, 3.60 GHz CPU processor and 16 GB RAM, and it took about 4 seconds.

4 Sensitivity study

To investigate the effect of measurement data amount on the BSL method, an additional measurement scenario is discussed for the 3D non-stationary q_t data example. In this scenario, grid sampling with 20 data points along three directions are sampled (i.e., M=20×20×20). The sampling ratio (i.e., 8000/500000) increases to 1.6%. Due to the page limit, the comparison between 3D cubes and cross sections are not provided here. The performances at the two selected testing profiles are shown in the Figures 4(b) and 5(b), respectively. For both the vertical and horizontal profiles, the best estimate profiles converge to the original ones. The quantified uncertainty reduces to almost zeros. The results indicate that the BSL framework is data-driven and can evolve with the measurement data amount.

5 Conclusions

This paper presents a non-parametric and data-driven Bayesian supervised learning (BSL) framework for probabilistic characterization of 3D non-stationary geotechnical spatial variability from limited data. The mathematical formulation of BSL is firstly introduced. A simulated cone penetration test data example is then provided to illustrate the method. Results suggest the BSL performs well. The non-stationary 3D trend function is well characterized from limited data without imposing an assumed parametric trend function. The statistical uncertainty is also properly quantified. BSL performances are further evaluated at one vertical profile and one horizontal profile where measurement data are not available. The results also suggest the BSL performs well. Sensitivity study also shows the BSL can converge to reality when measurement data amount increases. The computational effort of BSL framework is feasible for personal computers due to its sparsity-enhanced feature.

Acknowledgments

The work described in this paper was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CityU 11213119 and C6006-20G). The financial supports are gratefully acknowledged.

References

Baecher, G.B. and Christian, J.T. (2003). Reliability and Statistics in Geotechnical Engineering, *Wiley, Chichester, U.K*.

Ching, J., Huang, W. H., & Phoon, K. K. (2020). 3D probabilistic site characterization by sparse Bayesian learning. *Journal of Engineering Mechanics*, 146(12), 04020134.

Hu, Y., Wang, Y., Zhao, T., & Phoon, K. K. (2020). Bayesian supervised learning of site-specific geotechnical spatial variability from sparse measurements. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 6(2), 04020019.

Phoon, K. K., Ching, J., & Shuku, T. (2021). Challenges in data-driven site characterization. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, https://doi.org/10.1080/17499518.2021.1896005.

Shi, C., & Wang, Y. (2021). Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. *Geoscience Frontiers*, 12(1), 339-350.

Wang, Y., Zhao, T., Hu, Y., & Phoon, K. K. (2019). Simulation of random fields with trend from sparse measurements without detrending. *Journal of Engineering Mechanics*, 145(2), 04018130.

Wang, Y., Hu, Y., & Phoon, K. K. (2021). Non-parametric modelling and simulation of spatiotemporally varying geo-data. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, https://doi.org/10.1080/17499518.2021.1971258.

Xiao, T., Li, D. Q., Cao, Z. J., & Zhang, L. M. (2018). CPT-based probabilistic characterization of three-dimensional spatial variability using MLE. *Journal of Geotechnical and Geoenvironmental Engineering*, 144(5), 04018023.

Zhao, T., & Wang, Y. (2020). Non-parametric simulation of non-stationary non-gaussian 3D random field samples directly from sparse measurements using signal decomposition and Markov Chain Monte Carlo (MCMC) simulation. *Reliability Engineering & System Safety*, 203, 107087.

Zhao, T., & Wang, Y. (2021). Statistical interpolation of spatially varying but sparsely measured 3D geo-data using compressive sensing and variational Bayesian inference. *Mathematical Geosciences*, 53(6), 1171-1199.