Bayesian Learning of Site-Specific Spatial Variability Using Sparse Geotechnical Data

Yue Hu[1], Tengyuan Zhao[1], and Yu Wang[1]

[1]Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Hong Kong SAR.
E-mail: yuehu47-c@my.cityu.edu.hk
tengyzhao2@cityu.edu.hk
yuwang@cityu.edu.hk

Abstract: Geo-materials (e.g., soils and rocks) are natural materials, and their properties are affected by many spatially varying factors during their complex geological formation process, such as the textures of their parent materials, weathering processes, transportation agents, and sedimentation conditions. Geo-materials therefore may have different properties at different locations at a specific site. This spatial variability, to some extent, are unique and deterministic in every site as an outcome of the previous geological processes that the soils and rocks at a specific site have undergone. In recent years, random field model has been a popular tool to model spatial variability of geotechnical properties. In practice, measurement data at a specific site are used to estimate random field parameters, such as mean and standard deviation, as well as parameters (e.g., scale of fluctuation) of a pre-specified parametric form of correlation function. Estimation of these random field parameters generally require extensive measurements from a specific site, which are usually not available in geotechnical engineering practice. In addition, the underlying correlation function form is often unknown for a specific site, therefore it is also challenging to select the most suitable parametric function form of the correlation function when only limited measurements are available. Moreover, even if geotechnical properties are measured at every location, the conventional random field still produces stochastic sample in each realization rather than converging to the unique and deterministic spatial variation in the specific site. To address these issues, this paper presents a Bayesian learning method for random field modeling of site-specific spatial variability of geotechnical properties, which is data-driven and non-parametric. The method is developed based on supervised machine learning, and it bypasses the estimation of parametric correlation function. The learning results are demonstrated by generating random field samples of site-specific spatial variability. Simulated data is used to illustrate the performance of the method. Comparative study on conventional random field model is also provided. The results show that the Bayesian learning method performs well even when measurements are sparse.

Keywords: site investigation; machine learning; random field; site-specific geotechnical variability.

1 Introduction

Geo-materials (e.g., soils and rocks) are natural materials, and their properties are affected by many spatially varying factors during their complex geological formation process (e.g., Baecher and Christian 2003). Geo-materials therefore may have different properties at different locations at a specific site. This spatial variability, to some extent, are unique and deterministic in every site as an outcome of the previous geological processes that the soils and rocks at a specific site have undergone (e.g., Webster 2007). In recent years, random field model is a popular tool to model spatial variability of geotechnical properties (e.g., Vanmarcke 2010). It treats geotechnical properties at different locations as a series of correlated random variables and makes statistical inference of the geotechnical properties at unmeasured locations. Usually, measurement data at a specific site are used to estimate random field parameters, such as mean and standard deviation, as well as parameters (e.g., scale of fluctuation) of a pre-specified parametric form of correlation function. Estimation of these random field parameters generally require extensive measurements from a specific site, which are usually not available in geotechnical practice. In addition, it is also challenging to select the most suitable parametric function form of the correlation function when only limited data are available. Moreover, even if geotechnical properties are measured at every location, the conventional random field still models the geotechnical spatial variability using the estimated random field parameters, producing stochastic sample in each realization, rather than converging to the unique and deterministic spatial variability in the specific site.

A compressive sampling (CS) - based Bayesian learning method is presented in this paper to characterize site-specific spatial variability from sparse measurements. This Bayesian learning method addresses the above two problems in conventional random field model. It does not require estimation of parametric form of correlation function. Moreover, the interpreted results converge to the site-specific variability through learning process when the number of measurements increases. A set of simulated geotechnical data is adopted to illustrate and validate the CS-based Bayesian learning method. In addition, comparative study on conventional random field is also provided.

Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR)
Editors: Jianye Ching, Dian-Qing Li and Jie Zhang

Published by Research Publishing, Singapore.
ISBN: 978-981-11-2725-0; doi:10.3850/978-981-11-2725-0_MS2-3-cd

2 Compressive Sampling (CS)–Based Bayesian Learning

2.1 Supervised machine learning

Machine learning is to program computers to optimize a model using example data or past experiences for making prediction or gaining knowledge from the data (e.g., Alpaydin 2010). Generally speaking, machine learning can be divided into two categories: supervised learning and unsupervised learning. Supervised learning aims to produce a model function to map the data after training of data pairs. Regression can be considered as a classical example of supervised learning, and it aims to find a model function to represent the relationship between/among variables given available data pairs. From a machine learning point of view, the regression model can be considered as weighted summation of a series of basis functions, which can be expressed as (e.g., Tipping 2001):

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^{\mathrm{N}} \omega_i \boldsymbol{B}_i(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{f}(\boldsymbol{x})$ are input index and target vector respectively, both with length of N; $\boldsymbol{B}_i(\boldsymbol{x})$ (i = 1, 2, …, N) represents fixed basis function (e.g., wavelet function) also with length N; ω_i denotes the weight corresponding to $\boldsymbol{B}_i(\boldsymbol{x})$. Note that ω_i are the adjustable parameters of the model, and different ω_i leads to different models. For a specific problem, ω_i is unknown and needs to be learned by some supervised learning algorithms from the measured data pairs, such as M data pairs (x_m, $f(x_m)$) (m = 1, 2, 3, …, M, and M<<N) measured from the target vector. One of the popular supervised learning methods is sparse Bayesian learning (SBL), which can effectively learn the parameters (i.e., ω_i) in Eq. (1) using sparse measurements. Note that SBL is mathematically very similar to compressive sensing/sampling (CS) in signal processing (e.g., Williams et al. 2005). The following subsection introduces Bayesian compressive sampling (BCS) under a supervised learning framework.

2.2 CS-based Bayesian learning

Bayesian compressive sampling/compressed sensing (BCS) is a probabilistic extension of CS to reconstruct a complete signal from partial measurements on that signal (e.g., Ji et al. 2008; Wang and Zhao 2016; Wang and Zhao 2017; Zhao et al. 2018a). A signal herein is defined as a quantity which varies with time or space, e.g., geotechnical data profile (e.g., Zhao et al. 2018b). The signal is denoted by a real-valued vector $\boldsymbol{f}$, which is equivalent to the complete target data profile (i.e., $\boldsymbol{f}(\boldsymbol{x})$ in Eq. (1)) under a supervised learning framework. Sparse measurements on $\boldsymbol{f}$ are expressed as a column vector $\boldsymbol{y}$ which has a length of M (M<N). Under a supervised learning framework, $\boldsymbol{y}$ versus its measurement locations are taken as the training data set. The relationship between the training data and weights is expressed as:

$$\boldsymbol{y} = \boldsymbol{\Psi f} + \varepsilon \boldsymbol{I} = \boldsymbol{\Psi B \omega} + \varepsilon \boldsymbol{I} = \mathbf{A}\boldsymbol{\omega} + \varepsilon \boldsymbol{I} \tag{2}$$

where $\boldsymbol{\Psi}$ is a measurement matrix with dimension M×N, reflecting the location of measurement data $\boldsymbol{y}$ in $\boldsymbol{f}$; $\mathbf{B}$ is an N×N basis function matrix with N orthonormal columns as basis functions (e.g., discrete wavelet functions), i.e., the $\boldsymbol{B}_i(\boldsymbol{x})$ in Eq. (1) in a matrix form; $\boldsymbol{\omega}$ represents unknown weight ω_i (i = 1, 2, 3,…, N) in Eq. (1) in terms of a column vector with a length of N; $\boldsymbol{I}$ is an M-length column vector with elements being unity; ε is measurement error, which is modeled as a Gaussian random variable with a zero mean and unknown variance. Once the weight vector $\boldsymbol{\omega}$ is learned, the regression model is obtained. Then, data at unmeasured locations can be predicted using the regression model obtained. Because of compressibility of natural signal, there are only S (S<<N) non-trivial components with significant magnitudes in $\boldsymbol{\omega}$. The non-trivial weight vector is denoted as $\boldsymbol{\omega_s}$. It has been shown that the weight vector $\boldsymbol{\omega_s}$ follows a multivariate Student's t distribution, with a degree of freedom of $2c_n$, a mean of $\boldsymbol{\mu}_{\omega_s}$ and covariance matrix of $\mathbf{COV}_{\omega_s}$, which are expressed as (Wang and Zhao 2017):

$$\begin{aligned} \boldsymbol{\mu}_{\omega_s} &= \mathbf{H}\mathbf{A}^{\mathrm{T}}\boldsymbol{y} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{D}\right)^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{y} \\ \mathbf{COV}_{\omega_s} &= \frac{d_n \mathbf{H}}{c_n - 1} = \frac{d_n \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{D}\right)^{-1}}{c_n - 1} \end{aligned} \tag{3}$$

in which $\mathbf{H} = (\mathbf{A}^{\mathrm{T}}\mathbf{A}+\mathbf{D})^{-1}$; c_n = M/2+c; $d_n = d+(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}- \boldsymbol{\mu}_{\omega_s}{}^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\mu}_{\omega_s})/2$; c and d are small non-negative constants (e.g., $c = d = 10^{-4}$); $\mathbf{D}$ = a diagonal matrix with diagonal elements being $D_{i,i} = \alpha_i$ (i = 1, 2, …, N); α_i (i = 1, 2, …, N) are unknown non-negative variables which can be determined by maximizing the marginal likelihood of $\boldsymbol{y}$ (e.g., Tipping 2001; Ji et al. 2008; Wang and Zhao 2017). Using the learned $\boldsymbol{\omega_s}$ above and Eq. (1), the target vector $\boldsymbol{f}$ may be approximated as $\hat{\boldsymbol{f}}$, with its mean $\boldsymbol{\mu}_{\hat{f}}$ and covariance matrix $\mathbf{COV}_{\hat{f}}$, expressed as:

$$\boldsymbol{\mu}_{\hat{f}} = E(\hat{\boldsymbol{f}}) = E(\mathbf{B}\boldsymbol{\omega}_s) = \mathbf{B}E(\boldsymbol{\omega}_s) = \mathbf{B}\boldsymbol{\mu}_{\omega_s}$$
$$\mathbf{COV}_{\hat{f}} = E[(\hat{\boldsymbol{f}} - \boldsymbol{\mu}_{\hat{f}})(\hat{\boldsymbol{f}} - \boldsymbol{\mu}_{\hat{f}})^{\mathrm{T}}] = \mathbf{B}\mathbf{COV}_{\omega_s}\mathbf{B}^{\mathrm{T}} \tag{4}$$

The mean vector $\boldsymbol{\mu}_{\hat{f}}$ reflects the expectation of the complete data profile; while the covariance matrix objectively incorporates estimation errors and correlations. The results learned from sparse measurements can be further expressed by random field samples through Karhunen–Loève (KL) expansion (e.g., Wang et al. 2018):

$$\hat{\boldsymbol{f}} = \boldsymbol{\mu}_{\hat{f}} + \mathbf{V}\sqrt{\mathbf{COV}_{\hat{f},d}}\boldsymbol{Z} \tag{5}$$

where $\mathbf{V}$ is the eigenvector matrix of $\mathbf{COV}_{\hat{f}}$; $\mathbf{COV}_{\hat{f},d}$ is a diagonal matrix that records the eigenvalues of $\mathbf{COV}_{\hat{f}}$; $\boldsymbol{Z}$ is a set of uncorrelated standard random variables, with zero-mean and unit-variance. For implementation, $\boldsymbol{Z}$ is often taken as a random vector with uncorrelated standard Gaussian numbers. According to Eq. (5), random field sample (RFS) of $\hat{\boldsymbol{f}}$ can be obtained readily by realization of uncorrelated random vector $\boldsymbol{Z}$. In this study, the Bayesian learning results are illustrated through the generated RFS.

A schematic of the proposed Bayesian learning method is shown in the Figure 1. An example of geotechnical data (e.g., the X) along depth is used as an illustration. When limited measurements are taken (i.e., the open circles), BCS can provide a best estimate of complete data X (e.g., the grey scale column), and a covariance matrix (e.g., the grey scale map). Thereafter, RFS (e.g., three grey lines) can be readily generated through KL expansion using $\boldsymbol{\mu}_{\hat{f}}$ and $\mathbf{COV}_{\hat{f}}$.

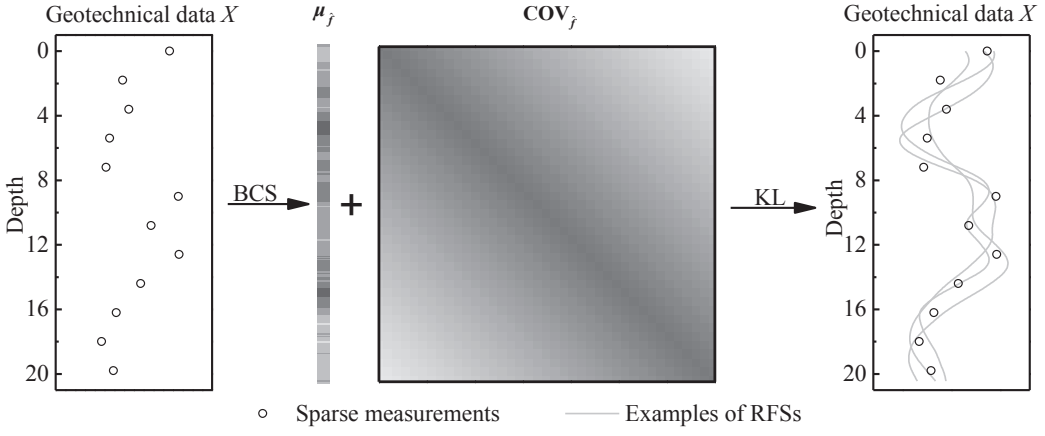
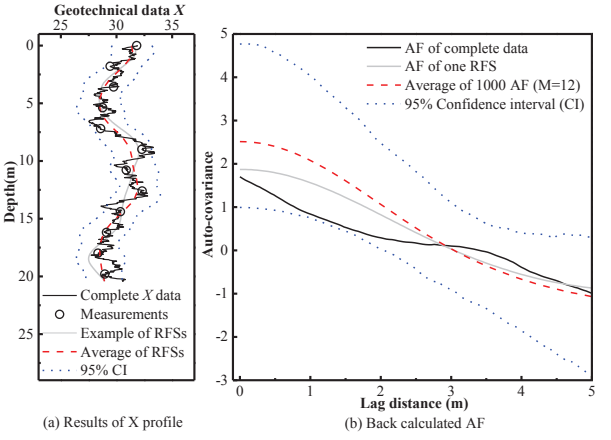


Figure 1. Schematic of the Bayesian learning method.

3 Simulated Geotechnical Data Example

In this section, a set of simulated geotechnical data is adopted to illustrate the Bayesian learning method in characterizing site-specific spatial variability. In this example, a geotechnical data X profile within a 20.44m-thick soil layer at a specific site is simulated with a vertical resolution of 0.04m along depth, as shown in Figure 2(a) by a black solid line. This profile has 512 data points, and it represents spatial variability of X at a specific site.

In this example, the X profile is simulated from a stationary Gaussian random field with constant mean value of 30 and standard deviation (SD) of 2. In addition, a cosine exponential auto-correlation function (CSX) is adopted and the scale of fluctuation is taken as 4m. To investigate the performances of the Bayesian learning method, as well as the conventional random field models, limited points (e.g., M=12 points shown in Figure 2(a)) are chosen as measurement input. The results are demonstrated in the following subsections.



(a) Results of X profile (b) Back calculated AF

Figure 2. Results of Bayesian learning from sparse measurements.

3.1 Results of CS-based Bayesian learning

Following the procedure described in Figure 1, 1000 RFS of X profile can be generated to illustrate the Bayesian learning results. An example of RFS of X is shown in Figure 2(a) by a gray line. Note that the RFS properly reflects the general spatial variation of the underlying site-specific data. Pointwise statistics (e.g., mean and SD) at each depth are calculated and shown in the Figure 2(a). The dashed line indicates the averaged profile of 1000 RFS. The two dotted lines correspond to the 2.5th and 97.5th percentiles of 1000 RFS values at each depth. It can be noted that the averaged profile follows a consistent trend with the measurement data, and almost all local variations of the black solid line fall within the 95% CI. This implies that the spatial variation of X is rationally characterized by RFS generated from sparse measurements.

The spatial auto-correlation embedded in the X profiles can be quantified by the auto-covariance function (AF), which is calculated using Eq. (6):

$$C(\delta) = \frac{1}{\mathrm{N}-\delta}\sum_{i=1}^{\mathrm{N}-\delta}[(\hat{f}_i - \mu)(\hat{f}_{i+\delta} - \mu)] \tag{6}$$

where $C(\delta)$ represents the estimator of AF at δ; δ is the lag distance between the i-th and the $(i + \delta)$-th element; and μ represents the mean of $\hat{\boldsymbol{f}}$. AF of the underlying X profile is shown in Figure 2(b) by a solid line. The AF for each of the simulated 1000 RFS can be calculated using Eq. (6). This leads to 1000 AF. Then the average and 95% CI of 1000 AF can be obtained at each lag distance, as shown in Figure 2(b) by a dashed line and a pair of dotted lines, respectively. AF of the RFS example in Figure 2(a) is also plotted as grey line in Figure 2(b). The results show that both the gray and dashed lines decrease as the lag distance increases, following a pattern similar to that of the black solid line. Moreover, all local variations of the solid lines fall within the 95% CI. These agreements indicate that the spatial auto-correlation embedded in the complete X profile, together with quantitative assessment of uncertainty, is properly estimated by 1000 RFS generated from sparse measurements.

3.2 Comparative study on conventional random field model

This subsection investigates conventional random field model using the same measurements input used in the previous subsection. The frequently used random field model requires the random field parameters (e.g., mean, variance, scale of fluctuation) to be determined before RFS generation.

Usually, experimental auto-correlation is estimated using measurements after de-trending (Wang et al. 2019). The experimental AF at three lag distances are obtained from Eq. (6) and shown in Figure 3(a) by open circles. In this case, the underlying AF function form i.e., CSX is assumed to be known already to best fit the experimental AF. The fitting AF and associated parameters are shown by red dash line in Figure 3(a). After that, covariance matrix can be constructed based on those parameters. Subsequently, RFS can be generated using Cholesky decomposition. To be consistent with the previous subsection, 1000 RFS are simulated, using which the average of RFS and 95% CI at each depth are calculated and shown in Figure 3(b) by a dashed line and two dotted lines, respectively. It is shown that the dashed line (i.e., the average of RFS) is almost a constant, failing to incorporate site-specific spatial variability. The results indicate that the RFS obtained from conventional random field modeling might not be representative of the spatial variability of X in this specific site, even when parametric form of AF is known already.

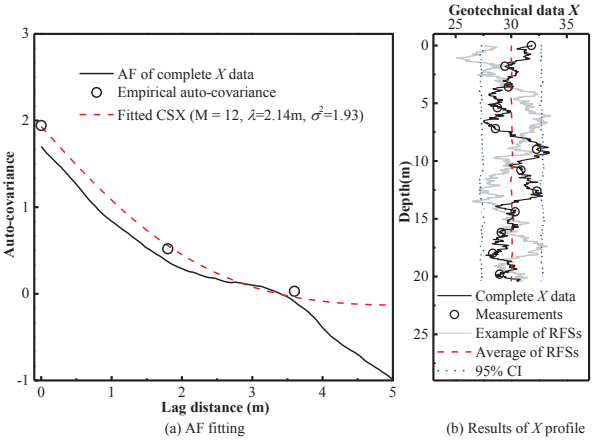


Figure 3. Results of conventional random field model from sparse measurements.

4 Effect of Number of Measurement Points

4.1 Convergence of CS-based Bayesian learning method

To further explore the effect of number of measurement points M, three more measurement scenarios, i.e., M = 25, M = 64, M = 512 are investigated. For each scenario, evenly-spaced M measurements are set as input to the CS-based Bayesian learning method. Following the same procedures above, 1000 RFS of X are generated for each M scenario. Figure 4 summarizes the results for four M scenarios.

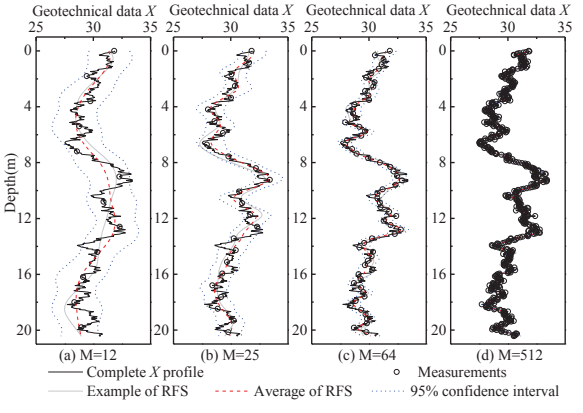


Figure 4. Effect of number of measurement points on the CS-based Bayesian learning method.

Figure 4 shows that as M increases, the average of 1000 sets of RFS (i.e., red dash lines) becomes increasingly similar to the complete X profile. In addition, the 95% CI gradually reduces and approaches to zero effectively. When M=512, the generated RFS converge to the complete X profile and become deterministic.

4.2 Non-convergence of conventional random field model

The effect of number of measurement points M is also investigated for the conventional random field models. The same measurement scenarios in the previous subsection are applied. Following the same procedure stated in the previous section, 1000 RFS are simulated for each M scenario. The results are plotted in Figure 5.

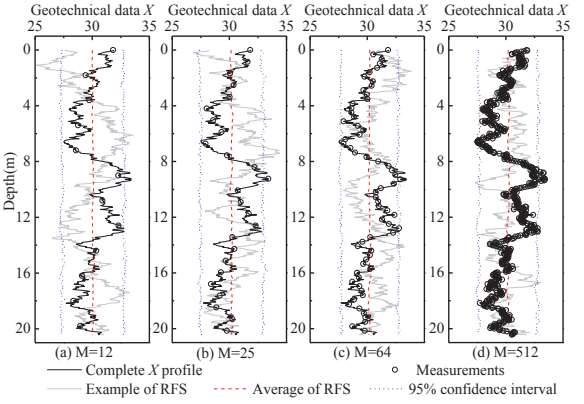


Figure 5. Effect of number of measurement points on conventional random field model.

Figure 5 shows that the RFS obtained from conventional random field model are not sensitive to the measurement quantity. The averaged profiles remain more or less constant, despite of the increase of measurement point number. When measurement at every location is available, it still produces stochastic RFS rather than deterministic spatial variability.

5 Conclusions

This paper presents a compressive sensing-based Bayesian learning method for characterizing site-specific spatial variability from sparse measurements. A numerical example is provided to investigate the effectiveness of the method. Performance of conventional random field models is also compared in the paper. The results show that the method performs well in characterizing site-specific spatial variability. Moreover, as the number of measurements points increases, the underlying spatial variability at a specific site is learned progressively through the method.

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 No. 9042331 (CityU 11225216) and Project No. 8779012 (T22-603/15N)). The financial support is gratefully acknowledged.

References

Alpaydin, E. (2010). *Introduction to Machine Learning*, MIT press.

Baecher, G.B. and Christian, J.T. (2003). *Reliability and Statistics in Geotechnical Engineering*, John Wiley & Sons, Hoboken, New Jersey, 228-239.

Ji, S., Xue, Y., and Carin, L. (2008). Bayesian compressive sensing. *IEEE Trans. Signal Process.*, 56(6), 2346-2356.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1, 211-244.

Vanmarcke, E. (2010). *Random Fields: Analysis and Synthesis*, World Scientific.

Wang, Y. and Zhao, T. (2016). Interpretation of soil property profile from limited measurement data: A compressive sampling perspective. *Can. Geotech. J.*, 53(9), 1547-1559.

Wang, Y. and Zhao, T. (2017). Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique*, 67(6), 523-536.

Wang, Y., Zhao, T., and Phoon, K.K. (2018). Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. *Can. Geotech. J.*, 55(6), 862-880.

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

Zhao, T., Hu, Y., and Wang, Y. (2018a). Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. *Engineering Geology*, 246, 162-175.

Zhao, T., Montoya-Noguera, S., Phoon, K.-K., and Wang, Y. (2018b). Interpolating spatially varying soil property values from sparse data for facilitating characteristic value selection. *Canadian Geotechnical Journal*, 55(2), 171-181.

Webster, R. and Oliver, M.A. (2007). *Geostatistics for Environmental Scientists. 2nd Edition*, John Wiley & Sons, Hoboken, New York, 50-53.

Williams, O., Blake, A., and Cipolla, R. (2005). Sparse Bayesian learning for efficient visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1292-1304.