Flexible Modeling of Trend for Geotechnical Spatial Variability by Gaussian Process Regression

Jianye Ching
Professor, Dept. of Civil Engineering, National Taiwan University, Taipei, Taiwan

Ikumasa Yoshida
Professor, Department of Urban and Civil Engineering, Tokyo City University, Tokyo, Japan

ABSTRACT: The spatial variation of a soil parameter is usually modeled as the summation of spatial trend and spatial variability. Traditionally, the trend is assumed deterministic and is fitted by spatial data. However, previous studies showed that it is more reasonable to model the trend as uncertain rather than deterministic. This paper proposes a probabilistic trend model that is based on the Gaussian process regression (GPR). In this model, the spatial trend is represented as a stationary normal random field with an auto-correlation function modeled by the squared exponential (QExp) model, which produces smooth random field realizations in order to mimic the trend. In contrast, the spatial variability is modeled as a stationary normal random field modeled by the Whittle-Matern (WM) model. Numerical and real examples are adopted to illustrate the effectiveness of this GPR method.

The spatial variation of a soil parameter [e.g., cone penetration test (CPT) data] is usually modeled as the sum of the spatial trend (t) and spatial variability (ε) (Phoon and Kulhawy 1999). The spatial variability (ε) is typically modeled as a stationary random field (Vanmarcke 1977). In the geotechnical literature, the probabilistic modeling of the spatial trend (t) is not addressed until recently. In fact, the trend is customarily regarded as deterministic (e.g. Fenton 1999; Phoon et al. 2003). Ching and Phoon (2017) developed a method of probabilistically modeling the spatial trend based on the sparse Bayesian learning (SBL) approach (Tipping 2001). In this SBL method, the spatial trend is represented as the weighted sum of sparse basis functions (BFs). The weights are zero-mean normal random variables with variances (s_1^2, …, s_m^2) [many BFs are de-activated after screening, and there remains m active BFs (m is usually small)], i.e., the SBL method models the spatial trend as a zero-mean normal random field whose auto-covariance function is governed by (s_1^2, …, s_m^2). Ching et al. (2020, 2021) further modified the original SBL method by deriving Kronecker-product formulas to resolve the issue of high computational cost for three-dimensional (3D) problems.

Recently, Yoshida et al. (2021) adopted a different strategy of probabilistically modeling the spatial trend. The spatial trend is still modeled as a zero-mean normal random field (same as the SBL method), but the kernel strategy (Rasmussen and Williams 2006) is adopted to directly model the auto-correlation function of the trend as a kernel function. This strategy of modeling the spatial trend is denoted as the Gaussian process regression (GPR) in the current paper. Ching et al. (2023) further extended the GPR method to 3D problems by adopting Kronecker-product algorithms to minimize computational cost. The purpose of the current paper is to introduce the GPR method and also to demonstrate its use on numerical and real examples.

1. MODEL FOR SPATIALLY VARIABLE SOIL PARAMETER

Let $\boldsymbol{\xi} \in \mathbf{R}^{n\times 1}$ contain the soil parameters at n positions ($\mathbf{p}_1$, $\mathbf{p}_2$, …, $\mathbf{p}_n$) in space, where $\mathbf{p}_i = (\mathbf{h}_i, z_i)$ contains the horizontal coordinates $\mathbf{h}_i = (x_i, y_i)$ and vertical coordinate z_i. $\boldsymbol{\xi}$ is modeled as the

summation of the spatial trend **t** and a zero-mean spatial variability **ε**:

$$\boldsymbol{\xi} = \mathbf{t} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{t} \in \mathbf{R}^{n\times 1}$ and $\boldsymbol{\varepsilon} \in \mathbf{R}^{n\times 1}$ contain the trend and spatial variability, respectively.

1.1. Model for spatial variability

Let us denote the auto-covariance matrix for **ε** by $\boldsymbol{\Sigma} \in \mathbf{R}^{n\times n}$. In the current paper, the horizontal and vertical auto-correlations are assumed separable:

$$\boldsymbol{\Sigma} = \sigma^2 \times \begin{bmatrix} 1 & \cdots & \rho\left(\|\mathbf{h}_1 - \mathbf{h}_n\|\right)\rho\left(|z_1 - z_n|\right) \\ \vdots & \ddots & \vdots \\ \rho\left(\|\mathbf{h}_n - \mathbf{h}_1\|\right)\rho\left(|z_n - z_1|\right) & \cdots & 1 \end{bmatrix} \tag{2}$$

where σ is the standard deviation of the spatial variability; ρ(Δ) denotes the auto-correlation function for a separation distance of Δ. The Whittle-Matérn (WM) model (Guttorp and Gneiting 2006; Liu et al. 2017; Wang et al. 2018; Ching et al. 2019) is adopted to model both horizontal and vertical auto-correlations of **ε**:

$$\rho(\Delta) = \frac{2}{\Gamma(\nu)} \cdot \left(\frac{\sqrt{\pi} \cdot \Gamma(\nu + 0.5) \cdot |\Delta|}{\Gamma(\nu) \cdot \delta} \right)^{\nu} \cdot K_{\nu}\left(\frac{2\sqrt{\pi} \cdot \Gamma(\nu + 0.5) \cdot |\Delta|}{\Gamma(\nu) \cdot \delta} \right) \tag{3}$$

where ν is the smoothness parameter; δ is the scale of fluctuation (SOF); Γ is the Gamma function; and K_ν is the modified Bessel function of the second kind with order ν. There are two sets of scale of fluctuation (SOF) and smoothness: the horizontal ones are denoted by (δ_h, ν_h), and the vertical ones are denoted by (δ_z, ν_z). As a result, the auto-covariance parameters of the spatial variability (denoted by **θ**) include (σ^2, δ_h, δ_z, ν_h, ν_z).

1.2. GPR model for trend

The Gaussian Process Regression (GPR) method treats the trend (**t**) as random and models it as a zero-mean normal random field. Let us denote the auto-covariance matrix for **t** by $\boldsymbol{\Sigma}_\mathbf{t} \in \mathbf{R}^{n\times n}$, which depends on the auto-covariance parameters of the spatial trend, denoted by $\boldsymbol{\theta}_t$. As a result, **ξ** is zero-mean with auto-covariance matrix equal to $\boldsymbol{\Sigma}_\mathbf{t} + \boldsymbol{\Sigma}$:

$$f\left(\boldsymbol{\xi} \mid \boldsymbol{\theta}_t, \boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi}^n} \times \frac{1}{\sqrt{\left|\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}\right|}} \times \exp\left[-\frac{1}{2} \boldsymbol{\xi}^T \left(\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\xi} \right] \tag{4}$$

For the GPR method developed by Yoshida et al. (2021), the trend is modeled as a stationary zero-mean normal random field with auto-correlation defined by a Gaussian kernel function $k(\mathbf{p}_i, \mathbf{p}_j)$. In 3D, the Gaussian kernel can be written as

$$k\left(\mathbf{p}_i, \mathbf{p}_j\right) = \exp\left(-\pi \frac{\left\|\mathbf{h}_i - \mathbf{h}_j\right\|^2}{\delta_{t,h}^2} - \pi \frac{\left(z_i - z_j\right)^2}{\delta_{t,z}^2} \right) \tag{5}$$

where $\delta_{t,h}$ and $\delta_{t,z}$ are the scales of fluctuation (SOFs) of the spatial trend in the horizontal and vertical directions, respectively. For the GPR model, the trend auto-covariance matrix $\boldsymbol{\Sigma}_\mathbf{t} \in \mathbf{R}^{n\times n}$ is governed by $\boldsymbol{\theta}_t = (\sigma_t^2, \delta_{t,h}, \delta_{t,z})$:

$$\boldsymbol{\Sigma}_t = \sigma_t^2 \times \mathbf{K} = \sigma_t^2 \times \begin{bmatrix} k\left(\mathbf{p}_1, \mathbf{p}_1\right) & \cdots & k\left(\mathbf{p}_1, \mathbf{p}_n\right) \\ \vdots & \ddots & \vdots \\ k\left(\mathbf{p}_n, \mathbf{p}_1\right) & \cdots & k\left(\mathbf{p}_n, \mathbf{p}_n\right) \end{bmatrix} \tag{6}$$

where $\mathbf{K} \in \mathbf{R}^{n\times n}$ is the kernel matrix. Ching et al. (2023) further extended the GPR method to 3D problems by adopting Kronecker-product algorithms to minimize computational cost. These 3D algorithms are adopted in the current paper.

1.3. Bayesian analysis and conditional random field simulation

The GPR method developed by Ching et al. (2023) is Bayesian. For the Bayesian analysis, prior probability density functions (PDFs) for $\boldsymbol{\theta}_t$ and **θ** are required. In the examples presented later, the logarithms of $\boldsymbol{\theta}_t$ and **θ** follow uniform prior PDFs with upper and lower bounds. Conditioning on the observed data (**D**), the transitional Markov chain

Monte Carlo (TMCMC) algorithm (Ching and Chen 2007) is used to draw posterior samples of $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}$. Then, the conditional trend and random field samples (conditioning on **D**) can be further simulated.

2. SIMULATED EXAMPLE

Consider the following 1D example: $\xi(z) = t(z) + \varepsilon(z)$, where $t(z) = 100 + 100\times z - 1000\times z^2 + 1000\times z^3$, the spatial variability $\varepsilon(z)$ is zero-mean, with standard deviation $\sigma = 20$ and SOF (δ) = 0.1, and simulated by a single exponential auto-correlation model (Vanmarcke 1977). The $\boldsymbol{\xi}$ data are obtained at equally spacing depths (z_1, z_2, …, z_n) = (0, 0.01, 0.02, …, 1.0) (n = 101). Once the $\boldsymbol{\xi}$ data are simulated (the simulated data are shown in Figure 1a as the dark line), the actual values of the auto-correlation parameters $\boldsymbol{\theta}$ and trend are treated as unknown. The simulated $\boldsymbol{\xi}$ data are analyzed by the GPR method to draw posterior samples of $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}$ as well as to simulate the conditional trend and random field samples. The upper and lower bounds of the uniform prior PDF for $\boldsymbol{\theta}$ are conservatively chosen so that they are away from the actual values of $\boldsymbol{\theta}$. The prior PDF for $\ln(\sigma_t)$ is chosen to be uniform over ln[(sample mean of ξ)/3] to ln[(sample mean of ξ)×3]. The prior PDF for $\ln(\delta_t)$ is chosen to be uniform over ln(0.1) to ln(10).

The TMCMC analysis results are shown in Figure 1. The red solid and dashed lines in Figure 1a show the posterior median and 95% confidence interval (CI) of the trend, respectively. Figure 1b shows the posterior samples of $\boldsymbol{\theta}_t$, whereas Figures 1c,d show the posterior samples of $\boldsymbol{\theta}$. The posterior samples of trend and $\boldsymbol{\theta}$ exhibit some consistency in the sense that the 95% CI or sample clouds mostly cover their actual values.

3. REAL CASE STUDIES

3.1. 2D example – South Parklands, Adelaide (South Australia)

This case study is a test site at South Parklands, Adelaide (South Australia) with more than two hundreds CPT soundings performed in a stiff, overconsolidated clay known as Keswick Clay (Jaksa 1995; Jaksa et al. 1999) to the depth of about 5 m. Among the soundings, 51 soundings were conducted along a line on the ground surface with a horizontal interval of 0.5 m. The 51 soundings spanning a horizontal extent (x direction) of 25 m. The problem is 2D because the y coordinate is fixed. The cone tip resistance data between z = 1.5 m and z = 5 m (sampling depth interval = 0.02 m) for the 51 soundings are analyzed. Figure 2 shows the actual cone tip resistance data on the (x,z) plane. The grey lines in the figure are the actual data. The surface plot in the figure is obtained by interpolation in the x direction. All data of the 51 soundings are analyzed together as a 2D problem. Uniform prior PDFs are adopted for the logarithms of ($\boldsymbol{\theta}_t$, $\boldsymbol{\theta}$) [e.g., $\ln(\delta)$ is uniform over ln(0.01 m) to ln(1 m), $\ln(\delta_{t,z})$ is uniform over ln(0.1 m) to ln(10 m), and $\ln(\delta_{t,h})$ is uniform over ln(0.1 m) to ln(100 m)]. Figure 3 shows a posterior sample of the trend, and Figure 4 shows the posterior samples of $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}$.

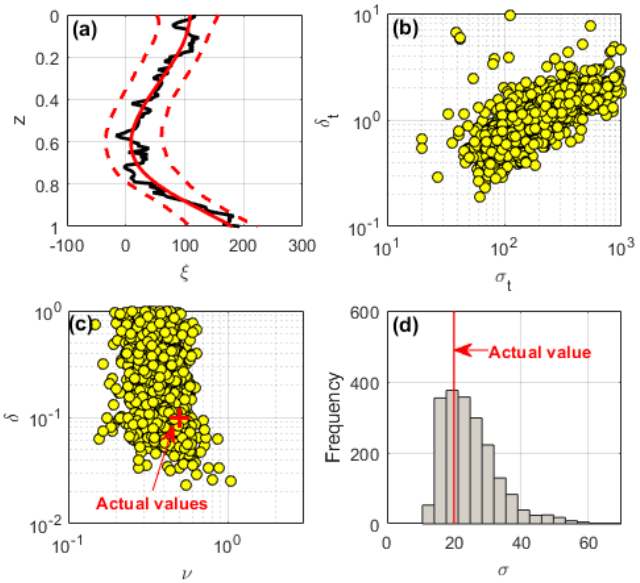

Figure 1: TMCMC results for the 1D simulated example: (a) posterior median and 95% CI of the trend; (b) posterior samples of $\boldsymbol{\theta}_t$; (c,d) posterior samples of $\boldsymbol{\theta}$.

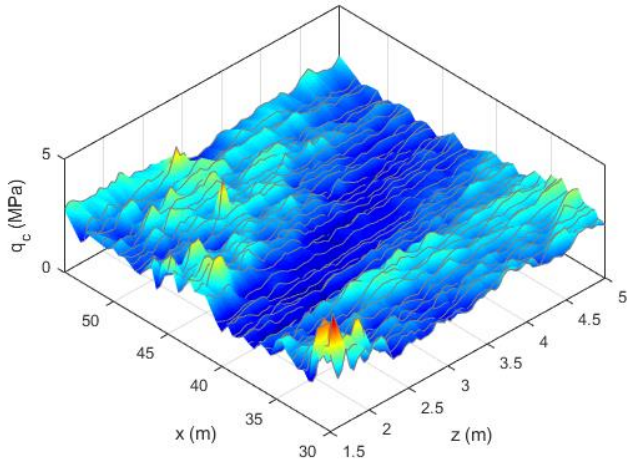

Figure 2: Actual cone tip resistance data shown on the (x,z) plane (South Parklands) (Ching et al. 2020).

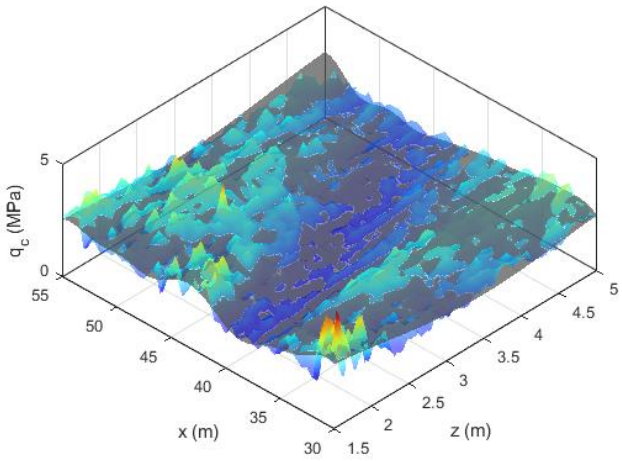

Figure 3: Posterior sample (grey surface) of the trend.

3.2. 3D example – Hollywood, South Carolina (United States)

This case study is a test site in Hollywood, South Carolina (Stuedlein et al. 2016) with multiple cone penetration test (CPT) soundings conducted in an area of 26 m × 3 m = 78 m^2. The top view of the test site is shown in Figure 5. There are in total five CPT clusters, and each cluster contains 5 CPTs (triangles in the figure indicates the horizontal locations of the CPTs). The CPT data are shown in Figure 6. The sampling depth interval is dz = 0.02 m. The q_t data of the 25 soundings in the 6 m thick layer (from 5 m to 11 m; dark lines in Figure 6) of loose to medium dense sand are analyzed. Uniform prior PDFs are adopted for the logarithms of (θ_t, θ). Figure 7 shows the 95% CI for the posterior samples of the trend for sounding P1-1 (the results for other soundings are not shown for brevity). Figure 8 shows the posterior samples of θ_t and θ.

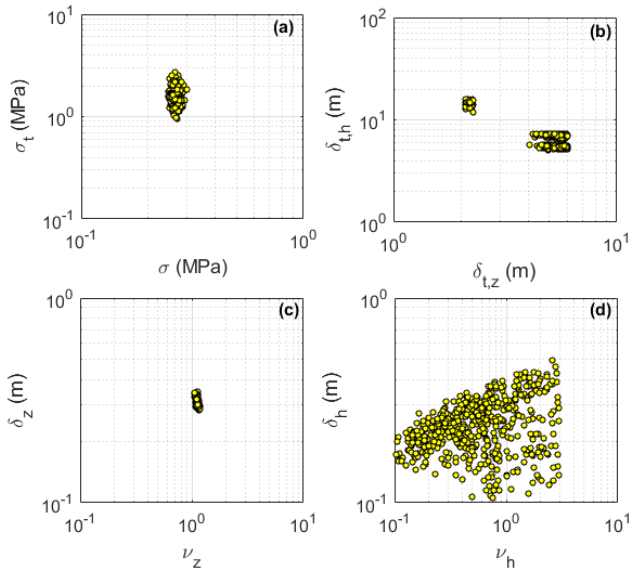

Figure 4: Posterior samples of θ_t and θ (South Parklands): (a,b) posterior samples of θ_t; (c,d) posterior samples of θ.

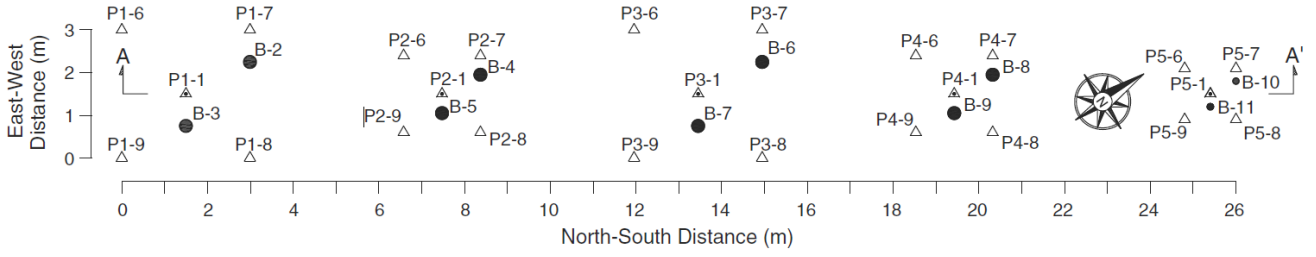

Figure 5: Top view of the Hollywood site (Bong and Stuedlein 2018). The open triangles denote CPTs, whereas the solid circles are boreholes (borehole data not used in analysis).

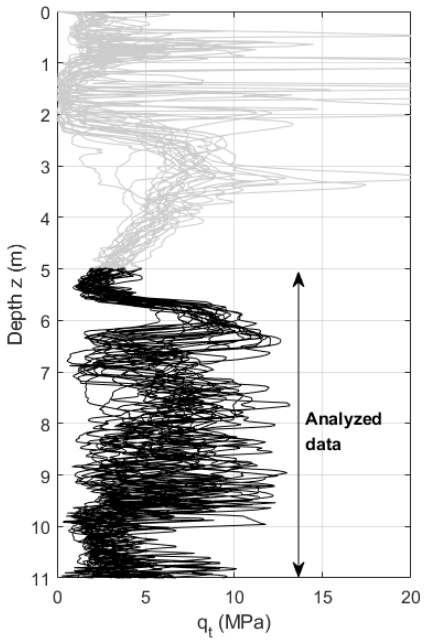

Figure 6: CPT data for the 25 soundings (Hollywood).

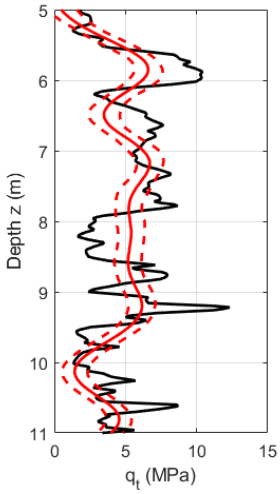

Figure 7: Posterior median and 95% CI of the trend (P1-1; Hollywood).

4. CONCLUSIONS

This paper proposes a probabilistic trend model that is based on the Gaussian process regression (GPR), where the spatial trend is represented as a stationary normal random field with an auto-correlation function modeled by the squared exponential (QExp) model in its prior. Note that the (updated) conditional trend is no longer stationary. The spatial variability is modeled as a stationary normal random field modeled by the Whittle-Matern (WM) model. Numerical and real examples are adopted to illustrate the effectiveness of this GPR method. The numerical example shows that the GPR method can obtain consistent auto-covariance parameter (θ) estimates without the need to prescribe the functional form of the trend. The real examples show that the posterior samples of trend are sensible, following the general trend of the data. The proposed GPR method can provide a flexible model for the trend of geotechnical data.

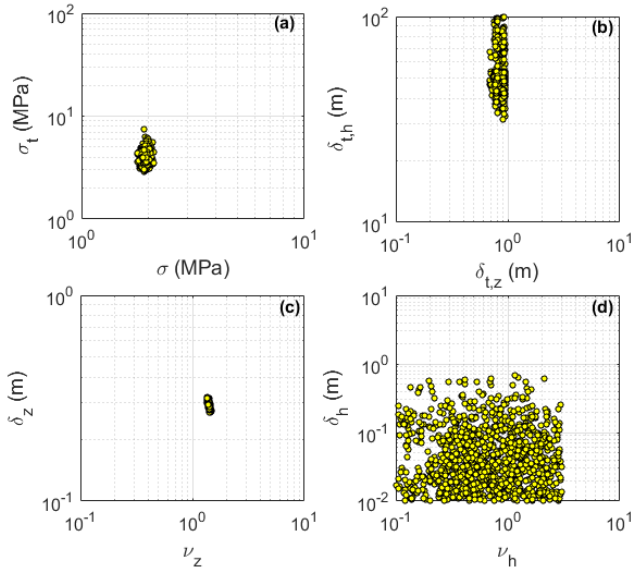

Figure 8: Posterior samples of $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}$ (Hollywood): (a,b) posterior samples of $\boldsymbol{\theta}_t$; (c,d) posterior samples of $\boldsymbol{\theta}$.

5. REFERENCES

Bong, T. and Stuedlein, A.W. (2018). Effect of cone penetration conditioning on random field model parameters and impact of spatial variability on liquefaction-induced differential settlements. *ASCE Journal of Geotechnical and Geoenvironmental Engineering*, 144(5), 04018018.

Ching, J. and Chen, Y.C. (2007). Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection and model averaging. *ASCE Journal of Engineering Mechanics*, 133(7), 816-832.

Ching, J. and Phoon, K.K. (2017). Characterizing uncertain site-specific trend function by sparse

Bayesian learning, *ASCE Journal of Engineering Mechanics*, 143(7), 04017028.

Ching, J. and Phoon, K.K. (2019). Impact of auto-correlation function model on the probability of failure. *ASCE Journal of Engineering Mechanics*, 145(1), 04018123.

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

Ching, J., Yang, Z.Y., and Phoon, K.K. (2021). Dealing with non-lattice data in three-dimensional probabilistic site characterization. *ASCE Journal of Engineering Mechanics*, 147(5), 06021003.

Ching, J., Yoshida, I., and Phoon, K.K. (2023). Comparison of trend models for geotechnical spatial variability: Sparse Bayesian Learning vs. Gaussian Process Regression. *Gondwana Research*, in press.

Fenton, G.A. (1999). Random field modeling of CPT data. *ASCE Journal of Geotechnical and Geoenvironmental Engineering*, 125(6), 486-498.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika*, 93(4), 989-995.

Jaksa, M. (1995). *The Influence of Spatial Variability on the Geotechnical Design Properties of a Stiff, Overconsolidated Clay*. Ph.D. Dissertation, University of Adelaide, Australia.

Jaksa, M.B., Kaggwa, W.S., and Brooker, P.I. (1999). Experimental evaluation of the scale of fluctuation of a stiff clay. *Proceedings of the 8th International Conference on Application of Statistics and Probability*, A.A. Balkema, Rotterdam, 415-422.

Liu, W. F., Leung, Y. F., and Lo, M. K. (2017). Integrated framework for characterization of spatial variability of geological profiles. *Canadian Geotechnical Journal*, 54(1), 47-58.

Phoon, K.K. and Kulhawy, F.H. (1999). Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36(4), 612-624.

Phoon, K.K., Quek, S.T., and An, P. (2003). Identification of statistically homogeneous soil layers using modified Bartlett statistics. *ASCE Journal of Geotechnical and Geoenvironmental Engineering*, 129(7), 649-659.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, London.

Stuedlein, A.W., Gianella, T.N., and Canivan, G.J. (2016). Densification of granular soils using conventional and drained timber displacement piles. *ASCE Journal of Geotechnical and Geoenvironmental Engineering*, 142(12), 04016075.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 211-244.

Vanmarcke, E. H. (1977). Probabilistic modeling of soil profiles. *ASCE Journal of Geotechnical Engineering*, GT11, 1227-1246.

Wang, H., Wang, X., Wellmann, J.F., and Liang, R.Y. (2018). Bayesian stochastic soil modeling framework using Gaussian Markov random fields. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 4(2), 04018014.

Yoshida, I., Tomizawa, Y. and Otake, Y. (2021). Estimation of trend and random components of conditional random field using Gaussian process regression. *Computers and Geotechnics*, 136, 104179.