# Similarity quantification of soil spatial variability between two cross-sections using auto-correlation functions

Yue Hu[1], Yu Wang[2,*], Kok-Kwang Phoon[3], and Michael Beer[4,5,6]

[1] Research Fellow, Institute for Risk and Reliability, Leibniz Universität Hannover, Hannover, Germany. (Email): yue.hu@irz.uni-hannover.de

[2] Professor, Department of Architecture and Civil Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China. (Tel): 852-3442-7605 (Fax): 852-3442-0427 (Email): yuwang@cityu.edu.hk ([*]Corresponding author)

[3] Cheng Tsang Man Chair Professor and Provost, Singapore University of Technology and Design, Singapore. (Email): kkphoon@sutd.edu.sg

[4] Professor and Head, Institute for Risk and Reliability, Leibniz Universität Hannover, Hannover, Germany. (Email): beer@irz.uni-hannover.de

[5] Professor, Institute for Risk and Uncertainty, University of Liverpool, United Kingdom.

[6] Guest Professor, International Joint Research Center for Resilient Infrastructure & International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China.

## Abstract

In geotechnical engineering, an appreciation of local geological conditions from similar sites is beneficial and can support informed decision-making during site characterization. This practice is known as "site recognition", which necessitates a rational quantification of site similarity. This paper proposes a data-driven method to quantify the similarity between two cross-sections based on the spatial variability of one soil property from a spectral perspective. Bayesian compressive sensing (BCS) is first used to obtain the discrete cosine transform (DCT) spectrum for a cross-section. Then DCT-based auto-correlation function (ACF) is calculated based on the obtained DCT spectrum using a set of newly derived ACF calculation equations. The cross-sectional similarity is subsequently reformulated as the cosine similarity of DCT-based ACFs between cross-sections. In contrast to the existing methods, the proposed method explicitly takes soil property spatial variability into account in an innovative way. The challenges of sparse investigation data, non-stationary and anisotropic spatial variability, and inconsistent spatial dimensions of different cross-sections are tackled effectively. Both numerical examples and real data examples from New Zealand are provided for illustration. Results show that the proposed method can rationally quantify cross-sectional similarity and associated statistical uncertainty from sparse investigation data. The proposed method advances data-driven site characterization, a core application area in data-centric geotechnics.



Keywords: Geotechnical site investigation; Site similarity; Auto-correlation; Bayesian compressive sensing

# 1. Introduction

Reliable site characterization is a cornerstone of effective geotechnical designs and construction safety. However, it is often subject to significant uncertainty due to the spatially variable geological conditions and a sparsity of investigation points (e.g., boreholes, in-situ tests). In practice, to supplement limited knowledge at a target site and to mitigate the resultant uncertainty, engineers usually attempt to refer to and review available information (e.g., interpreted soil cross-sections) of previous construction sites in the neighborhood where geological conditions are expected to be similar to the target project site. This practice is also known as "site recognition" which helps engineers to better understand the site-specific features at the target site and plays an important role in data-driven site characterization and informed decision-making in the presence of inter-site variabilities (e.g., Fenton, 1999b; Phoon et al., 2022; Yang et al., 2022; Phoon and Zhang, 2023; Shi et al., 2023; Zhao et al., 2023).

Consider, for example, a two-dimensional (2D) soil property cross-section, which has been commonly adopted in practical engineering designs and analyses for explicitly representing site geological conditions along both depth and horizontal directions. Interpretation of a target 2D cross-section may be underpinned and supplemented by referring to 2D cross-section interpretations available from other pre-existing and documented construction sites. To this end, before introducing knowledge from other 2D cross-sections to inform decision making at a target site, it is desirable to assess the similarity between the target 2D cross-section and the other available 2D cross-sections. Geotechnical engineers primarily do this *qualitatively*, because there are no quantitative methods that are tractable/effective in the presence of sparse and incomplete data to name a few data attributes. From a geotechnical engineering viewpoint, 2D cross-sectional similarity shall be closely related to the similarity of corresponding 2D soil property spatial variability, which is a natural product of complicated geological formation processes (e.g., erosion, weathering, deposition) undergone by the corresponding sites (e.g., Fenton, 1999a; Phoon and Kulhawy, 1999; Baecher and Christian, 2003; Juang et al., 2019; Wang et al., 2022). However, a direct and quantitative comparison of spatial variability in different 2D cross-sections is challenging because of the following issues: (1) the available site investigation data (e.g., borehole and in-situ test data) in both a target 2D cross-section and existing 2D cross-sections are usually sparse and are often not measured over an identical sampling grid (e.g., Xu et al., 2021; Guan and Wang, 2023). It is unlikely that the site investigation plans for two sites are identical (e.g., boreholes or CPT soundings layout). Therefore, classical statistical correlation analysis may not be applicable to quantify the similarity between two such cross-sections; (2) spatial variability in 2D cross-sections may exhibit non-stationary trends and spatial variability anisotropy. Accurate identification of the underlying trend and spatial variability anisotropy for different cross-sections is critical for cross-sectional similarity quantification, but challenging in the presence of sparse data (e.g., Ching and Phoon, 2017; Ching et al., 2017; Hu et al., 2019; Wang et al., 2019; Ching et al., 2020; Shuku et al., 2020; Yoshida et al., 2021; Ching et al., 2022; Katsman and Painuly, 2022); and (3) the dimensions of different 2D cross-sections along depth and horizontal directions are often different due to projects occupying different footprints and extending to different depths. Directly comparing 2D cross-sections with different spatial dimensions is often a tricky task (e.g., Shechtman and Irani, 2007; Simakov et al., 2008; Shi and Wang, 2021b). Therefore, how to quantitatively evaluate the similarity between two given 2D cross-sections from their sparse site investigation data measured over different grids and covering different spatial dimensions remains unsolved.

Recently, the topic of site similarity, or site retrieval, has been investigated from different perspectives. For example, Ching and Phoon (2020) proposed a Bayesian method for measuring similarity between data records (e.g., two or more soil parameter values at a location and or depth) at a target site and data records from other sites. Sharma et al. (2022) further developed a novel hierarchical Bayesian model for measuring similarity between the target site and database sites, achieving site similarity quantification beyond solely data record similarity. Phoon and Ching (2022) presented a summary of different methods for similarity measures. Their frameworks focused on the MUSIC data attributes framework (Multivariate, Uncertain and Unique, Sparse, Incomplete, and potentially Corrupted) (e.g., Phoon et al., 2022) and treated the likelihood function of past data records given site-specific data records as an index of the similarity. Only cross correlations between different soil parameters are considered. The spatial variability of a soil parameter was not considered in these studies, although it was recognized as a critical aspect. In addition, Han et al. (2022) used confidence ellipses to quantify the similarity of soil parametric data using existing databases. Their framework required abundant data over identical depth ranges to be compiled at every site. The performance was highly dependent on the specific volume of available data at different depths. More importantly, the geotechnical spatial variability might not be fully preserved after preprocessing of the data. Shi and Wang (2021a, 2021b) proposed to use training images to incorporate and summarize past geological knowledge on stratigraphy and to quantify the site similarity by measuring the similarity of edge orientation statistics of soil layer boundaries between site-specific borehole data and geological training images. The spatial variability of soil property was not considered. Currently, there is no rational method available for quantifying similarity between 2D cross-sections of soil property from sparse site investigation data with explicit consideration of spatial variability.

This paper proposes a novel method for data-driven quantification of 2D cross-sectional similarity from a spectral perspective. This paper attempts to fill an important gap in the site recognition challenge (e.g., Phoon et al., 2022). A non-parametric method called Bayesian compressive sensing (BCS) is used to directly approximate the sparse spectrum of 2D cross-section. A new efficient and robust formulation of 2D auto-correlation function (ACF) is derived for a unified representation of 2D spatial variability based on a sparse spectrum approximated by BCS. The 2D cross-sectional similarity is then quantified by the similarity between ACFs of the corresponding 2D cross-sections. In contrast to current methods in the literature, cross-sectional similarity quantification in this study deals explicitly with soil property spatial variability. Theoretical derivation suggests that the spatial variability patterns in a 2D cross-section, either stationary or non-stationary, spatially isotropic or anisotropic, can be quantified concisely by 2D ACF. The three challenges highlighted above are solved by the proposed method. The proposed method also has significant practical relevance in geotechnical site recognition. For example, given a global geotechnical database (e.g., Ching et al., 2023) containing a wealth of information from different sites, the proposed method can efficiently pick up a limited number of similar and informative records for a target site, which is also referred to as a “quasi-regional clustering” strategy (e.g., Phoon and Ching, 2022; Guan et al., 2023b).

The rest of this paper is organized as follows. Section 2 briefly illustrates the background and practical significance of 2D cross-sectional similarity using real examples. The proposed method for data-driven quantification of 2D cross-sectional similarity from sparse site investigation data is described in Section 3. The implementation procedure is provided in Section 4, followed by illustrative examples in Section 5. In Section 6, a real case study is used to demonstrate the application of the proposed method.

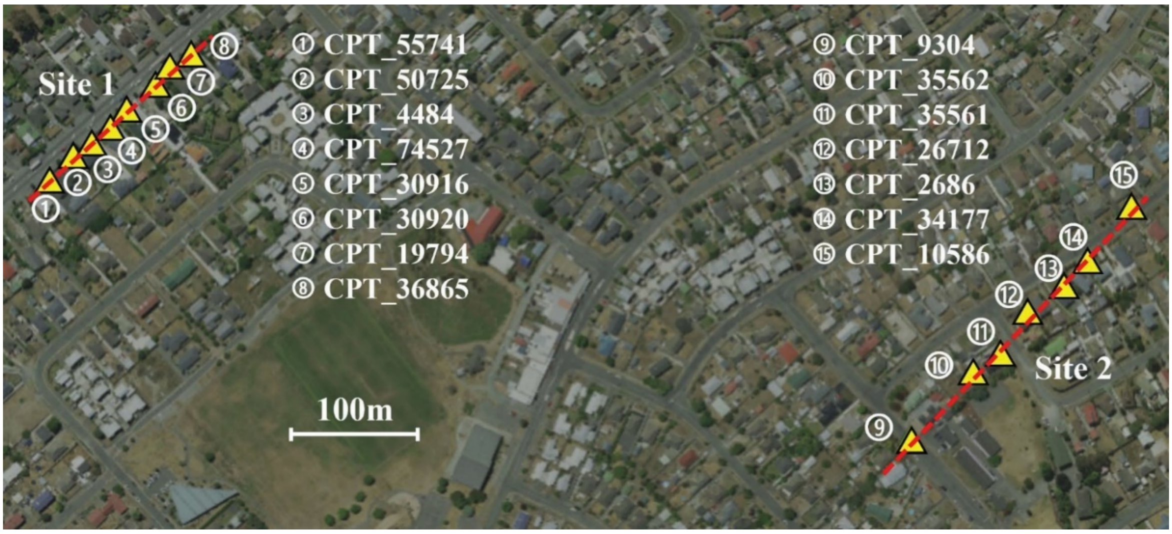



Figure 1. A layout of cone penetration tests (CPTs) performed in two cross-sections in Christchurch, New Zealand (NZGD, 2023)

## 2. Similarity between two 2D cross-sections of a soil property

A 2D cross-section in this study refers to 2D spatial variability of one soil property within a single soil layer. To illustrate the 2D soil property cross-sectional similarity, Figure 1 shows a map with a layout of 15 cone penetration tests (CPTs) performed in Christchurch, New Zealand. The CPTs data are obtained from the New Zealand Geotechnical Database (e.g., NZGD, 2023). In Figure 1, the CPTs are denoted by yellow triangles and numbered from #1 to #15. It is seen that these CPTs were performed at two separate sites, i.e., Site 1 and Site 2, respectively. Eight CPTs (CPT #1 to CPT #8) were performed in Site 1 (see the left-hand side of Figure 1), while seven CPTs (CPT #9 to CPT #15) were carried out in Site 2 (see the right-hand side of Figure 1). The actual IDs of these CPTs used in the NZGD database are also provided in the map. Note that the CPTs at these two sites are roughly laid alone a straight line, leading to two cross-sections denoted by two red dashed lines. Engineers may assess the cross-sections at Site 1 and Site 2 to be similar since the distance between the two cross-sections is only roughly 700m. Figures 2a and 2b show the corrected cone resistance ($q_t$) data of available CPTs by color-coded columns in the two cross-sections, respectively. It appears that the general patterns of $q_t$

data at Sites 1 and 2 are comparable, both with $q_t$ values fluctuating but generally increasing with depth. For example, as shown in Figures 2c and 2d, the $q_t$ profiles of CPT #2 (CPT_50725) from Site 1 and CPT #11 (CPT_35561) from Site 2 exhibit comparable variation patterns. Between these two specific 1D $q_t$ profiles, similarity quantification can be conducted directly and readily. Mathematically, this may be routinely achieved by calculating the cross-correlation between these two 1D $q_t$ profiles after re-configuring the data with an identical sampling interval, or by comparing the corresponding estimated spectrum of these two $q_t$ profiles (e.g., Priestley, 1981; Dai et al., 2022; Guan and Wang, 2023). However, it is very challenging to quantify the $q_t$ data cross-sectional similarity between Site 1 and Site 2, as shown in Figures 2a and 2b. This real example clearly demonstrates the three challenges for direct quantification of 2D cross-sectional similarity mentioned above in concrete terms. First, in Figures 2a and 2b, the available CPT soundings are sparse within these two cross-sections with non-uniform horizontal spacing and sounding depths. Second, the $q_t$ data shown in these two cross-sections exhibit evidence of non-stationarity and spatial variability anisotropy. Third, these two cross-sections have different spatial dimensions. The cross-section of Site 1 has a length of around 145m, while the cross-section of Site 2 has a length of 245m. The challenges highlighted above cannot be addressed by existing methods in literature (e.g., Ching and Phoon, 2020; Han et al., 2022). This next section addresses these challenges by proposing a novel data-driven approach.

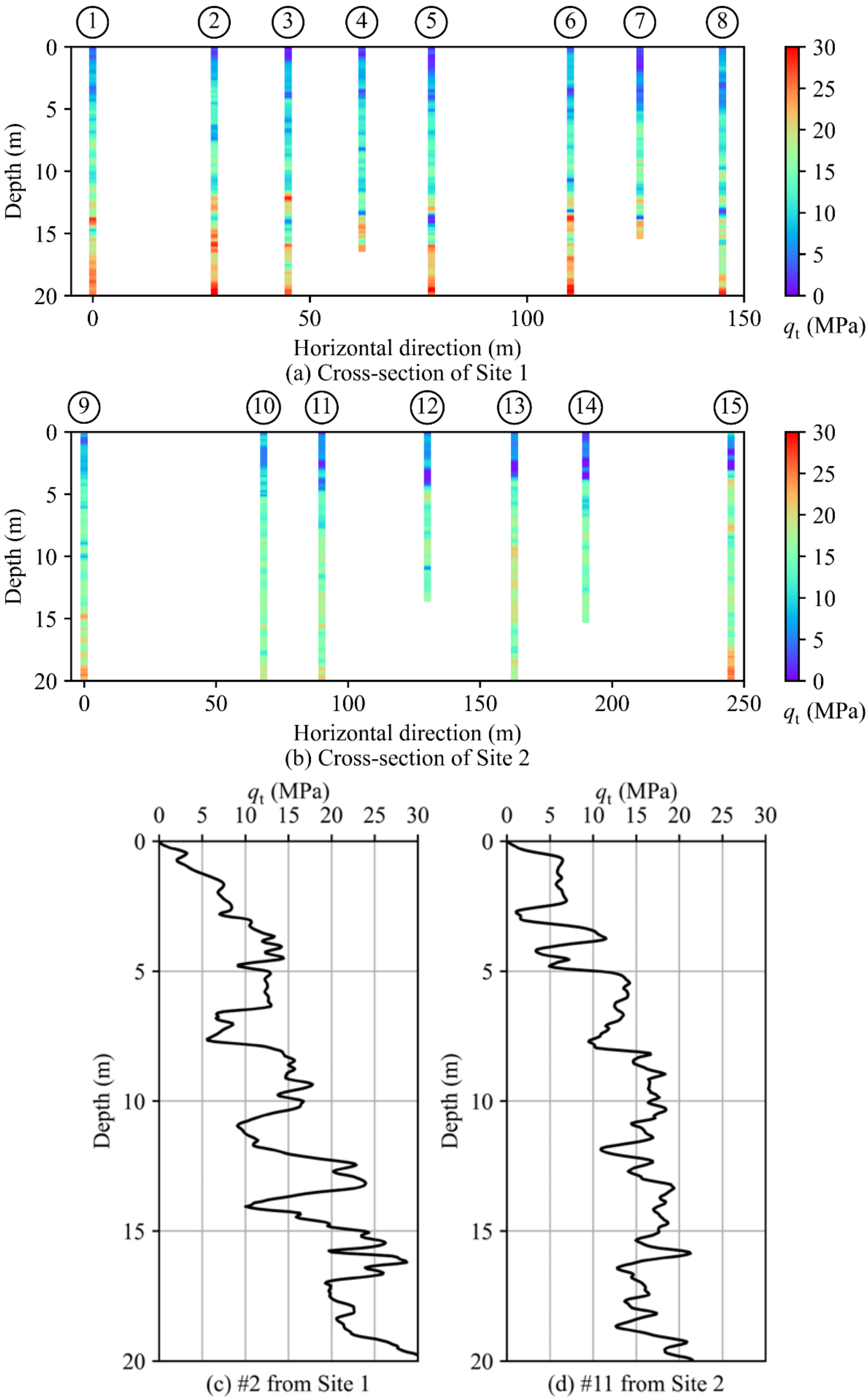



Figure 2. Illustration of CPT example: (a) Cross-section of Site 1; (b) Cross-section of Site 2; (c) $q_t$ data profile of CPT #2 from Site 1; and (d) $q_t$ data profile of CPT #11 from Site 2

## 3. Proposed method for quantification of cross-sectional similarity

The concept for quantification of cross-sectional similarity is to treat soil property cross-sections as images and then compare them from a spectral perspective. Note that image spectral analysis is able to identify non-stationary patterns, spatial variability anisotropy, and spatial shift-invariant patterns (e.g., Shalvi and Weinstein, 1996; Wen and Gu, 2004; Blumensath and Davies, 2006). The results of image spectral analysis are also independent of image dimension. In the proposed method, the following two steps shall be performed before similarity quantification. First, Bayesian compressive sensing (BCS) is adopted to obtain the discrete cosine transform (DCT) spectrum of a 2D soil property cross-section directly from sparse data. It has been shown in past research that BCS can deal with non-stationarity (e.g., Wang et al., 2019; Zhao and Wang, 2020), spatial variability anisotropy of soil property (e.g., Hu et al., 2019), and the associated statistical uncertainty quantification (e.g., Wang et al., 2022). Second, to tackle the difficulty in comparing target 2D cross-sections with different dimensions, a novel and efficient 2D DCT-based ACF is developed to facilitate a unified representation of 2D soil property spatial variability. The DCT-based ACF is utilized in this study as a data-driven surrogate to represent 2D cross-sections and enables direct pattern comparison between cross-sections with different spatial dimensions. Subsequently, cross-sectional similarity is quantified by DCT-based ACF similarity between two cross-sections. Details of the proposed method are elaborated in the following three subsections. The approximation of sparse DCT spectrum by BCS will be described in Subsection 3.1. A unified representation of 2D spatial variability using DCT-based ACF is then derived in Subsection 3.2. Quantification of DCT-based ACF similarity is established in Subsection 3.3.

### *3.1. Approximation of DCT spectrum from sparse data using BCS*

Compressive sensing (CS) is a technique for efficiently acquiring and reconstructing signals or images (e.g., Candès et al., 2006; Donoho, 2006; Candès and Wakin, 2008). Utilizing the sparsity featured by many signals or images after adopting appropriate basis functions, CS is able to reconstruct a signal or image from far fewer measurement data points than the number indicated by conventional Nyquist sampling theorem (e.g., Shannon, 1948; Candès et al., 2006). From a spectral perspective, complicated soil property spatial variability in terms of a 1D profile or 2D cross-section (or image) can be sparsely represented after transformation using basis functions. For example, the DCT functions, which have been widely used in digital signal processing and data compression (e.g., Rao and Yip, 1990; Wallace, 1992), are used to construct basis functions in this study. The commonly used type-II 1D DCT basis function is defined as:

$$\mathrm{B}_t(x) = \begin{cases} \dfrac{1}{\sqrt{N}} & for\ t = 1; x = 1,2,\cdots,N \\ \sqrt{\dfrac{2}{N}} cos\pi \dfrac{(t-1)(2x-1)}{2N} & for\ t = 2,\cdots,N; x = 1,2,\cdots,N \end{cases} \tag{1}$$

in which $x$ represents the 1D index ($x$=1, 2, …, $N$); $t$ indicates the order of $\mathrm{B}_t(x)$. In Figure 3, the first five DCT basis functions (i.e., $t$ = 1, 2, 3, 4, 5) with $N$ = 200 are illustrated by colored lines with different styles. The frequency of these DCT basis function $\mathrm{B}_t(x)$ is controlled by $t$ and increases with $t$. Based on the 1D DCT basis functions in Equation (1), 2D DCT basis functions may be constructed by a tensor product of two 1D DCT basis functions (e.g., Itskov, 2007):

$$\mathrm{B}_{t,s}(x_1, x_2) = \mathrm{B}_t(x_1) \times \mathrm{B}_s(x_2) \tag{2}$$

in which $\mathrm{B}_t(x_1)$ and $\mathrm{B}_s(x_2)$ are two basis functions along two directions, respectively; $t$ and $s$ indicate the corresponding orders ($t$=1, 2, …, $N_1$; $s$=1, 2, …, $N_2$). For example, Figure 4 illustrates the construction process of 25 2D DCT basis functions $\mathrm{B}_{t,s}(x_1, x_2)$. Each 2D DCT basis function is constructed by a tensor product of two 1D DCT basis functions of the same length at each frequency (e.g., $t$, $s$=1, 2, 3, 4, 5). Using the 2D DCT basis function $\mathrm{B}_{t,s}(x_1, x_2)$, soil property spatial variability in a cross-section can be regarded as an image F with size $N_1 \times N_2$, which is formulated as (e.g., Tipping, 2001; Candès and Wakin, 2008):

$$\mathrm{F}(x_1, x_2) = \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} \omega_{t,s}^{2D} \mathrm{B}_{t,s}(x_1, x_2) \tag{3}$$

in which $\mathrm{F}(x_1, x_2)$ is the 2D spatial variability in the cross-section; $x_1$, $x_2$ are indexes along two directions, respectively ($x_1$=1, 2, …, $N_1$; $x_2$=1, 2, …, $N_2$); $\omega_{t,s}^{2D}$ is the weight coefficient of $\mathrm{B}_{t,s}(x_1, x_2)$. The weight coefficients and their corresponding frequencies collectively form the DCT spectrum of $\mathrm{F}(x_1, x_2)$.

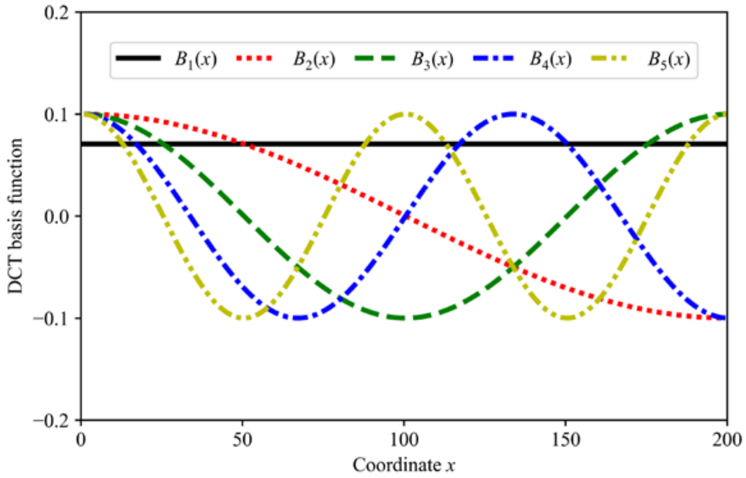



Figure 3. Illustration of five 1D discrete cosine transform (DCT) basis functions

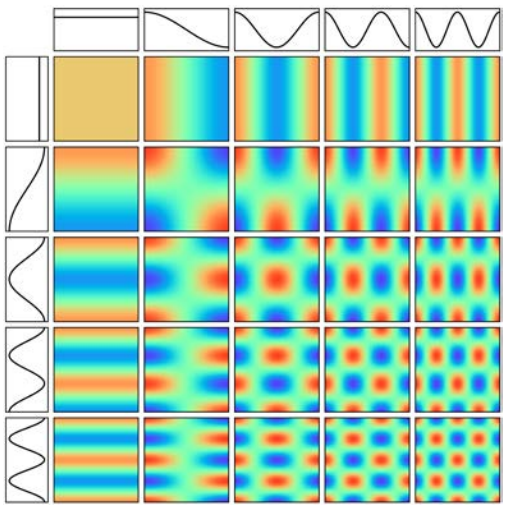


Figure 4. Construction of 2D DCT basis functions from 1D DCT basis functions

The DCT spectrum enables an effective representation of variability patterns at various frequencies along two directions. Note that both non-stationarity and spatial variability anisotropy may be preserved by a combination of various $\mathrm{B}_{t,s}(x_1, x_2)$ with large weight coefficients and specific patterns (e.g., see various 2D DCT basis functions in Figure 4). In addition, due to the spatial correlation contained in soil property spatial variability, $\mathrm{F}(x_1, x_2)$, usually leads to a sparse representation in its DCT spectrum. In other words, most weight coefficients in Equation (3) have negligible magnitudes and only limited weight coefficients are significant, or non-trivial, after adopting proper basis functions (e.g., Candès and Wakin, 2008; Zhao et al., 2018; Hu et al., 2019). It is therefore feasible to approximate those limited non-trivial weight coefficients in DCT spectrum using sparse data Y, which is expressed as:

$$\mathrm{Y}(x_1, x_2) = \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} \omega_{t,s}^{2D} \mathrm{A}_{t,s}(x_1, x_2) \tag{4}$$

in which $\mathrm{Y}(x_1, x_2)$ is the measured data points from $\mathrm{F}(x_1, x_2)$, and $\mathrm{A}_{t,s}(x_1, x_2)$ are corresponding values extracted from the 2D basis function $\mathrm{B}_{t,s}(x_1, x_2)$ based on the measurement locations. In the context of geotechnical site investigation, generally the number, $M$, of available measurement $\mathrm{Y}(x_1, x_2)$, is much smaller than the size of $\mathrm{F}(x_1, x_2)$ (i.e., $N_1 \times N_2$), and this leads to an underdetermined system in Equation (4), which cannot be solved directly. Numerical algorithms, e.g., orthogonal matching pursuit (OMP), may be used to approximate the solution $\widehat{\omega}_{t,s}^{2D}$ in Equation (4) by minimizing the error between the measured data $\mathrm{Y}(x_1, x_2)$ and the estimated values at measured locations (e.g., Pati et al., 1993; Wang and Zhao, 2016). The idea of the OMP algorithm is to iteratively find out $\mathrm{B}_{t,s}(x_1, x_2)$ that can well match the $\mathrm{Y}(x_1, x_2)$. However, since the available site investigation data $\mathrm{Y}(x_1, x_2)$ are sparse, CS may not produce a perfect reconstruction of the spatial variability $\mathrm{F}(x_1, x_2)$, and the approximated $\widehat{\omega}_{t,s}^{2D}$ contain significant statistical uncertainty. To quantify the associated statistical uncertainty, CS may be integrated with the Bayesian framework (i.e., Bayesian compressive sensing, BCS) to estimate the non-trivial weight coefficients. In BCS, the prior of non-trivial weight coefficients is formulated as independent normal random variables with relatively large variance to achieve an uninformative prior. The posterior distribution of $\widehat{\omega}_{t,s}^{2D}$ also follows a normal distribution and can be solved efficiently by a Markov chain Monte Carlo (MCMC) simulation, leading to a series of random samples of $\widehat{\omega}_{t,s}^{2D}$ (e.g., Zhao and Wang, 2020; Wang et al., 2022; Lyu et al., 2023). After repeatedly generating random samples of $\widehat{\omega}_{t,s}^{2D}$ $N_B$ times, the best estimate of DCT spectrum can be approximated by taking the mean of $N_B$ random samples of $\widehat{\omega}_{t,s}^{2D}$. In a MCMC simulation, the statistical independence of $N_B$ random samples of $\widehat{\omega}_{t,s}^{2D}$ is guaranteed by taking only one sample in every larger number (e.g., 20 or 50) Markov chain samples as the random sample. The associated statistical uncertainty can be quantified using the standard deviation (SD) of $N_B$ samples. Another noteworthy advantage of BCS is that it is applicable to a non-uniform measurement grid, a scenario commonly encountered in site investigation (e.g., Zhao and Wang 2020; Guan et al., 2023a). Both the CS and BCS algorithms have been compiled into a user-friendly free download software which is available from the corresponding author's website (https://sites.google.com/site/yuwangcityu/software-download/bayesian-compressive-samplingsensing-bcs). The approximated DCT spectrum for the cross-section allows subsequent development of a unified representation of 2D spatial variability and quantification of a cross-sectional similarity.

### *3.2. Unified representation of 2D spatial variability using DCT-based ACF*

Note that each weight coefficient $\omega_{t,s}^{2D}$ in DCT spectrum corresponds to a specific basis function $\mathrm{B}_{t,s}(x_1, x_2)$, which is defined over specific $N_1$ and $N_2$ (see Equations (1) and (2)). In other words, the DCT spectrum is relative to the dimension $N_1$ and $N_2$, which are essentially determined by both the cross-section dimension and discretization resolution. This indicates that direct comparison of the DCT spectrums of spatial variability $\mathrm{F}(x_1, x_2)$ obtained in different 2D cross-sections may not be feasible. To this end, a unified representation of 2D spatial variability using DCT-based ACF is developed, which enables direct comparison between 2D cross-sections with different spatial dimensions.

In signal processing, ACF is an effective tool for evaluating the correlation structure of signals (e.g., Vanmarcke, 2010; Onyejekwe et al., 2016). ACF essentially measures the correlation of a signal with a shifted version of itself. Mathematically, it describes how the correlation between two points varies as the lag distance between the two points changes. In the context of 2D cross-sectional spatial variability, ACF of $\mathrm{F}(x_1, x_2)$ can be calculated as (e.g., Webster and Oliver, 2007; Vanmarcke, 2010):

$$\mathrm{ACF}[\mathrm{F}(x_1, x_2), \tau_1, \tau_2] = \frac{E\{[\mathrm{F}(x_1, x_2) - \mu_{\mathrm{F}(x_1,x_2)}][\mathrm{F}(x_1+\tau_1, x_2+\tau_2) - \mu_{\mathrm{F}(x_1,x_2)}]\}}{\sigma^2_{\mathrm{F}(x_1,x_2)}} \tag{5}$$

in which $\tau_1, \tau_2$ are the lag distances along $x_1$ and $x_2$ directions, respectively; $\mu_{\mathrm{F}(x_1,x_2)}$ and $\sigma_{\mathrm{F}(x_1,x_2)}$ are the mean value and SD of all data points in $\mathrm{F}(x_1, x_2)$, respectively. Equation (5) does not assume stationarity. The ACF reflects the auto-correlation structure of 2D spatial variability with respect to its mean value. It is a normalized and non-parametric measure because it is normalized by the variance value and it is not fitted to any parametric function form. In many fields, ACF has been widely used to identify predominant patterns/frequencies embedded in the signals or images of interest (e.g., Priestley, 1981; Rafiee and Tse, 2009; Zhang et al., 2021). Mathematically, ACF is closely related to the spectrum, and they form a Wiener–Khinchin transform pair (e.g., Priestley, 1981). Although ACF may be used as a surrogate to represent patterns of 2D cross-sectional spatial variability, it is worth noting that calculating 2D ACF accurately and efficiently is usually difficult. Conventional approach of calculating 2D ACF using Equation (5) often yields instable ACF values at large lag distances and may be subject to significant computational efforts when dealing with high-resolution images/matrices (e.g., Phoon and Fenton, 2004). To tackle this issue, this subsection derives a new efficient formulation of ACF based on the DCT spectrum obtained from BCS:

$$\mathrm{ACF}[\mathrm{F}(x_1,x_2),\tau_1,\tau_2] = \frac{1}{\sum_{t=1}^{N_1}\sum_{s=1}^{N_2}{\omega_{t,s}^{2D}}^2}\sum_{t=1}^{N_1}\sum_{s=1}^{N_2}{\omega_{t,s}^{2D}}^2\,\mathrm{ACF}\big[\mathrm{B}_{t,s}(x_1,x_2),\tau_1,\tau_2\big] \qquad (6)$$

$$(t,s \neq 1\ concurrently)$$

in which $\mathrm{ACF}\big[\mathrm{B}_{t,s}(x_1,x_2),\tau_1,\tau_2\big]$ is the ACF of $\mathrm{B}_{t,s}(x_1,x_2)$. Step-by-step derivation of Equation (6) is provided in Appendix. Equation (6) shows that the ACF of 2D spatial variability $\mathrm{F}(x_1,x_2)$ is a weighted summation of the ACFs of 2D DCT basis functions, which are functions of lag distances $\tau_1$ and $\tau_2$. The weight is the corresponding squared weight coefficient in DCT spectrum. Note that Equation (6) establishes a theoretical basis for the unified representation of 2D cross-sectional spatial variability, including non-stationarity and spatial variability anisotropy. Equation (6) can also be interpreted as a special case of covariance decomposition in traditional Karhunen-Loève expansion (e.g., Huang et al., 2001). Moreover, DCT-based ACF enables a direct and convenient comparison between different 2D cross-sections. Using $N_B$ random samples of DCT spectrum, $N_B$ DCT-based ACFs are obtained by substituting $N_B$ random samples of $\hat{\omega}_{t,s}^{2D}$ into Equation (6). The best estimate DCT-based ACF is calculated as the mean of $N_B$ DCT-based ACFs. Statistical uncertainty of approximated DCT spectrum also propagates to DCT-based ACFs and can be quantified using SD of the $N_B$ DCT-based ACFs.

### *3.3. Quantification of DCT-based ACF similarity between cross-sections*

With the DCT-based ACFs of two cross-sections determined in Section 3.2, cross-sectional similarity can be quantified by the similarity between the corresponding DCT-based ACFs. Consider, for example, two cross-sections A and B. Note that the actual dimensions of cross-sections A and B may be different, leading to the different dimensions and patterns of the corresponding DCT-based ACFs. To enable fair and effective comparison between two cross-sections, only the largest overlapped sections with common lag distances of DCT-based ACFs of cross-sections A and B are used accordingly. For example, if the sizes of spatial variability matrices for cross-sections A and B are (200, 300) and (300, 200), respectively, only the overlapped DCT-based ACFs with an identical range of lag distances, i.e., $\tau_1$ = 0, 1, 2, …, 199 and $\tau_2$ = 0, 1, 2, …, 199, are considered for similarity quantification. The overlapped ACFs called effective ACFs offer a benchmark for comparison of two different cross-sections in a statistical manner. Mathematically, a generalized cosine similarity between the effective DCT-based ACFs of cross-sections A and B is calculated as (e.g., Dong et al., 2006; Nguyen and Bai, 2011; Hu and Wang 2024):

$$\rho_{\mathrm{AB}} = \frac{tr\left(\mathbf{ACF}_{\mathrm{A}} \cdot \mathbf{ACF}_{\mathrm{B}}^{\mathrm{T}}\right)}{\sqrt{tr\left(\mathbf{ACF}_{\mathrm{A}} \cdot \mathbf{ACF}_{\mathrm{A}}^{\mathrm{T}}\right)}\sqrt{tr\left(\mathbf{ACF}_{\mathrm{B}} \cdot \mathbf{ACF}_{\mathrm{B}}^{\mathrm{T}}\right)}} \tag{7}$$

in which $\rho_{\mathrm{AB}}$ is defined as the similarity value between cross-sections A and B; $\mathbf{ACF}_{\mathrm{A}}$ and $\mathbf{ACF}_{\mathrm{B}}$ are matrix representations of the effective 2D DCT-based ACFs of cross-sections A and B, respectively; "T" is a transpose operation of a matrix; "$tr$" is the trace operation of a matrix. This formula is equivalent to calculating the sum of element-wise product of $\mathbf{ACF}_{\mathrm{A}}$ and $\mathbf{ACF}_{\mathrm{B}}$, divided by the product of the Frobenius norms of $\mathbf{ACF}_{\mathrm{A}}$ and $\mathbf{ACF}_{\mathrm{B}}$. $\rho_{\mathrm{AB}}$ is therefore defined over the range of [-1, 1]. Equation (7) essentially treats $\mathbf{ACF}_{\mathrm{A}}$ and $\mathbf{ACF}_{\mathrm{B}}$ as high-dimensional vectors and measures the cosine value of the angle between the two vectors. High $\rho_{\mathrm{AB}}$ indicates closeness between the two vectors and hence high similarity between cross-sections A and B, and vice versa.

Note that a deterministic $\rho_{\mathrm{AB}}$ is obtained when substituting the corresponding best estimate DCT-based ACFs of cross-sections A and B into Equation (7). To consider the associated statistical uncertainty in both cross-sections A and B simultaneously, Equation (7) may be used in a probabilistic manner. A random sample of $\rho_{\mathrm{AB}}$ is calculated using Equation (7) by substituting a pair of random samples of DCT-based ACFs of cross-sections A and B respectively into Equation (7). After repeating the $\rho_{\mathrm{AB}}$ calculation for all $N_{\mathrm{B}}$ pairs of DCT-based ACFs of two cross-sections, statistical analysis is performed on the obtained $N_{\mathrm{B}}$ $\rho_{\mathrm{AB}}$ values. The statistical uncertainty associated with the cross-sectional similarity quantification is expressed by the SD of the $\rho_{\mathrm{AB}}$ samples. The SD reflects the variability of the cross-sectional similarity quantification in the presence of uncertainties in both cross-sections. Note that in engineering practice, the required number of $N_B$ depends on the characteristics of the spatial variability in the two cross-sections. The optimum value of $N_B$ may be identified by examining the convergence behavior of the obtained similarity values.

## 4. Implementation procedures

To facilitate its applicability in engineering practice, this section summarizes the implementation procedure of the proposed method for cross-sectional similarity quantification. For example, two cross-sections, e.g., A and B, are to be evaluated. Five steps are involved in implementing the cross-sectional similarity quantification between A and B, as described below:

**Step 1**: Obtain the actual spatial dimensions (i.e., depths and horizontal lengths), and determine the corresponding spatial resolutions $N_1 \times N_2$ of 2D spatial variability $\mathrm{F}(x_1, x_2)$ for cross-sections A and B, respectively. For example, if a cross-section has a depth of 20m and a horizontal distance of 30m, spatial resolutions of 0.1m along both directions will lead to a discretized 2D cross-section of shape 200×300.

**Step 2**: Compile the available soil property data within the cross-sections A and B as measurement data $\mathrm{Y}(x_1, x_2)$, which is a subset of $\mathrm{F}(x_1, x_2)$. This step leads to two sets of $\mathrm{Y}(x_1, x_2)$ for cross-sections A and B, respectively.

**Step 3**: Perform BCS simulation to generate $N_B$ (e.g., $N_B$=500) random samples of DCT spectrum from the corresponding measurement data $\mathrm{Y}(x_1, x_2)$ in two cross-sections, respectively.

**Step 4**: Calculate the $N_B$ DCT-based ACFs using Equation (6) and $N_B$ random samples of DCT spectrum for cross-sections A and B, respectively.

**Step 5**: Perform probabilistic cross-sectional similarity quantification. $N_\mathrm{B}$ DCT-based ACFs of cross-section A are randomly paired with the $N_\mathrm{B}$ DCT-based ACFs of cross-section B. One similarity value is then obtained using Equation (7) for each pair of DCT-based ACFs. $N_\mathrm{B}$ pairs of DCT-based ACFs lead to $N_\mathrm{B}$ similarity values. Statistical analysis is then performed on the obtained $N_\mathrm{B}$ similarity values.

With the mean of these $N_\mathrm{B}$ similarity values, the similarity between cross-sections A and B can be evaluated based on a pre-specified threshold. The threshold is purpose-dependent and problem-specific. The question of how to select an optimal threshold is an interesting research topic and will be investigated in a future study. One possible approach is to develop characteristic values of similarity for a specific geotechnical problem based on many case studies performed in similar geological settings. The statistical uncertainty of cross-sectional similarity is quantified using SD of $N_\mathrm{B}$ similarity values. Note that geotechnical analysis is purpose-dependent and problem-specific. Different engineering projects might be sensitive to the spatial variability of soil properties to different extents. This study only considers the statistical similarity (e.g., the ACF similarity) of 2D cross-sectional spatial variability of one soil property. This study does not consider the response of structures installed in the soil as a result of spatial variability. Geotechnical analyses of different projects still need to resort to specific domain knowledge from geotechnical engineers, even for the same or highly similar site but with different purposes (e.g., deep foundation design versus liquefaction assessment) (e.g., Leung, 2023). In the following section, the proposed method is illustrated using numerical examples.

# 5. Illustrative examples

### *5.1. Numerical examples of soil property cross-sections*

In this section, three cross-sections, i.e., namely A, B, and C, are simulated for illustration of the proposed method. The configurations of these three cross-sections are summarized in Table 1. These three cross-sections have different spatial dimensions. Cross-section A has a depth of 20m and a width of 20m, i.e., a 20m×20m cross-section. Cross-sections B and C are 30m×20m and 20m×30m cross-sections, respectively. A discretization resolution of 0.1m is adopted for these three cross-sections, leading to discretized cross-sections with spatial resolution $N_1 \times N_2$ = 200×200, 300×200, and 200×300, respectively. Non-stationary undrained shear strength $s_u$ data are simulated using random field (RF) models for these three discretized cross-sections. The non-stationary $s_u$ random fields are realized by adding a non-stationary trend function to a 2D zero-mean random field. As summarized in Table 1, in these three cross-sections, the $s_u$ data have different trend functions, which are formulated as the sum of 50 kPa and a scaled cosine function term (in kPa) in different frequencies or phases. The cosine trend functions adopted herein are to model the periodic property of geological depositional conditions (e.g., Einsele et al., 1996). Note that the trend functions of cross-section A and cross-section B exhibit the same frequency, i.e., 0.5, which is equivalent to a period of around 12.5m, while cross-section B incorporates an additional spatial shift of 10m. Cross-section B may be interpreted as a cross-section exhibiting similar geological depositional conditions to A, but occurring at another elevation level, as a scenario commonly encountered in engineering practice. Cross-section C contains a frequency of 1, which is higher than cross-sections A and B, and does not incorporate any spatial shift. The trend functions of the three cross-sections are illustrated in Figures 5a-c, respectively. For each cross-section, an SD of 5 kPa and a Gaussian auto-correlation structure are adopted for the 2D zero-mean random field. Different correlation lengths along vertical and horizontal directions are configured for each cross-section. As shown in Table 1, the vertical correlation lengths for three cross-sections are 1m, 1.5m, and 1m, respectively; the horizontal correlation lengths for three cross-sections are 3.5m, 3m, and 2.5m, respectively, leading to different spatial variability anisotropy structures for $s_u$ data. For each cross-section, one $s_u$ data cross-section is realized and used for illustration, as shown in Figures 5d-f, respectively. Each $s_u$ data cross-section is a realization of a random field simulated by the spectral representation method (e.g., Shinozuka and Deodatis 1991; Müller et al., 2022).

Table 1. Configurations of simulated undrained shear strength $s_u$ data for three cross-sections

| Parameters | Cross-section A | Cross-section B | Cross-section C |
| --- | --- | --- | --- |
| Depth (m) | 20 | 30 | 20 |
| Width (m) | 20 | 20 | 30 |
| Trend $s_u(z)$ versus depth (kPa) | 50+5×cos(0.5×z) | 50+ 5×cos(0.5×(z+10)) | 50+5×cos(z) |
| Standard deviation (kPa) | 5 | 5 | 5 |
| Correlation function | Gaussian | Gaussian | Gaussian |
| Vertical correlation length (m) | 1 | 1.5 | 1 |
| Horizontal correlation length (m) | 3.5 | 3 | 2.5 |

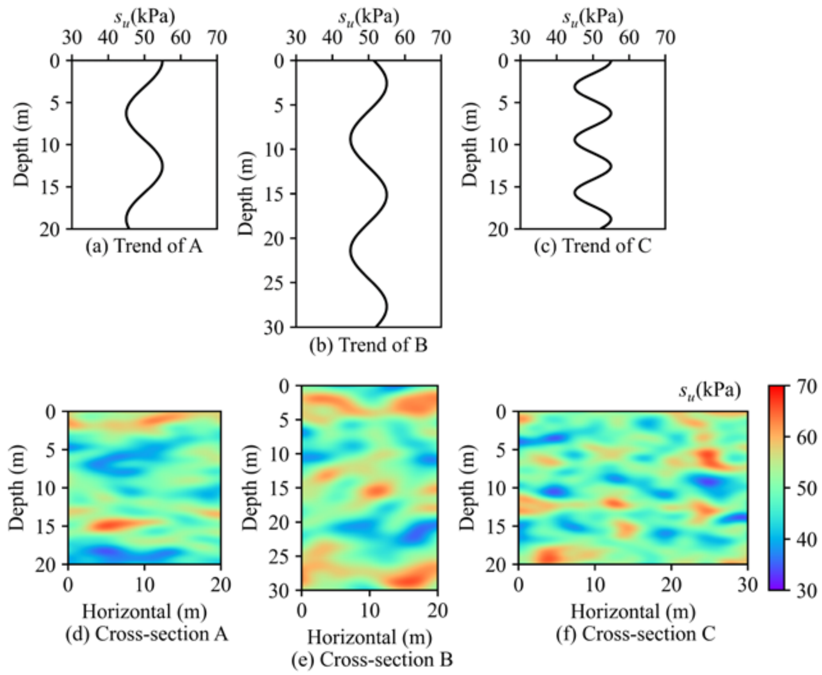



Figure 5. Simulated undrained shear strength ($s_u$ in kPa) data cross-sections A, B, and C

Note that the three cross-sections are configured to illustrate the challenges of cross-sectional similarity quantification. In Figure 5, it is seen that cross-sections A, B, and C have different spatial dimensions and show different non-stationary and spatial variability anisotropy patterns. It is very challenging to rationally quantify the similarity among these cross-sections using conventional statistical methods. In this study, the derived DCT-based ACF tackles this challenge and offers an effective way to quantify the cross-sectional similarity among these three cross-sections. For each of the three cross-sections, the associated DCT spectrum can be readily obtained using Equation (3), and subsequently, the associated DCT-based ACF can be calculated using Equation (6). Note that the three cross-sections are respectively synthesized by adding up a non-stationary trend function and a 2D zero-mean random field. Therefore, the ACFs of $s_u$ cross-sections are controlled by both the underlying trends and zero-mean RFs. The ACFs of the trend functions for the three cross-sections are shown in Figures 6a-6c, respectively, while the corresponding theoretical RF ACFs are shown in Figures 6d-6f, respectively. Figures 6g-6i, respectively, show the ACFs of the three synthesized $s_u$ cross-sections. The 2D ACFs are plotted as a colormap versus varying lag distances along two directions. Since the auto-correlation decreases as the lag distance increases, only half of the maximum lag distances along both directions are considered as shown in the figure. In Figures 6a-6c, it is shown that the DCT-based ACFs of the trend functions behave also like cosine functions, with ACF values fluctuating at corresponding frequencies along the depth direction. In Figures 6d-6f, theoretical RF ACFs decay along both directions accordingly to the corresponding RF parameters. Note that in Figures 6g-6i, the DCT-based ACFs of the three $s_u$ cross-sections show combined patterns exhibiting features of the ACFs from the trend functions and the ACFs from the RFs. This indicates that ACF not only may be used to characterize a zero-mean RF, but also simultaneously characterize the underlying deterministic trend function (e.g., Brockwell and Davis, 1991).

Note that the three DCT-based ACFs have different shapes. To fairly compare these DCT-based ACFs, the largest overlapped sections between any two cross-sections are selected, as delineated by red dashed lines in Figures 6g-6i. For any two cross-sections, a similarity value is calculated using Equation (7) and the corresponding overlapped DCT-based ACFs. The similarity values for different pairs of cross-sections are calculated as $\rho_{AB} = 0.97$, $\rho_{AC} = 0.27$, and $\rho_{BC} = 0.36$. The similarity values are consistent with the theoretical configurations of the three cross-sections. It has been indicated in Table 1 that cross-sections A and B demonstrate better spectral coherence since their non-stationary trend functions have an identical frequency, although the trend function of cross-section B has a spatial shift. However, the trend function of cross-section C contains a higher frequency than A and B. From a spectral perspective, cross-section C may not be similar to cross-sections A and B. Note that the similarity quantification using DCT-based ACF is invariant to the spatial offset values of spatial variability. Similar examples with different spatial offset values had been analyzed, and consistent results were obtained. In this paper, only the cross-sections configured in Table 1 are presented for brevity.

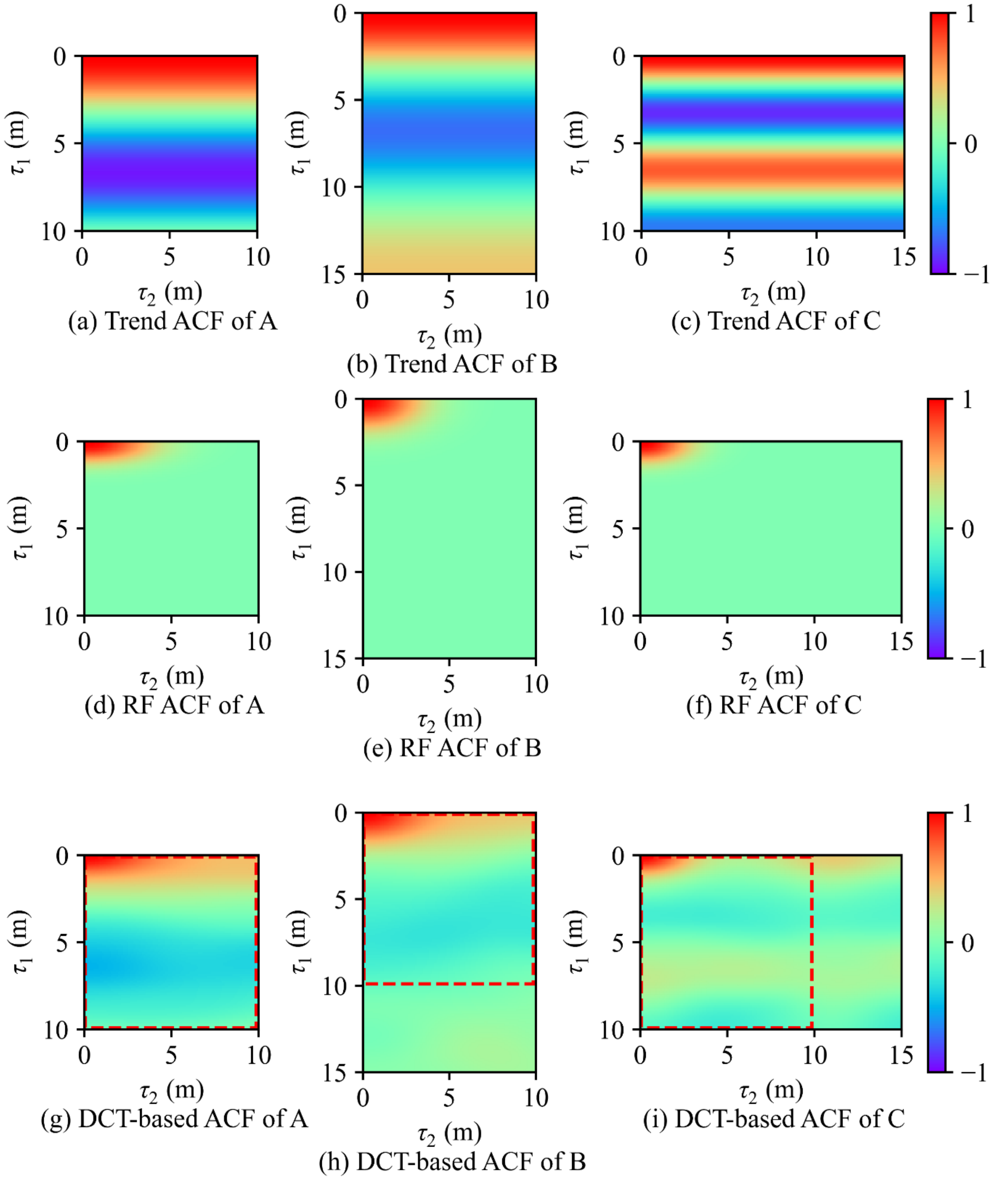



Figure 6. DCT-based ACF in cross-sections A, B, and C: (a) Trend ACF of A; (b) Trend ACF of B; (c) Trend ACF of C; (d) RF ACF of A; (e) RF ACF of B; (f) RF ACF of C; (g) ACF of synthetic $s_u$ data in A; (h) ACF of synthetic $s_u$ data in B; (i) ACF of synthetic $s_u$ data in C

Note that the above cross-sectional similarity quantification is based on the three simulated cross-sections with complete $s_u$ data. To illustrate the typical scenario of sparse investigation data, the following subsection performs cross-sectional similarity quantification using the same examples, but with sparsely measured data.

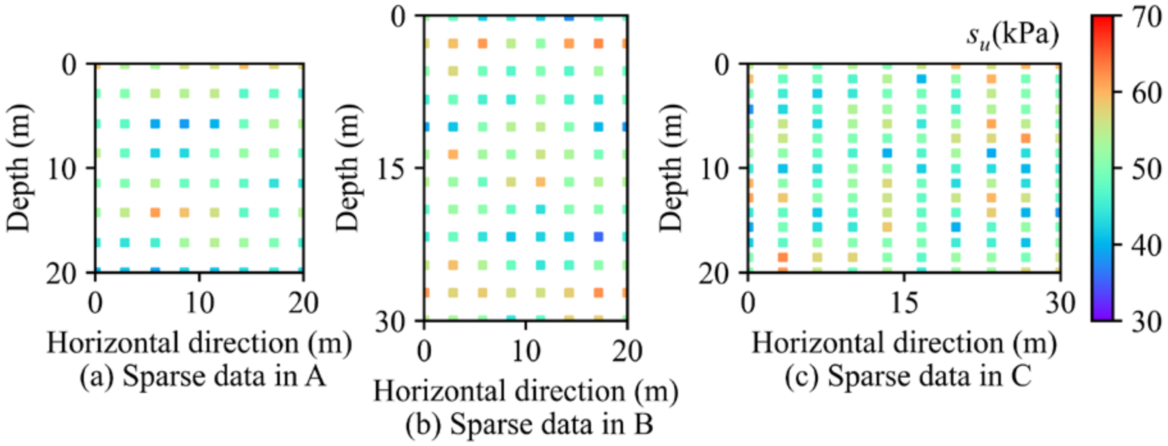



Figure 7. Sparse data measured from cross-sections A, B, and C

### *5.2. Similarity quantification between simulated cross-sections using sparse data*

To illustrate the challenge of sparse investigation data in cross-sections, selective $s_u$ data are sampled from the simulated cross-sections as measurement data, which are subsequently used for probabilistic quantification of similarity between cross-sections with consideration of statistical uncertainty. As shown in Figure 7, uniform grid sampling is implemented in the three cross-sections. In cross-sections A, B, and C, 8×8, 12×8, and 15×10 $s_u$ data are measured, respectively. The measured data account for around 0.16%, 0.16%, and 0.25%, respectively, of the corresponding discretized cross-section. The sampling ratios adopted are generally comparable to the engineering practice of site investigation, where the ratio of the volume of sampled soils over the volume of soils loaded/affected is normally around or less than 0.1%, depending on the project requirements, site complexity, and the level of details needed (e.g., Look 2014; Guan and Wang, 2020). For the three cross-sections, the spatial resolutions $N_1 \times N_2$ are set as the original simulated cross-section, i.e., 200×200 for A, 300×200 for B, and 200×300 for C. The corresponding measurement data are then adopted as $\mathrm{Y}(x_1, x_2)$, and subsequently, BCS simulation is performed to generate $N_{\mathrm{B}}$=500 samples of DCT spectrum for each cross-section.

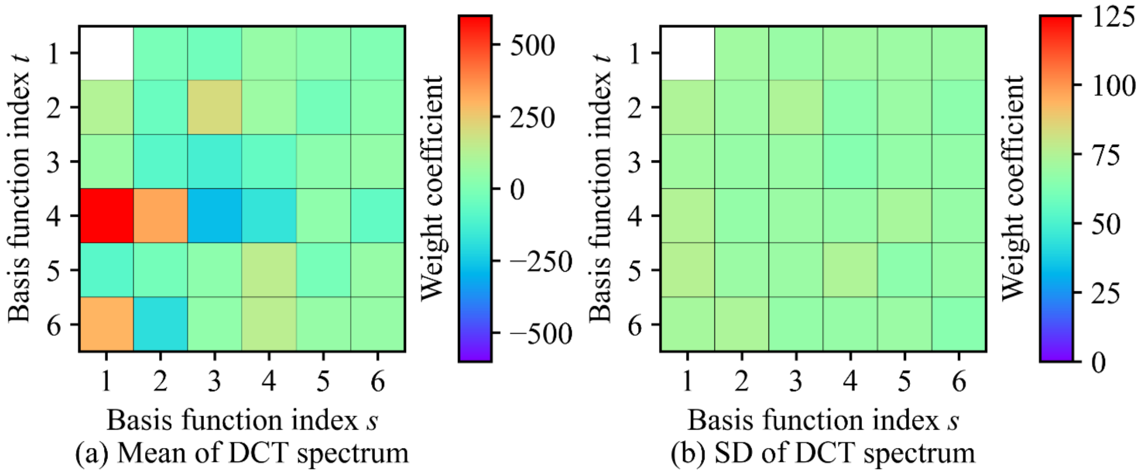



Figure 8. Statistics of DCT spectrum in cross-section A

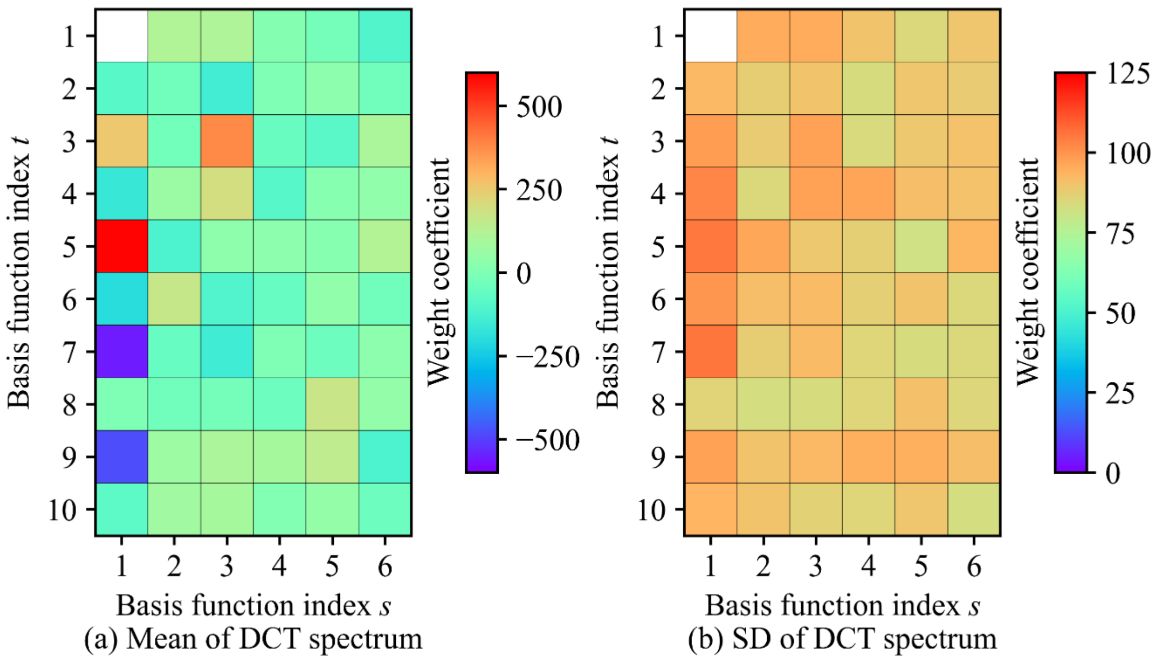
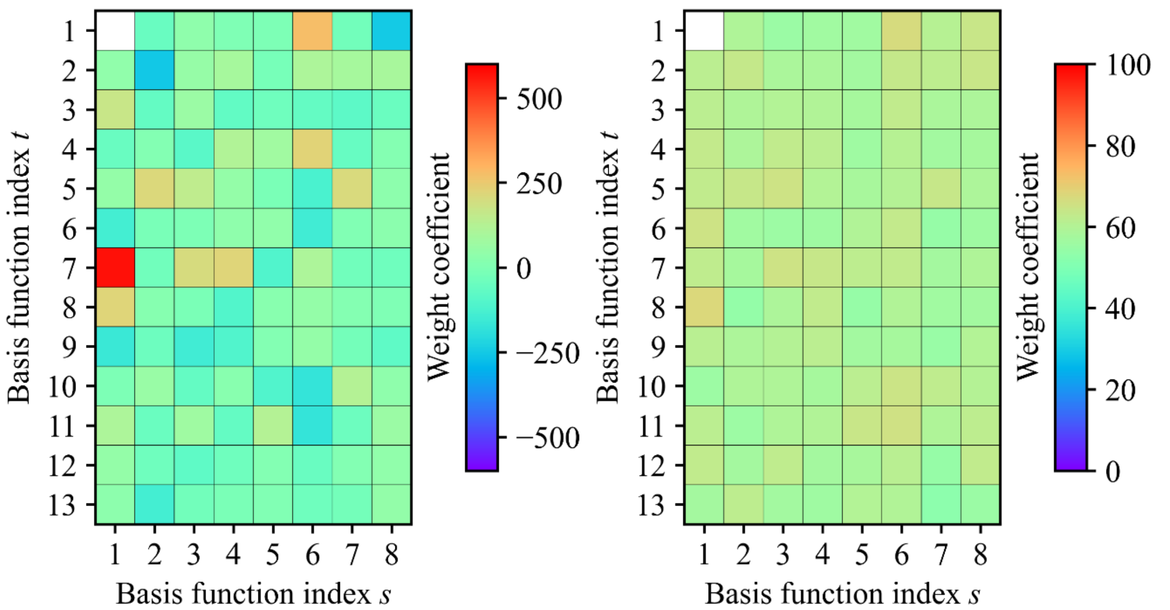



Figure 9. Statistics of DCT spectrum in cross-section B




(a) Mean of DCT spectrum (b) SD of DCT spectrum

Figure 10. Statistics of DCT spectrum in cross-section C

Figures 8a and 8b show the statistics of DCT spectrum for cross-sections A in the form of colored meshes. As shown in Figure 8a, 36 weight coefficients are identified from sparse data using BCS, and they are characterized by a 6×6 matrix of DCT spectrum. Each mesh is color-coded using the mean value of the weight coefficient with indexes $t$ and $s$, which are indexes (or frequencies) of the corresponding 2D basis functions, as indicated in Equation (2). Note that the $\omega_{1,1}^{2D}$ corresponding to the contribution of the constant basis function $\mathrm{B}_{1,1}(x_1, x_2)$ is much greater than the other coefficients and is not required for the calculation of DCT-based ACF (see Equation (6)). Therefore, the upper left cell in Figure 8 is shown as empty for visualization clarity. It is observed that among these 35 weight coefficients, only a few of them are significant, with the remaining ones close to zeros. Figure 8b shows the SD values of the corresponding weight coefficients in Figure 8a, which reflect the statistical uncertainty of the DCT spectrum. Figures 9 and 10 show the statistics of $N_{\mathrm{B}}$ samples of DCT spectrum for cross-sections B and C, respectively. Figures 8-10 display that the number of weight coefficients, as well as the indexes of significant coefficients, in the approximated DCT spectrum are different in different cross-sections. The DCT spectrums are not directly comparable, since they correspond to cross-sections with different spatial dimensions. DCT-based ACF is subsequently calculated for a unified representation of 2D spatial variability in different cross-sections.

Using Equation (6), $N_B$ samples of DCT-based ACFs are obtained for each cross-section. Figure 11 shows the statistics of the obtained DCT-based ACFs in cross-section A by colormaps. Subplot (a) demonstrates one sample of DCT-based ACF, which portrays a possible representation of the 2D spatial variability patterns of $s_u$ data in cross-section A. The associated non-stationarity and spatial variability anisotropy are represented in DCT-based ACF in a unified manner. Subplots (b) and (c) show the mean and the SD of $N_B$ samples of DCT-based ACFs. Subplot (d) reveals the absolute residual between the mean in Subplot (b) and the original DCT-based ACF in Figure 6g. The mean of DCT-based ACFs in Subplot (b) is interpreted as the best estimate for the cross-section A in the presence of sparse investigation data (e.g., see Figure 7a). Figures 12 and 13 show the statistics of the obtained DCT-based ACFs in cross-sections B and C, respectively, following the same presentation format. It is observed from Figures 11-13 that, for all three cross-sections, the colormaps in Subplots (c) and (d) are generally comparable, indicating the quantified statistical uncertainty of DCT-based ACF is rational.

To quantify the similarity between any two cross-sections with consideration of statistical uncertainty, $N_B$ DCT-based ACF samples from one cross-section are randomly paired with $N_B$ DCT-based ACF samples from another cross-section. For each pair of DCT-based ACFs, a similarity value is calculated using Equation (7), leading to $N_B$ similarity values. Probabilistic cross-sectional similarity quantification is then performed by statistical analysis of these $N_B$ similarity values. Figure 14 shows the obtained similarity values by blue histograms. Figure 14a presents the similarity values between cross-sections A and B. It shows that for cross-sections A and B, the associated histogram of similarity values peaks at a value approaching 1. The mean of the similarity values is calculated as 0.88 and is close to the true similarity value 0.97, which is denoted by a vertical red dashed line in Figure 14a. High cross-sectional similarity between A and B is reasonably quantified from sparse data. Figure 14b presents the similarity values between cross-sections A and C. The mean of the similarity values is calculated as 0.26 which is close to the true value 0.27. Although the correlation lengths of cross-sections A and B are slightly different as indicated in Table 1, the similarity quantification may be dominated by the respective trend functions that have the same frequency. The low cross-sectional similarity between cross-sections A and C is also quantified accurately. In Figure 14, the SD of similarity values can be interpreted as the statistical uncertainty of quantified cross-sectional similarity, which integrates the statistical uncertainty of DCT spectrum from both concerned cross-sections. The SD values are calculated as 0.06 for cross-sections A and B, and 0.08 for cross-sections A and C. It suggests that for the above two comparisons, the associated statistical uncertainty is relatively small. In other words, the three cross-sections, particularly the associated trend functions, may be characterized well by sparse data.

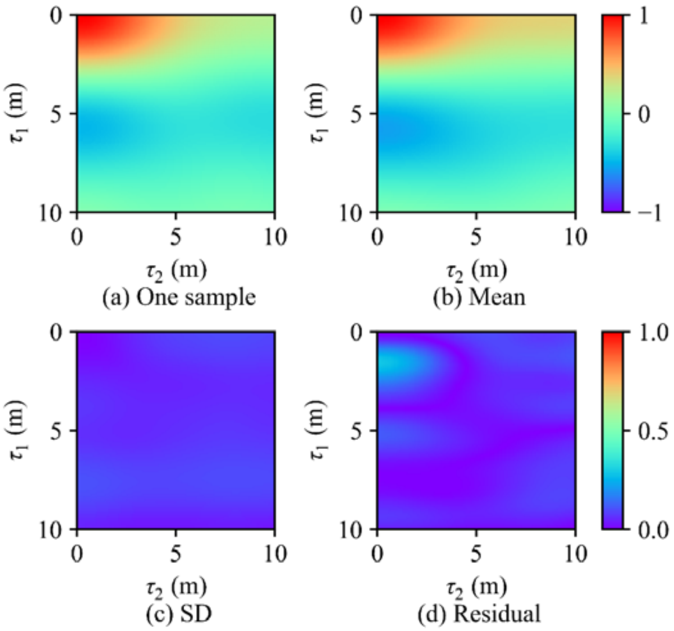



Figure 11. Statistics of DCT-based ACFs in cross-section A

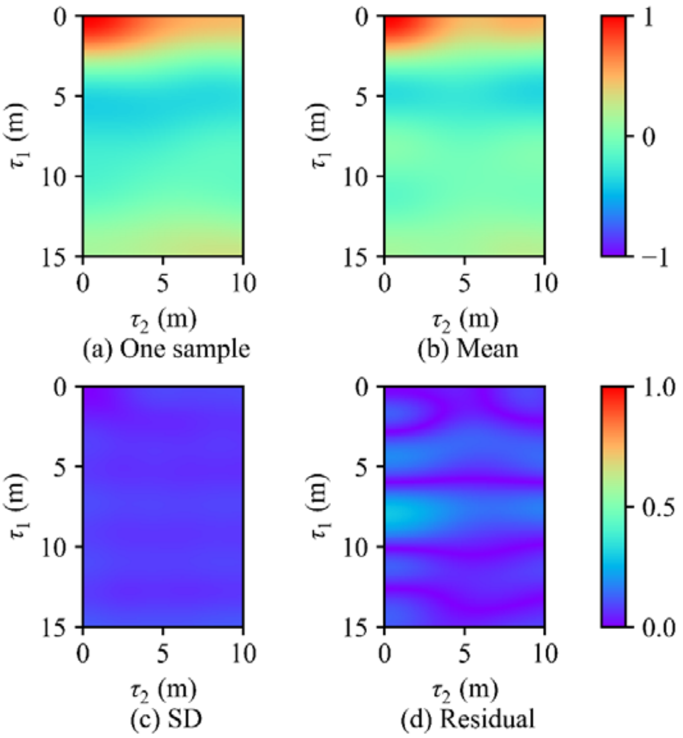



Figure 12. Statistics of DCT-based ACFs in cross-section B

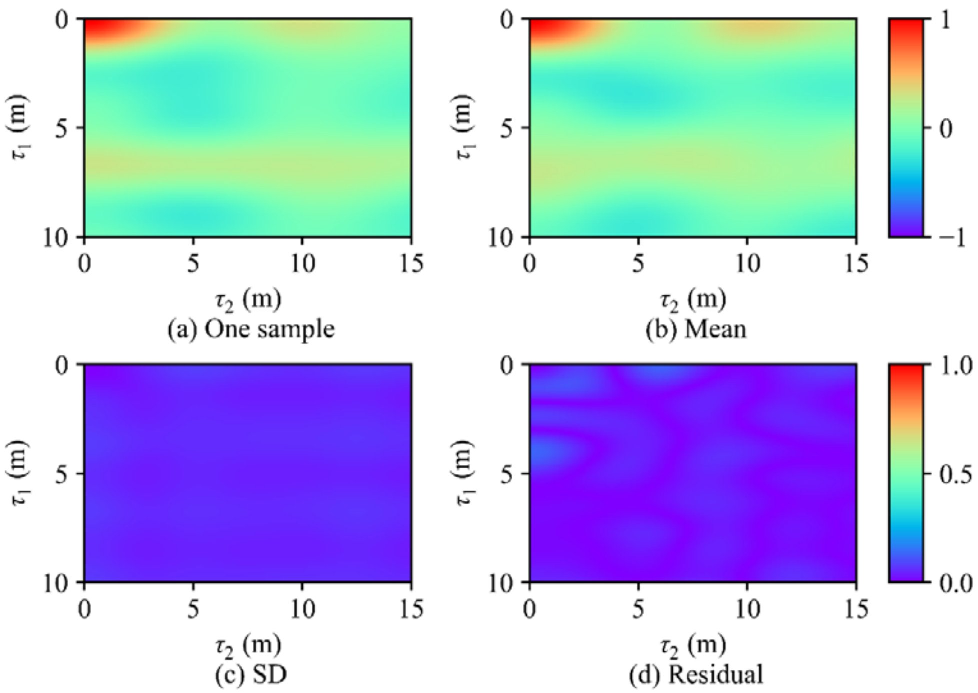



Figure 13. Statistics of DCT-based ACFs in cross-section C

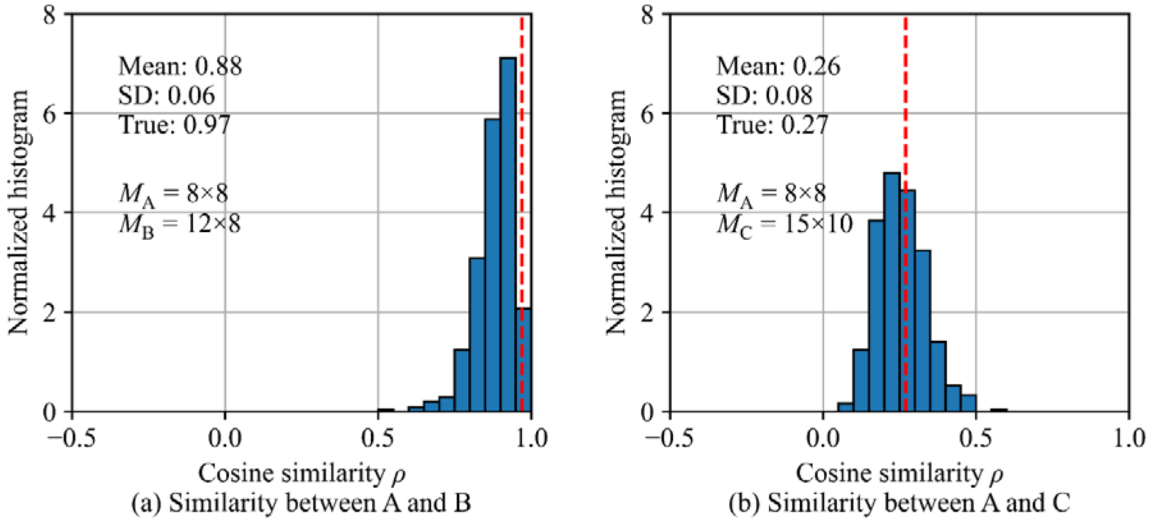



Figure 14. Normalized histogram of cosine similarity values between cross-sections: (a) A and B; (b) A and C

Table 2. Statistics of cross-sectional similarity values obtained from measurement data of cross-sections A, B, and C

| *M* scenario | Cross-sections | A | B | C |
|---|---|---|---|---|
| $M_A$=6×6 | A | 1 | 0.68(0.21) | 0.24(0.16) |
| $M_B$=10×6 | B | - | 1 | 0.35(0.14) |
| $M_C$=10×8 | C | - | - | 1 |
| $M_A$=8×8 | A | 1 | 0.88(0.06) | 0.26(0.08) |
| $M_B$=12×8 | B | - | 1 | 0.47(0.09) |
| $M_C$=15×10 | C | - | - | 1 |
| $M_A$=15×15 | A | 1 | 0.96(0.01) | 0.23(0.04) |
| $M_B$=20×15 | B | - | 1 | 0.36(0.03) |
| $M_C$=20×15 | C | - | - | 1 |

Data format: Mean (Standard deviation)

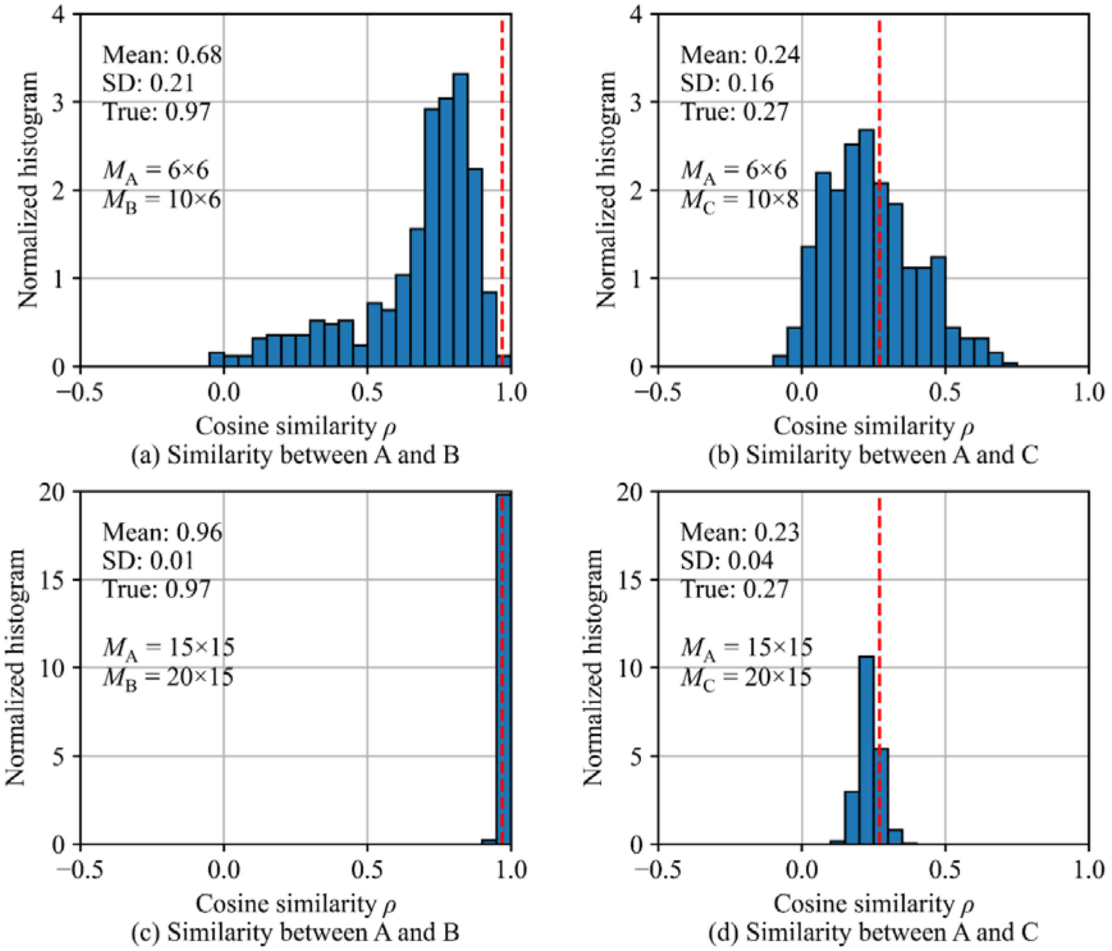



Figure 15. Normalized histogram of cosine similarity values between cross-sections under different measurement scenarios ($M$)

### *5.3. Effect of the number of measured data points*

This subsection investigates the effect of the number, $M$, of measured data points on the performance of cross-sectional similarity quantification. Two more measurement scenarios for the three cross-sections are added. One added scenario has a smaller number of measured $s_u$ data, i.e., $M$=6×6, 10×6, and 10×8 $s_u$ data with a uniform grid sampling are measured in cross-sections A, B, and C, respectively. Another added scenario has a larger number of measured $s_u$ data, i.e., $M$=15×15, 20×15, and 20×15 $s_u$ data are measured in the three cross-sections respectively. For each added scenario, cross-sectional similarity quantifications are performed, following the implementation procedures described in Section 4.

Figure 15 shows the histograms of cross-sectional similarity values for two added scenarios, between A and B, and between A and C, respectively. Figure 15a shows the similarity between cross-sections A and B, when the number of measurement data is relatively small. In comparison to Figure 14a, it is observed that the mean of similarity values decreases significantly from 0.88 to 0.68. In addition, the SD of similarity values increases significantly from 0.08 to 0.21. Figure 15c corresponds to the similarity between cross-sections A and B, when the number of measurement data is relatively large. It shows that the histogram is narrowed down significantly and almost overlaps with the true value. Similar observations are also obtained for similarity values between cross-sections A and C, where the corresponding small and large measurement data number scenarios are shown in Figures 15b and 15d, respectively. For cross-sectional similarity between A and C, it appears that the true similarity value, which is as low as 0.27, can be accurately identified using extremely sparse data. Both high cross-sectional similarity between A and B and low cross-sectional similarity between A and C can be quantified effectively using sparse data. The results of this sensitivity study, as well as the cross-sectional similarity between B and C, are summarized in Table 2. The results indicate that the performance of the proposed method for quantifying cross-sectional similarity depends on the number $M$ of available measured data in corresponding cross-sections. When $M$ is low in the two cross-sections to be compared, the associated statistical uncertainty might become dominant in the subsequent cross-sectional similarity quantification. In this case, additional site investigation might be required to get more measurements and insights into the spatial variability in concerned cross-sections. As $M$ increases, the quantified cross-sectional similarity converges to the true value. Moreover, the proposed method also applies to scenarios where the amount of measured data differs significantly in the two cross-sections, e.g., one cross-section characterized with limited data while another characterized with much more data. This enables the proposed method to be performed in a data-driven manner in practical site investigation. Note that the relationship between the number $M$ of available measurement data and the cross-sectional similarity is problem-specific and might not necessarily be a general one that can possibly be applied to other cross-sections.

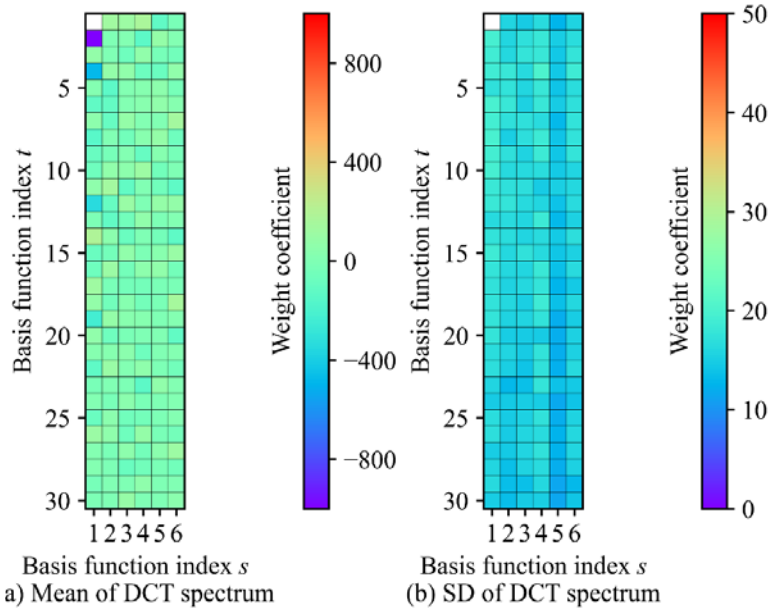



Figure 16. Statistics of DCT spectrum at Site 1

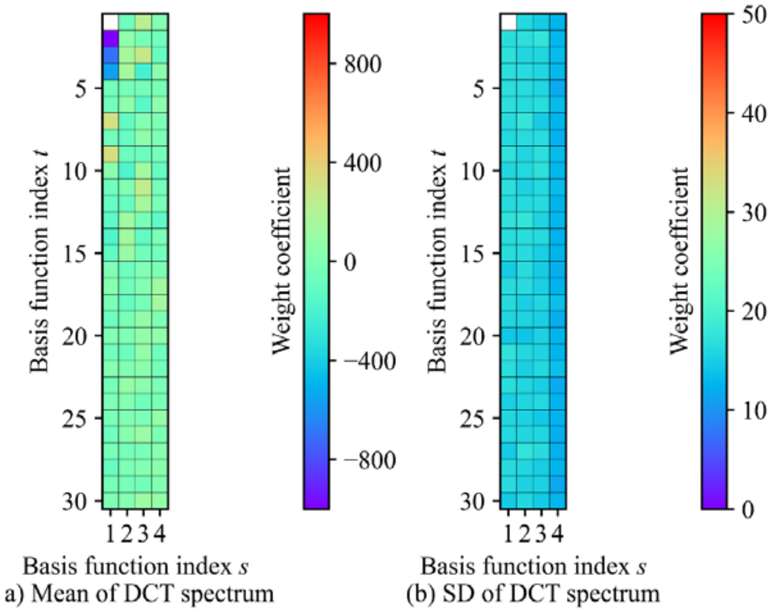



Figure 17. Statistics of DCT spectrum at Site 2

## 6. Real examples

This section demonstrates an application of the proposed method to the real examples shown in Figure 1. Probabilistic quantification of cross-sectional similarity between Site 1 and Site 2 is performed, following the implementation procedures in Section 4. Sites 1 and 2 have a length of 145m and 245m, respectively, and they are both 20m deep. In step 1, a vertical resolution of 0.05m and a horizontal resolution of 0.5m are adopted to discretize the cross-sections at the two sites, leading to a cross-section image with a size of 400×290 for Site 1, and a cross-section image with a size of 400×490 for Site 2. In step 2, the CPT data (e.g., corrected cone resistance $q_t$ in this example) within these two cross-sections are obtained. As shown in Figures 2a and 2b, eight $q_t$ data profiles are within cross-section at Site 1 and seven profiles are within cross-section at Site 2. In step 3, $N_B$=500 samples of DCT spectrum are generated from $q_t$ data for each site using BCS. Figures 16 and 17 show the statistics of DCT spectrum at Site 1 and Site 2, respectively. It is seen that for both sites, the numbers of identified weight coefficients are different, with 179 coefficients for Site 1 and 119 coefficients for Site 2. The indexes of significant coefficients for the two sites are also apparently different. In step 4, $N_B$ DCT-based ACFs are calculated based on $N_B$ samples of DCT spectrum for both sites. Figures 18a and 18b show the mean and SD of $N_B$ samples of DCT-based ACFs at Site 1. Figures 19a and 19b show the corresponding results at Site 2. It is evident that the means of DCT-based ACFs in Figures 18a and 19a have generally consistent patterns, i.e., predominant spatial variability patterns along vertical directions and relatively minor variability patterns along horizontal directions. The SD maps in Figures 18b and 19b show similar magnitudes, suggesting the statistical uncertainty for these two sites is comparable. In step 5, to perform a probabilistic quantification of cross-sectional similarity, $N_B$ DCT-based ACFs at Site 1 are randomly paired with $N_B$ DCT-based ACFs at Site 2. Since the horizontal lengths are different for these two sites (i.e., 145m for Site 1 and 245m for Site 2), the associated DCT-based ACFs of these two sites have different dimensions, as shown in Figures 18 and 19. Therefore, the overlapped sections of Sites 1 and 2, denoted by red dashed lines in Figures 18a and 19a, are used for cross-sectional similarity quantification. Using $N_B$ pairs of DCT-based ACFs and Equation (6), $N_B$ similarity values are obtained, which are presented by a histogram in Figure 20. Note that the histogram of similarity values is narrow and mainly located at the $\rho$ range of about 0.97 to 0.98. The mean and SD of the $N_B$ similarity values are calculated as 0.977 and 0.002, respectively. According to the results, the proposed method suggests that Sites 1 and 2 are highly similar, and the associated statistical uncertainty is insignificant. Although the numbers of CPTs soundings are sparse from both sites, the spatial variability patterns of $q_t$ data are prominent and consistent. The increasing trends of $q_t$ data profiles at both sites, as shown in Figure 2 are comparable and properly identified. A systematic study on spatial variability with increasing trend functions is worth exploring in a future study to clearly demonstrate generalizability of the proposed method. In addition, the results indicate that the proposed method performs well even for cross-sections with non-uniformly measured data. This scenario can be regarded as incomplete data, because any non-uniform measurement grid can be derived from a uniform measurement grid by removing measurements from selected points. Hence, this scenario refers to one aspect of MUSIC, which is "I" for incomplete data.

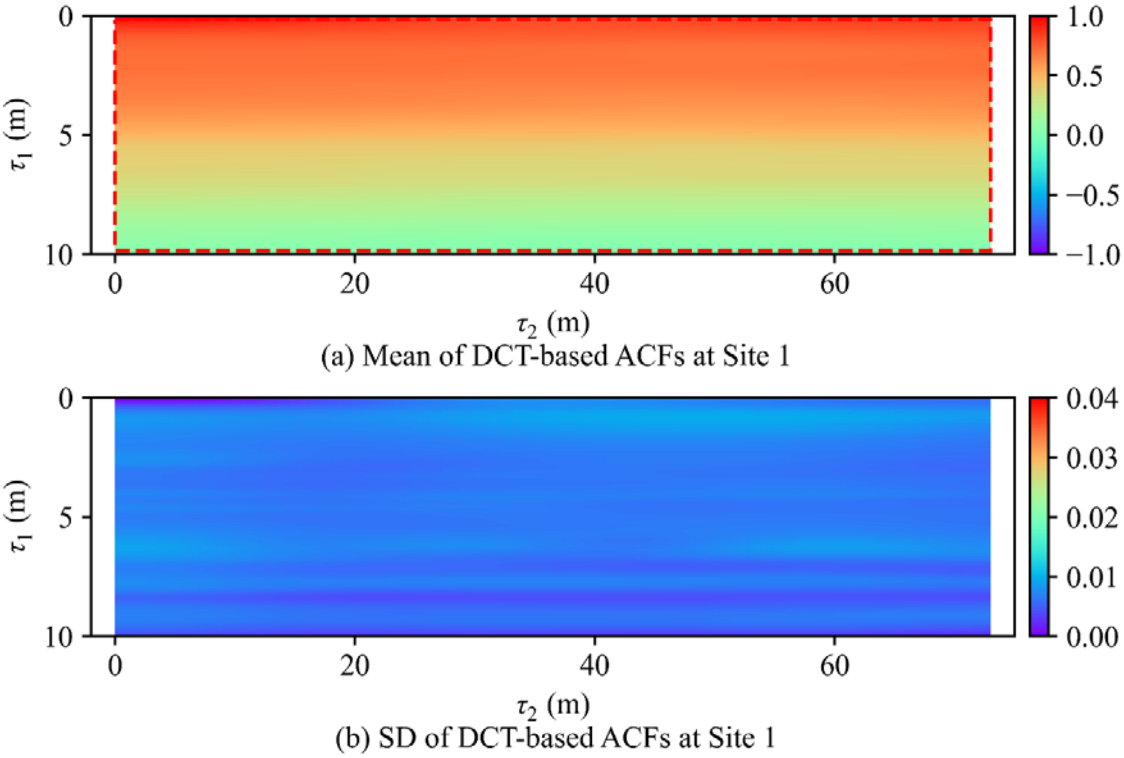



Figure 18. Statistics of DCT-based ACFs at Site 1

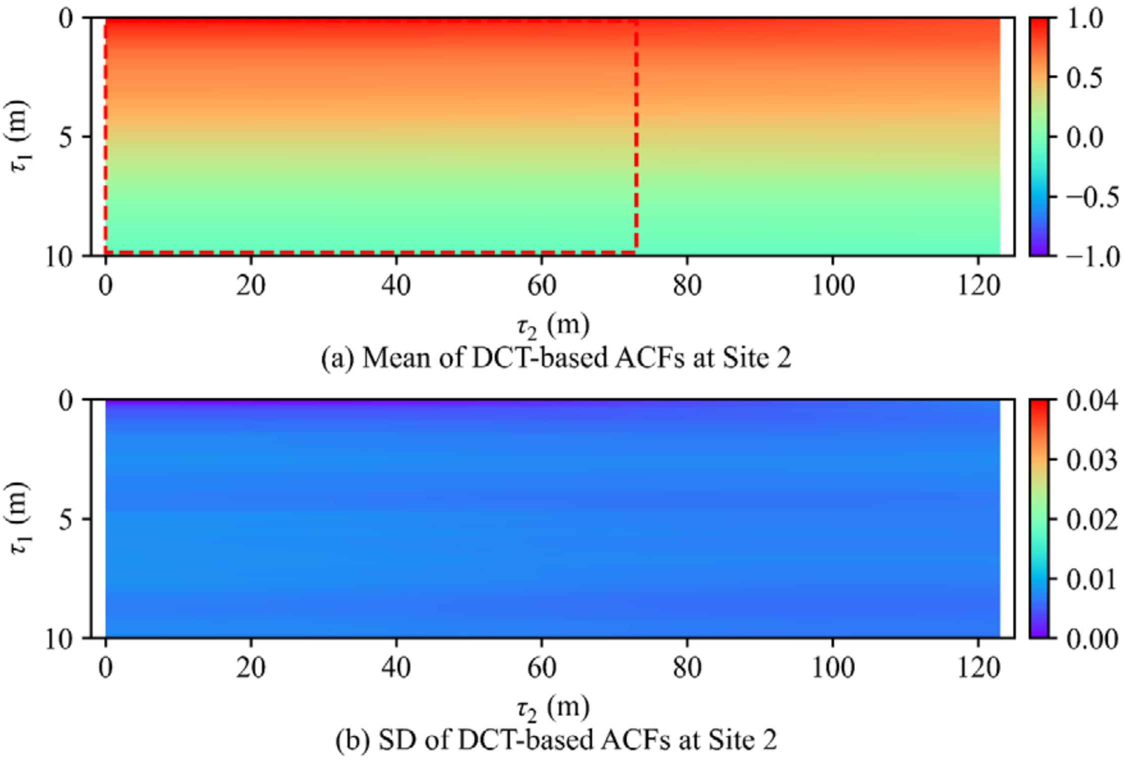



Figure 19. Statistics of DCT-based ACFs at Site 2

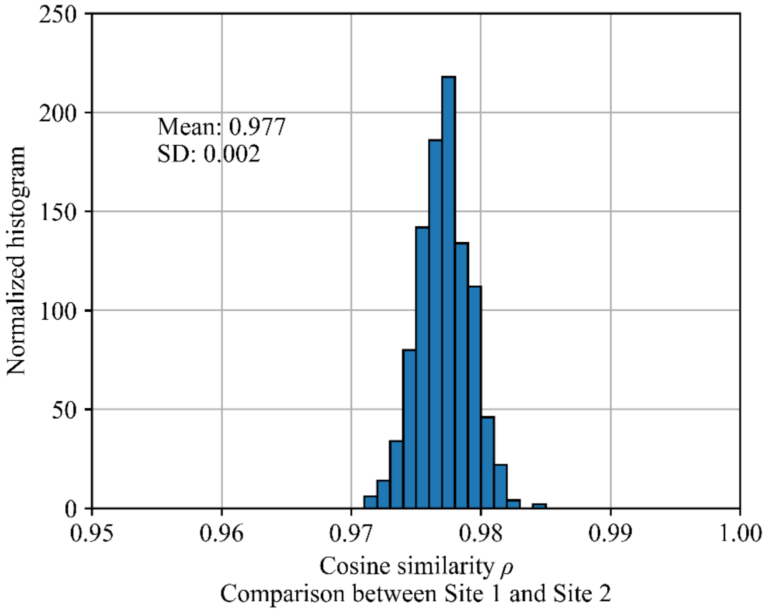
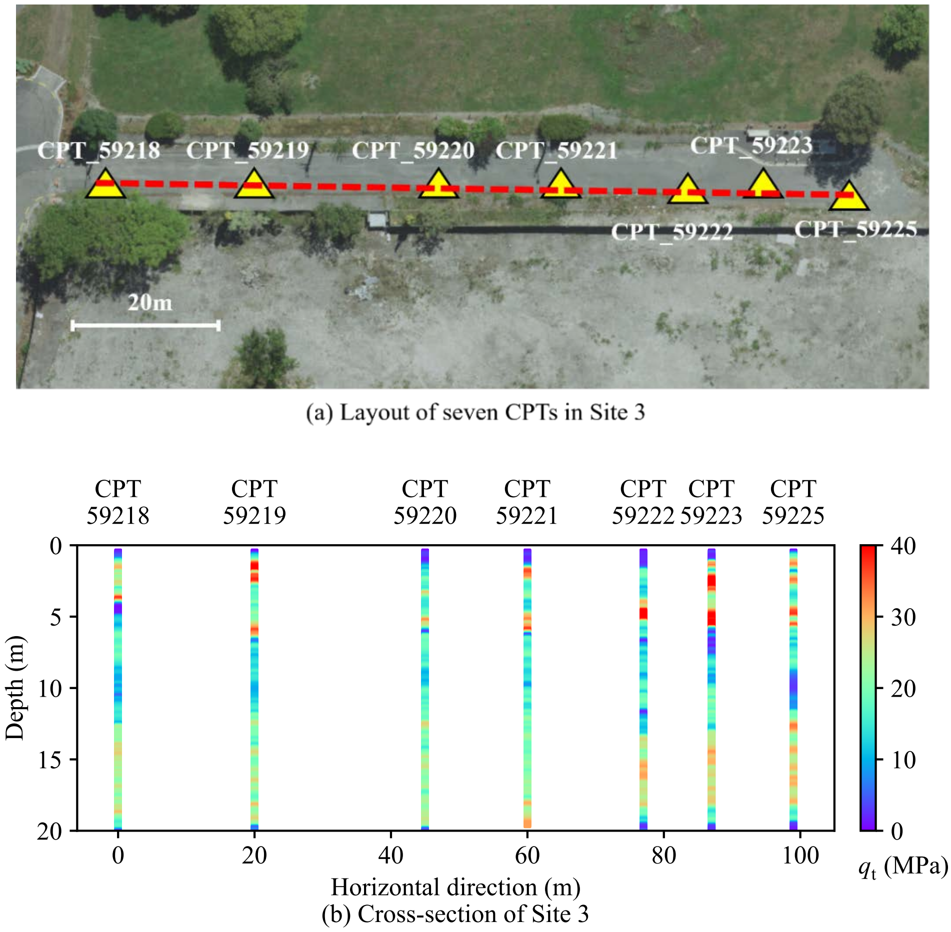



Figure 20. Normalized histogram of cosine similarity values between Sites 1 and 2 in the real data example




(a) Layout of seven CPTs in Site 3

(b) Cross-section of Site 3

Figure 21. Cone penetration tests (CPTs) performed in Site 3: (a) layout map of seven CPTs; (b) Cross-section of seven CPTs

To further demonstrate the performance of the proposed method, another cross-section example at Site 3 is compared with the cross-section at Site 1. As shown in Figure 21a, Site 3 is a cross-section with a horizontal length of around 100m, and seven CPTs were performed in this cross-section. In contrast to Site 2, which is only approximately 700m away from Site 1, Site 3 is relatively far from Site 1, and they are roughly 5km apart. In view of this spatial distance, CPT data at Site 3 might exhibit different spatial variability patterns from that of Site 1. Figure 21b shows the $q_t$ data profiles in this cross-section. Figure 2a and Figure 21b are visually different. Probabilistic quantification of cross-sectional similarity between Sites 1 and 3 is performed, following the implementation procedures described in Section 4. After configuring the same spatial resolutions for Site 3 as Site 1 (i.e., vertical resolution of 0.05m and a horizontal resolution of 0.5m), $N_B$=500 samples of DCT spectrum are generated for Site 3 using BCS. Subsequently, DCT-based ACFs are calculated for Site 3. Figures 22a and 22b show the mean and SD of DCT-based ACFs at Site 3, respectively. It shows that the mean of DCT-based ACFs in Figure 22a differs significantly from Figure 18a. The ACF at Site 3 decays and fluctuates faster than the ACF at Site 1 along the depth direction. This suggests that the correlation length along the depth direction at Site 3 is much smaller than that at Site 1. In addition, the SD map in Figure 22b shows a higher magnitude than the SD map in Figure 18b, suggesting the statistical uncertainty of $q_t$ data at Site 3 is greater than Site 1. Then, $N_B$ DCT-based ACFs at Site 3 are randomly paired with $N_B$ DCT-based ACFs at Site 1. The resulting $N_B$ similarity values are calculated and shown in Figure 23. The mean and SD of the $N_B$ similarity values are calculated as 0.036 and 0.008, respectively. The difference between Figure 20 (similarity between Sites 1 and 2) and Figure 23 (similarity between Sites 1 and 3) is stark. According to the results, the proposed method suggests that Sites 3 and 1 are not similar.

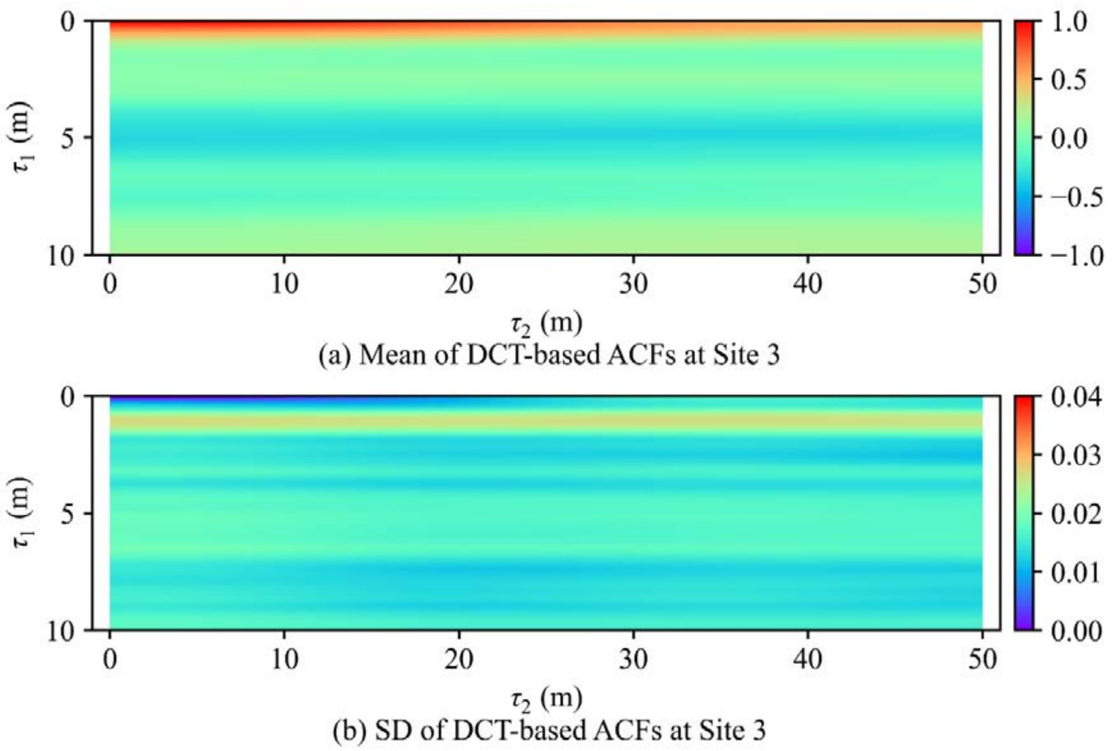



Figure 22. Statistics of DCT-based ACFs at Site 3

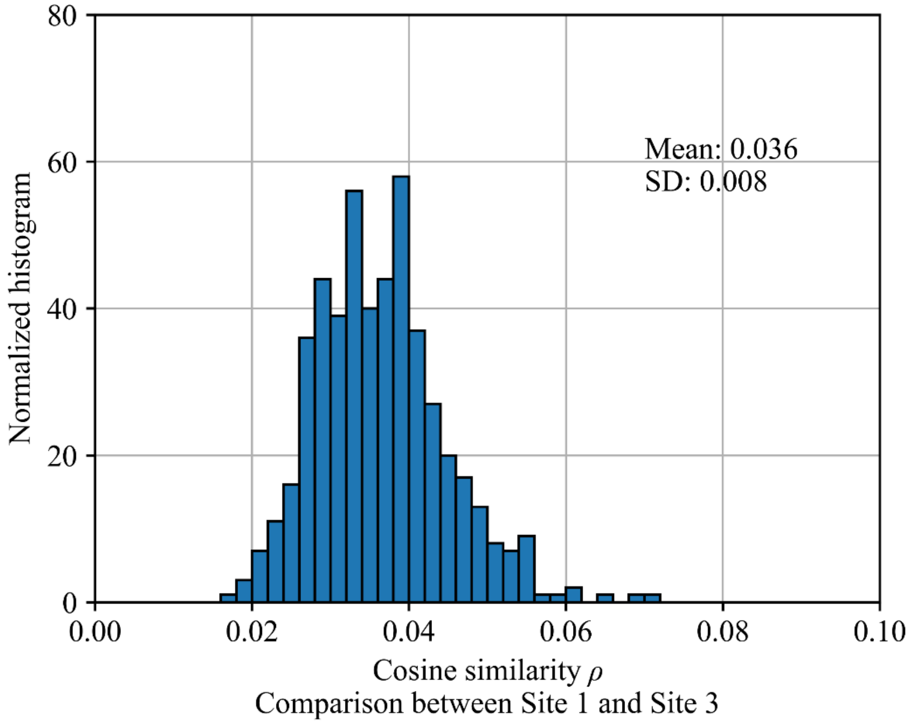



Figure 23. Normalized histogram of cosine similarity values between Site 1 and Site 3

## 7. Conclusions

In this study, a novel data-driven method was proposed to quantify 2D cross-sectional similarity based on the spatial variability of one soil property. To the best of the authors' knowledge, the proposed method is the first method to quantify geotechnical site similarity with explicit consideration of 2D spatial variability. It tackled the challenges of sparse investigation data, non-stationary spatial variability, and inconsistent spatial dimensions of different 2D cross-sections. A unified representation framework of 2D spatial variability using DCT-based ACF was developed. Cross-sectional similarity was quantified by the similarity of DCT-based ACFs between cross-sections. For a given 2D cross-section, BCS was adopted to approximate the DCT spectrum directly from sparse investigation data. The associated statistical uncertainty was also quantified by simulation of many random samples of the DCT spectrum. Samples of DCT-based ACF were then calculated using random samples of DCT spectrum and the newly derived equations. Then, cross-sectional similarity was quantified by the cosine similarity of DCT-based ACFs between two cross-sections. Theoretical derivation in the Appendix suggested that DCT-based ACF is an effective surrogate to represent 2D cross-sectional spatial variability from a spectral perspective and enables direct pattern comparison between cross-sections with different spatial dimensions.

Numerical examples of three soil property cross-sections were provided to illustrate the performance of the proposed method. The similarity between any two of the three cross-sections was quantified. Results indicated that the proposed method rationally quantifies the cross-sectional similarity and associated statistical uncertainty from sparse data in a data-driven manner. The quantified similarity values converged to the true value when the number of measured data increases. Real data examples from New Zealand were also used to demonstrate the application of the proposed method. High cross-sectional similarity was obtained between two sites which are approximately 700m apart. The proposed method also suggested low similarity between two sites which are 5km apart. The similarity quantification developed in this study assists engineering geologists and geotechnical engineers with an efficient identification of similar project sites and informed decision-making in site characterization. Geotechnical experiences may be effectively shared between the identified similar sites. In practice, the proposed method might also be implemented with masked geographical coordinate information, since such information may be restricted for confidentiality reasons. The proposed method can identify a quasi-regional cluster of CPT soundings in a global database that is more relevant to understanding the spatial variability of a soil property at a target site. It relieves the engineer from sole reliance on subjective judgment and tedious manual visual inspection to complete the same task.

# Appendix: Derivation of Equation (6)

Equation (3) suggests that the patterns of basis function $\mathrm{B}_{t,s}(x_1, x_2)$ and corresponding weight coefficient $\omega_{t,s}^{2D}$ jointly control the patterns of $\mathrm{F}(x_1, x_2)$. To investigate the effect of basis function $\mathrm{B}_{t,s}(x_1, x_2)$ patterns on the $\mathrm{F}(x_1, x_2)$ patterns when DCT basis function is adopted, the ACF of $\mathrm{B}_{t,s}(x_1, x_2)$ is firstly derived based on Equations (1), (2), and (5) (e.g., Shinozuka and Deodatis, 1991; Vanmarcke, 2010):

$$\mathrm{ACF}\left[\mathrm{B}_{t,s}(x_1, x_2), \tau_1, \tau_2\right] = \frac{E\left\{\left[\mathrm{B}_{t,s}^{2D}(x_1, x_2) - \mu_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}\right]\left[\mathrm{B}_{t,s}^{2D}(x_1+\tau_1, x_2+\tau_2) - \mu_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}\right]\right\}}{\sigma^2_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}} \tag{A.1}$$

in which $\mu_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}$ and $\sigma_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}$ are the mean value and standard deviation of $\mathrm{B}_{t,s}(x_1, x_2)$, respectively. It is seen from Equation (1) and Figure 4 that, when $t = s = 1$, $\mu_{\mathrm{B}_{1,1}^{2D}(x_1,x_2)}$ is a constant function, for which ACF is undefined. For $t \geq 2$, the mean value $\mu_{\mathrm{B}_t(x)}$ is zero because of the nature of cosine function:

$$\int_0^1 cos(t\pi)idi = 0 \quad \big(for\ t \in \mathbb{Z}, t \geq 2, and\ \ i \in (0,1)\big) \tag{A.2}$$

Therefore, for $t, s \neq 1$ concurrently, the $\mu_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}$ is reduced to zero according to the definition of $\mathrm{B}_{t,s}(x_1, x_2)$ in Equation (2). Equation (A.1) is then rewritten as:

$$\begin{aligned} &\mathrm{ACF}\big[\mathrm{B}_{t,s}(x_1, x_2), \tau_1, \tau_2\big] \\ &= \frac{1}{\sigma^2_{\mathrm{B}_{t,s}^{2D}(x_1,x_2)}} E\big\{\big[\mathrm{B}_{t,s}^{2D}(x_1, x_2)\big]\big[\mathrm{B}_{t,s}^{2D}(x_1{+}\tau_1, x_2{+}\tau_2)\big]\big\} \end{aligned} \tag{A.3}$$

Since the 2D basis function $\mathrm{B}_{t,s}^{2D}(x_1, x_2)$ is constructed by a tensor product of two 1D DCT basis functions along two orthonormal directions, respectively (see Equation (2)), two 1D DCT basis functions are independent of each other. $\mathrm{B}_{t,s}^{2D}(x_1, x_2)$ and $\mathrm{B}_{t,s}^{2D}(x_1{+}\tau_1, x_2{+}\tau_2)$ hence can be factorized, and Equation (A.3) is rewritten as:

$$\begin{aligned} &\mathrm{ACF}\big[\mathrm{B}_{t,s}(x_1, x_2), \tau_1, \tau_2\big] \\ &= \frac{1}{\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}} E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1 + \tau_1)]E[\mathrm{B}_s(x_2)\mathrm{B}_s(x_2 + \tau_2)] \end{aligned} \tag{A.4}$$

The first expectation term in Equation (A.4) for $x_1$ direction can be derived, after substituting 1D DCT basis function from Equation (1) to Equation (A.4):

$$\begin{aligned} &E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1 + \tau_1)] \\ &= E\left[\sqrt{\frac{2}{N_1}} cos\pi(t-1)\frac{(x_1 - 0.5)}{N_1}\sqrt{\frac{2}{N_1}} cos\pi(t-1)\frac{(x_1 + \tau_1 - 0.5)}{N_1}\right] \end{aligned} \tag{A.5}$$

Using product-to-sum identity, the product of two cosine functions can be rewritten as:

$$
\begin{aligned}
&E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1+\tau_1)] \\
&= \frac{1}{N_1}\left\{E\left[cos\pi(t-1)\frac{(2x_1+\tau_1-1)}{N_1}\right] + E\left[cos\pi(t-1)\frac{\tau_1}{N_1}\right]\right\}
\end{aligned} \tag{A.6}
$$

The first expectation term with index $x_1$ is reduced to zero after expectation operation for $x_1$. Therefore, Equation (A.6) is derived as a cosine function of $\tau_1$ at frequency $\pi(t-1)$:

$$
E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1+\tau_1)] = \frac{1}{N_1}cos\pi(t-1)\frac{\tau_1}{N_1} \tag{A.7}
$$

In a similar fashion, the second expectation in Equation (A.4) for $x_2$ direction can be derived as:

$$
E[\mathrm{B}_s(x_2)\mathrm{B}_s(x_2+\tau_2)] = \frac{1}{N_2}cos\pi(s-1)\frac{\tau_2}{N_2} \tag{A.8}
$$

Therefore, substituting Equations (A.7) and (A.8) into Equation (A.4) leads to:

$$
\begin{aligned}
&\mathrm{ACF}\left[\mathrm{B}_{t,s}(x_1,x_2),\tau_1,\tau_2\right] \\
&= \frac{1}{\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}}E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1+\tau_1)]E[\mathrm{B}_s(x_2)\mathrm{B}_s(x_2+\tau_2)] \\
&= \frac{1}{\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}}\frac{1}{N_1N_2}cos\pi(t-1)\frac{\tau_1}{N_1}cos\pi(s-1)\frac{\tau_2}{N_2}
\end{aligned} \tag{A.9}
$$

Note that the variance term $\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}$ can be derived based on the fact that the 2D DCT basis function is zero-mean and orthonormal (e.g., Rao and Yip, 1990). The Frobenius norm of 2D basis function $\mathrm{B}_{t,s}(x_1,x_2)$ is unity:

$$
\left\|\mathrm{B}_{t,s}(x_1,x_2)\right\| = \sqrt{\mathrm{B}_{t,s}(1,1)^2 + \mathrm{B}_{t,s}(2,1)^2 + \cdots + \mathrm{B}_{t,s}(N_1,N_2)^2} = 1 \tag{A.10}
$$

According to the definition of variance, the $\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}$ is derived as:

$$\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)} = \frac{1}{N_1N_2}\left[\mathrm{B}_{t,s}(1,1)^2 + \mathrm{B}_{t,s}(2,1)^2 + \cdots + \mathrm{B}_{t,s}(N_1,N_2)^2\right] = \frac{1}{N_1N_2} \tag{A.11}$$

Therefore, substituting Equation (A.11) into Equation (A.9) leads to the normalized ACF of 2D basis function:

$$\begin{aligned}
&\mathrm{ACF}\left[\mathrm{B}_{t,s}(x_1,x_2),\tau_1,\tau_2\right] \\
&= \frac{1}{\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}} E[\mathrm{B}_t(x_1)\mathrm{B}_t(x_1+\tau_1)]E[\mathrm{B}_s(x_2)\mathrm{B}_s(x_2+\tau_2)] \\
&= cos\pi(t-1)\frac{\tau_1}{N_1}cos\pi(s-1)\frac{\tau_2}{N_2} \qquad (t,s \neq 1\ concurrently)
\end{aligned} \tag{A.12}$$

Equation (A.12) shows that the ACF of a 2D DCT basis function $\mathrm{B}_{t,s}(x_1,x_2)$ is derived as the product of two cosine functions along $x_1$ and $x_2$ directions, respectively. These two cosine functions are of lag distances $\tau_1$ and $\tau_2$, respectively. Note that the indexes $x_1$ and $x_2$ are eliminated in Equation (A.12).

Based on Equations (2), (3), and (5), the ACF of 2D spatial variability $\mathrm{F}(x_1,x_2)$ can also be derived under DCT framework:

$$\begin{aligned}
&\mathrm{ACF}[\mathrm{F}(x_1,x_2),\tau_1,\tau_2] \\
&= \frac{E\left\{\left[\mathrm{F}(x_1,x_2) - \mu_{\mathrm{F}(x_1,x_2)}\right]\left[\mathrm{F}(x_1+\tau_1,x_2+\tau_2) - \mu_{\mathrm{F}(x_1,x_2)}\right]\right\}}{\sigma^2_{\mathrm{F}(x_1,x_2)}}
\end{aligned} \tag{A.13}$$

in which $\mu_{\mathrm{F}(x_1,x_2)}$ and $\sigma_{\mathrm{F}(x_1,x_2)}$ are mean value and variance of $\mathrm{F}(x_1,x_2)$. Note that $\mu_{\mathrm{F}(x_1,x_2)}$ can be replaced by the contribution of $\mathrm{B}_{1,1}(x_1,x_2)$, since $\mathrm{B}_{1,1}(x_1,x_2)$ is the only 2D DCT basis function with a non-zero mean value (see upper left 2D basis function in Figure 4). Substituting Equation (3) into Equation (5) leads to:

$$\mathrm{ACF}[\mathrm{F}(x_1,x_2),\tau_1,\tau_2] = \frac{E\left\{\left[\sum_{t=1}^{N_1}\sum_{s=1}^{N_2}\omega_{t,s}^{2D}\mathrm{B}_{t,s}(x_1,x_2)\right]\left[\sum_{t'=1}^{N_1}\sum_{s'=1}^{N_2}\omega_{t',s'}^{2D}\mathrm{B}_{t',s'}(x_1+\tau_1,x_2+\tau_2)\right]\right\}}{\sigma_{\mathrm{F}(x_1,x_2)}^2} \quad (t,s\neq 1\ concurrently;\ t',s'\neq 1\ concurrently) \tag{A.14}$$

The prime symbols are used to distinguish two factors along an individual direction. By utilizing the orthonormal property of $\mathrm{B}_{t,s}(x_1,x_2)$, Equation (A.14) can be expanded and rearranged as a multiple summation:

$$\mathrm{ACF}[\mathrm{F}(x_1,x_2),\tau_1,\tau_2] = \frac{1}{\sigma_{\mathrm{F}(x_1,x_2)}^2}\sum_{t=1}^{N_1}\sum_{s=1}^{N_2}\sum_{t'=1}^{N_1}\sum_{s'=1}^{N_2}\omega_{t,s}^{2D}\omega_{t',s'}^{2D}E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)]E[\mathrm{B}_s(x_2)\mathrm{B}_{s'}(x_2+\tau_2)] \quad (t,s\neq 1\ concurrently;\ t',s'\neq 1\ concurrently) \tag{A.15}$$

The expectation operations are performed for $x_1$ and $x_2$ directions, separately. For $x_1$ direction, after substituting 1D DCT basis function in Equation (1), the expectation operation $E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)]$ can be further expressed as:

$$E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)] = \begin{cases} E\left[\sqrt{\frac{2}{N_1}}\sqrt{\frac{2}{N_1}}cos\pi\frac{(t'-1)(x_1+\tau_1-0.5)}{N_1}\right]=0 & when\ t=1,t'\neq 1 \\ E\left[\sqrt{\frac{2}{N_1}}\sqrt{\frac{2}{N_1}}cos\pi\frac{(t-1)(x_1-0.5)}{N_1}\right]=0 & when\ t\neq 1,t'=1 \\ E\left[\begin{array}{c}\sqrt{\frac{2}{N_1}}\sqrt{\frac{2}{N_1}}cos\pi\frac{(t-1)(x_1-0.5)}{N_1} \\ cos\pi\frac{(t'-1)(x_1+\tau_1-0.5)}{N_1}\end{array}\right] & when\ t\neq 1,\ t'\neq 1 \end{cases} \tag{A.16}$$

Equation (A.16) consists of three scenarios, i.e., when $t=1, t'\neq 1$; $t\neq 1, t'=1$; and $t\neq 1$, $t'\neq 1$. Since the first two scenarios lead to single cosine functions of $x_1$, the associated terms are reduced to zeros after expectation operation on $x_1$. When $t\neq 1$, $t'\neq 1$, the expectation of product of two cosine functions yields zero when the frequencies of two basis function are not equal, i.e., $t \neq t'$. Only the product of two cosine functions with equal frequencies remain, i.e., $t = t'$:

$$E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)] = \\ = \begin{cases} 0 & when\ t \neq t' \\ \dfrac{2}{N_1} E\left[cos\pi \dfrac{(t-1)(x_1-0.5)}{N_1} cos\pi \dfrac{(t-1)(x_1+\tau_1-0.5)}{N_1}\right] & when\ t = t' \end{cases} \tag{A.17}$$

When $t = t'$, the product of two cosine functions can be rewritten using product-to-sum identity as:

$$E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)] \\ = \frac{1}{N_1} E\left[cos\pi \frac{2(t-1)(x_1-0.5)+(t-1)\tau_1}{N_1} + cos\pi(t-1)\frac{\tau_1}{N_1}\right] \tag{A.18}$$

Similarly, in Equation (A.18), the first cosine function term reduces to zeros after expectation operation on $x_1$. Therefore, Equation (A.18) is then derived as:

$$E[\mathrm{B}_t(x_1)\mathrm{B}_{t'}(x_1+\tau_1)] = \frac{1}{N_1} cos\pi(t-1)\frac{\tau_1}{N_1} \qquad when\ t = t' \tag{A.19}$$

In a similar fashion, the second expectation term in Equation (A.15) for $x_2$ direction is derived as:

$$E[\mathrm{B}_s(x_2)\mathrm{B}_{s'}(x_2+\tau_2)] = \frac{1}{N_2} cos\pi(s-1)\frac{\tau_2}{N_2} \qquad when\ s = s' \tag{A.20}$$

After substituting Equations (A.19) and (A.20) into Equation (A.15), Equation (A.15) can be rewritten as:

$$\mathrm{ACF}[\mathrm{F}(x_1, x_2), \tau_1, \tau_2]$$
$$= \frac{1}{N_1 N_2} \frac{1}{\sigma^2_{\mathrm{F}(x_1,x_2)}} \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} {\omega^{2D}_{t,s}}^2 \cos\pi(t-1)\frac{\tau_1}{N_1} \cos\pi(s-1)\frac{\tau_2}{N_2} \quad (t, s \neq 1 \ concurrently) \tag{A.21}$$

The variance term $\sigma^2_{\mathrm{F}(x_1,x_2)}$ can also be derived since the 2D DCT basis function $\mathrm{B}_{t,s}(x_1, x_2)$ is orthonormal and independent of each other. By definition, $\sigma^2_{\mathrm{F}(x_1,x_2)}$ is expressed as:

$$\sigma^2_{\mathrm{F}(x_1,x_2)} = E\left\{\left[\mathrm{F}(x_1, x_2) - \mu_{\mathrm{F}(x_1,x_2)}\right]^2\right\} = \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} {\omega^{2D}_{t,s}}^2 \sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)} \tag{A.22}$$

Note that in Equation (A.11), $\sigma^2_{\mathrm{B}_{t,s}(x_1,x_2)}$ is derived as $\frac{1}{N_1 N_2}$. Therefore, Equation (A.22) can be rewritten as

$$\sigma^2_{\mathrm{F}(x_1,x_2)} = \frac{1}{N_1 N_2} \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} {\omega^{2D}_{t,s}}^2 \tag{A.23}$$

Combining Equations (A.23) and (A.21) yields the normalized DCT-based ACF of $\mathrm{F}(x_1, x_2)$, which is provided as Equation (6) in the main text:

$$\mathrm{ACF}[\mathrm{F}(x_1, x_2), \tau_1, \tau_2]$$
$$= \frac{1}{\sum_{t=1}^{N_1} \sum_{s=1}^{N_2} {\omega^{2D}_{t,s}}^2} \sum_{t=1}^{N_1} \sum_{s=1}^{N_2} {\omega^{2D}_{t,s}}^2 \, \mathrm{ACF}\left[\mathrm{B}_{t,s}(x_1, x_2), \tau_1, \tau_2\right] \quad (t, s \neq 1 \ concurrently) \tag{A.24}$$

## References


Baecher, G.B. & Christian, J.T. 2003. Reliability and statistics in geotechnical engineering. John Wiley & Sons.

Blumensath, T. & Davies, M. 2006. Sparse and shift-Invariant representations of music. IEEE Transactions on Audio, Speech, and Language Processing, 14, 50-57, doi: 10.1109/TSA.2005.860346.

Brockwell, P.J., Davis, R.A. 1991. Stationary ARMA Processes. In: Time Series: Theory and Methods. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-0320-4_3

Candès, E.J., Romberg, J.K. & Tao, T. 2006. Stable signal recovery from incomplete and inaccurate measurements. Communications on pure and applied mathematics, 59, 1207-1223.

Candès, E.J. & Wakin, M.B. 2008. An introduction to compressive sampling. IEEE Signal Processing Magazine, 25, 21-30.

Ching, J. & Phoon, K.-K. 2017. Characterizing Uncertain Site-Specific Trend Function by Sparse Bayesian Learning. Journal of Engineering Mechanics, 143, 04017028, doi: doi:10.1061/(ASCE)EM.1943-7889.0001240.

Ching, J., Phoon, K.-K., Beck, J.L. & Huang, Y. 2017. Identifiability of Geotechnical Site-Specific Trend Functions. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 3, 04017021, doi: doi:10.1061/AJRUA6.0000926.

Ching, J., Huang, W.-H. & Phoon, K.-K. 2020. 3D Probabilistic Site Characterization by Sparse Bayesian Learning. Journal of Engineering Mechanics, 146, 04020134, doi: doi:10.1061/(ASCE)EM.1943-7889.0001859.

Ching, J. & Phoon, K.-K. 2020. Measuring Similarity between Site-Specific Data and Records from Other Sites. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 6, 04020011, doi: doi:10.1061/AJRUA6.0001046.

Ching, J., Uzielli, M., Phoon, K.-K. & Xu, X. 2023. Characterization of Autocovariance Parameters of Detrended Cone Tip Resistance from a Global CPT Database. Journal of Geotechnical and Geoenvironmental Engineering, 149, 04023090, doi: doi:10.1061/JGGEFK.GTENG-11214.

Ching, J., Yoshida, I. & Phoon, K.-K. 2022. Comparison of trend models for geotechnical spatial variability: Sparse Bayesian Learning vs. Gaussian Process Regression. Gondwana Research, 123, pp.174-183.

Dai, H., Zhang, R. & Beer, M. 2022. A new perspective on the simulation of cross-correlated random fields. Structural Safety, 96, 102201, doi: https://doi.org/10.1016/j.strusafe.2022.102201.

Dong, Y., Sun, Z. & Jia, H. 2006. A cosine similarity-based negative selection algorithm for time series novelty detection. Mechanical Systems and Signal Processing, 20, 1461-1472, doi: https://doi.org/10.1016/j.ymssp.2004.12.006.

Donoho, D.L. 2006. Compressed sensing. IEEE Transactions on Information Theory, 52, 1289-1306.

Einsele, G., Chough, S.K. & Shiki, T. 1996. Depositional events and their records—an introduction. Sedimentary Geology, 104, 1-9, doi: https://doi.org/10.1016/0037-0738(95)00117-4.

Fenton, G.A. 1999a. Estimation for Stochastic Soil Models. Journal of Geotechnical and Geoenvironmental Engineering, 125, 470-485, doi: doi:10.1061/(ASCE)1090-0241(1999)125:6(470).

Fenton, G.A. 1999b. Random Field Modeling of CPT Data. Journal of Geotechnical and Geoenvironmental Engineering, 125, 486-498, doi: doi:10.1061/(ASCE)1090-0241(1999)125:6(486).

Guan, Z. & Wang, Y., 2020. Statistical charts for determining sample size at various levels of accuracy and confidence in geotechnical site investigation. Géotechnique, 70(12), 1145-1159.

Guan, Z. & Wang, Y. 2023. Data-driven simulation of two-dimensional cross-correlated random fields from limited measurements using joint sparse representation. Reliability Engineering & System Safety, 238, 109408, doi: https://doi.org/10.1016/j.ress.2023.109408.

Guan, Z., Wang, Y. and Phoon, K.K., 2023a. Fusion of Sparse Non-co-located Measurements from Multiple Sources for Geotechnical Site Investigation. Canadian Geotechnical Journal, doi: https://doi.org/10.1139/cgj-2023-0289.

Guan, Z., Wang, Y. & Phoon, K.-K. 2023b. Dictionary learning of sparse ground investigation data from a specific site and existing database from other sites. (Manuscript in preparation).

Han, L., Wang, L., Ding, X., Wen, H., Yuan, X. & Zhang, W. 2022. Similarity quantification of soil parametric data and sites using confidence ellipses. Geoscience Frontiers, 13, 101280, doi: https://doi.org/10.1016/j.gsf.2021.101280.

Hu, Y., Zhao, T., Wang, Y., Choi, C. & Ng, C.W.W. 2019. Direct simulation of two-dimensional isotropic or anisotropic random field from sparse measurement using Bayesian compressive sampling. Stochastic Environmental Research and Risk Assessment, 33, 1477-1496, doi: 10.1007/s00477-019-01718-7.

Hu, Y. and Wang, Y., 2024. Evaluating statistical homogeneity of cone penetration test (CPT) data profile using auto-correlation function. Computers and Geotechnics, 165, p.105852.

Huang, S.P., Quek, S.T. & Phoon, K.K. 2001. Convergence study of the truncated Karhunen–Loeve expansion for simulation of stochastic processes. International Journal for Numerical Methods in Engineering, 52, 1029-1043, doi: https://doi.org/10.1002/nme.255.

Itskov, M. 2007. Tensor algebra and tensor analysis for engineers. Springer.

Juang, C.H., Zhang, J., Shen, M. & Hu, J. 2019. Probabilistic methods for unified treatment of geotechnical and geological uncertainties in a geotechnical analysis. Engineering Geology, 249, 148-161, doi: https://doi.org/10.1016/j.enggeo.2018.12.010.

Katsman, R. & Painuly, A. 2022. Influence of anisotropy in mechanical properties of muddy aquatic sediment on CH4 bubble growth direction and migration pattern. Engineering Geology, 299, 106565, doi: https://doi.org/10.1016/j.enggeo.2022.106565.

Leung, A.Y. 2023. Soil-Structure Interaction in Spatially Variable Ground. In Uncertainty, Modeling, and Decision Making in Geotechnics (pp. 427-440). CRC Press.

Lyu, B., Hu, Y. & Wang, Y. 2023. Data-Driven Development of Three-Dimensional Subsurface Models from Sparse Measurements Using Bayesian Compressive Sampling: A Benchmarking Study. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 9, 04023010.

Müller, S., Schüler, L., Zech, A. and Heße, F., 2022. GSTools v1. 3: a toolbox for geostatistical modelling in Python. Geoscientific Model Development, 15(7), pp.3161-3182.

Nguyen, H.V. & Bai, L. 2011. Cosine Similarity Metric Learning for Face Verification. *In*: Kimmel, R., Klette, R. & Sugimoto, A. (eds.) *Computer Vision – ACCV 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, 709-720.

NZGD. 2023. New Zealand Geotechnical Database. *In*: (EQC), N.Z.E.C. (ed.), https://www.nzgd.org.nz (accessed 3 October 2023)

Onyejekwe, S., Kang, X. & Ge, L. 2016. Evaluation of the scale of fluctuation of geotechnical parameters by autocorrelation function and semivariogram function. Engineering Geology, 214, 43-49, doi: https://doi.org/10.1016/j.enggeo.2016.09.014.

Pati, Y.C., Rezaiifar, R. & Krishnaprasad, P. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 40-44.

Phoon, K.-K. & Ching, J. 2022. Additional observations on the site recognition challenge. Journal of GeoEngineering, 17, 231-247.

Phoon, K.-K., Ching, J. & Shuku, T. 2022. Challenges in data-driven site characterization. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16, 114-126, doi: 10.1080/17499518.2021.1896005.

Phoon, K.-K. & Fenton, G.A. 2004. Estimating sample autocorrelation functions using bootstrap. *Ninth ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Albuquerque, New Mexico.

Phoon, K.-K. & Kulhawy, F.H. 1999. Characterization of geotechnical variability. Canadian Geotechnical Journal, 36, 612-624.

Phoon, K.-K. & Zhang, W. 2023. Future of machine learning in geotechnics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 17, 7-22, doi: 10.1080/17499518.2022.2087884.

Priestley, M.B. 1981. Spectral analysis and time series: probability and mathematical statistics.

Rafiee, J. & Tse, P.W. 2009. Use of autocorrelation of wavelet coefficients for fault diagnosis. Mechanical Systems and Signal Processing, 23, 1554-1572, doi: https://doi.org/10.1016/j.ymssp.2009.02.008.

Rao, K.R. & Yip, P. 1990. Discrete cosine transform: algorithms, advantages, applications. Academic Press Professional, Inc.

Shalvi, O. & Weinstein, E. 1996. System identification using nonstationary signals. IEEE Transactions on Signal Processing, 44, 2055-2063, doi: 10.1109/78.533725.

Shannon, C.E. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Sharma, A., Ching, J. & Phoon, K.-K. 2022. A Hierarchical Bayesian Similarity Measure for Geotechnical Site Retrieval. Journal of Engineering Mechanics, 148, 04022062.

Shechtman, E. & Irani, M. 2007. Matching Local Self-Similarities across Images and Videos. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1-8.

Shi, C. & Wang, Y. 2021a. Development of subsurface geological cross-section from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost. Journal of Geotechnical and Geoenvironmental Engineering, 147, 04021082.

Shi, C. & Wang, Y. 2021b. Training image selection for development of subsurface geological cross-section by conditional simulations. Engineering Geology, 295, 106415.

Shi, C., Wang, Y. & Kamchoom, V. 2023. Data-driven multi-stage sampling strategy for a three-dimensional geological domain using weighted centroidal voronoi tessellation and IC-XGBoost3D. Engineering Geology, 325, 107301, doi: https://doi.org/10.1016/j.enggeo.2023.107301.

Shinozuka, M. & Deodatis, G. 1991. Simulation of Stochastic Processes by Spectral Representation. Applied Mechanics Reviews, 44, 191-204, doi: 10.1115/1.3119501.

Shuku, T., Phoon, K.-K. & Yoshida, I. 2020. Trend estimation and layer boundary detection in depth-dependent soil data using sparse Bayesian lasso. Computers and Geotechnics, 128, 103845, doi: https://doi.org/10.1016/j.compgeo.2020.103845.

Simakov, D., Caspi, Y., Shechtman, E. & Irani, M. 2008. Summarizing visual data using bidirectional similarity. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1-8.

Tipping, M.E. 2001. Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1, 211-244.

Vanmarcke, E. 2010. Random Fields. World Scientific.

Wallace, G.K. 1992. The JPEG still picture compression standard. IEEE Transactions on Consumer Electronics, 38, xviii-xxxiv, doi: 10.1109/30.125072.

Wang, Y., Hu, Y. & Phoon, K.-K. 2022. Non-parametric modelling and simulation of spatiotemporally varying geo-data. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16, 77-97, doi: 10.1080/17499518.2021.1971258.

Wang, Y. & Zhao, T. 2016. Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. Canadian Geotechnical Journal, 53, 1547-1559.

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, 04018130, doi: doi:10.1061/(ASCE)EM.1943-7889.0001560.

Webster, R. & Oliver, M.A. 2007. Geostatistics for environmental scientists. John Wiley & Sons.

Wen, Y.K. & Gu, P. 2004. Description and Simulation of Nonstationary Processes Based on Hilbert Spectra. Journal of Engineering Mechanics, 130, 942-951, doi: doi:10.1061/(ASCE)0733-9399(2004)130:8(942).

Xu, J., Wang, Y. & Zhang, L. 2021. Interpolation of extremely sparse geo-data by data fusion and collaborative Bayesian compressive sampling. Computers and Geotechnics, 134, 104098.

Yang, H.Q., Zhang, L., Gao, L., Phoon, K.K. and Wei, X., 2022. On the importance of landslide management: Insights from a 32-year database of landslide consequences and rainfall in Hong Kong. Engineering Geology, 299, p.106578.

Yoshida, I., Tomizawa, Y. & Otake, Y. 2021. Estimation of trend and random components of conditional random field using Gaussian process regression. Computers and Geotechnics, 136, 104179, doi: https://doi.org/10.1016/j.compgeo.2021.104179.

Zhang, Y., Li, Y.E. & Ku, T. 2021. Soil/rock interface profiling using a new passive seismic survey: Autocorrelation seismic interferometry. Tunnelling and Underground Space Technology, 115, 104045, doi: https://doi.org/10.1016/j.tust.2021.104045.

Zhao, C., Gong, W., Juang, C.H., Tang, H., Hu, X., Wang, L., 2023. Optimization of site exploration program based on coupled characterization of stratigraphic and geo-properties uncertainties. Engineering Geology, 317, p.107081.

Zhao, T., Hu, Y. & Wang, Y. 2018. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. Engineering Geology, 246, 162-175, doi: https://doi.org/10.1016/j.enggeo.2018.09.022.

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.