Data-driven Recovery of Incomplete Geotechnical Dataset Using Low-rank Matrix Completion

Guan, Zheng; Wang, Yu; Phoon, Kok-Kwang

DOI
10.1139/cgj-2024-0781

Publication date
2025

Document Version
Final published version

Published in
Canadian Geotechnical Journal

Canadian Geotechnical Journal

OPEN ACCESS | Research Article

Data-driven recovery of incomplete geotechnical dataset using low-rank matrix completion

Zheng Guan[a], **Yu Wang**[b], **and Kok-Kwang Phoon**[c]

[a]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands; [b]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; [c]Information Systems Technology and Design/Architecture and Sustainable Design, Singapore University of Technology and Design, 8 Somapah Rd., Singapore

Corresponding author: **Yu Wang** (email: wang.yu@ust.hk)

Abstract

Real geotechnical data from a typical site might be characterized as MUSIC-X (i.e., multivariate, uncertain, unique, sparse, incomplete, and potentially corrupted, with X denoting spatial/temporal variability). One of the key challenges in developing site-specific statistical models for multiple geotechnical properties (i.e., multivariate) is missing (or incomplete) values from different tests at various depths/locations. This raises a critical question in geotechnical site investigations: how to recover the missing values in real geotechnical datasets from available measurements by leveraging the underlying structure of geotechnical datasets? Since different geotechnical properties are not only correlated among different properties, but also auto-correlated across different depths, this suggests that a simple underlying structure with only a limited number of important features/patterns might exist for multivariate geotechnical datasets. Leveraging on this observation, this study proposes a novel, data-driven method for predicting missing values by low-rank matrix completion. The proposed method exploits the auto- and cross-correlation structures of different test data. Missing values are then recovered using a singular value thresholding algorithm, and a *k*-fold cross-validation strategy is employed to determine the level of measurement noise. The method is illustrated and validated using a real geotechnical dataset. The results indicate that the proposed method can provide reliable predictions.

Key words: geotechnical site characterization, low-rank matrix, incomplete data, MUSIC-X, matrix completion

Résumé

Les données géotechniques réelles d'un site typique peuvent être caractérisées comme MUSIC-X (c.-à-d. multivariées, incertaines, uniques, clairsemées, incomplètes et potentiellement corrompues, X dénotant une variabilité spatiale et temporelle). L'un des principaux défis dans le développement de modèles statistiques spécifiques au site pour plusieurs propriétés géotechniques (c'est-à-dire multivariées) est que des valeurs manquantes (ou incomplètes) proviennent de différents essais à diverses profondeurs ou emplacements. Cela soulève une question cruciale dans les investigations géotechniques sur site : comment récupérer les valeurs manquantes dans les ensembles de données géotechniques réels à partir des mesures disponibles en exploitant la structure sous-jacente de ces ensembles de données? Étant donné que les différentes propriétés géotechniques sont non seulement corrélées entre elles, mais aussi autocorrélées à différentes profondeurs, cela suggère qu'une structure sous-jacente simple, composée d'un nombre limité de caractéristiques ou de motifs importants, pourrait exister pour les ensembles de données géotechniques multivariées. S'appuyant sur cette observation, cette étude propose une méthode innovante basée sur les données pour prédire les valeurs manquantes par complétion de matrice à faible rang. La méthode proposée exploite les structures d'autocorrélation et de corrélation croisée de différentes données d'essai. Les valeurs manquantes sont ensuite récupérées à l'aide d'un algorithme de seuillage des valeurs singulières, et une stratégie de validation croisée en k plis est utilisée pour déterminer le niveau de bruit de mesure. La méthode est illustrée et validée à l'aide d'un jeu de données géotechniques réelles. Les résultats montrent que la méthode proposée fournit des prédictions fiables.

Mots-clés : caractérisation géotechnique du site, matrice à faible rang, données incomplètes, MUSIC-X, complétion de matrice

1. Introduction

In geotechnical engineering practice, it is common that different types of field or laboratory tests are conducted during a site investigation program to identify the engineering properties of soils or rocks at different depths and locations (e.g., Ching and Phoon 2014). However, due to budget and/or

Fig. 1. Illustration of incomplete multivariate geotechnical dataset.

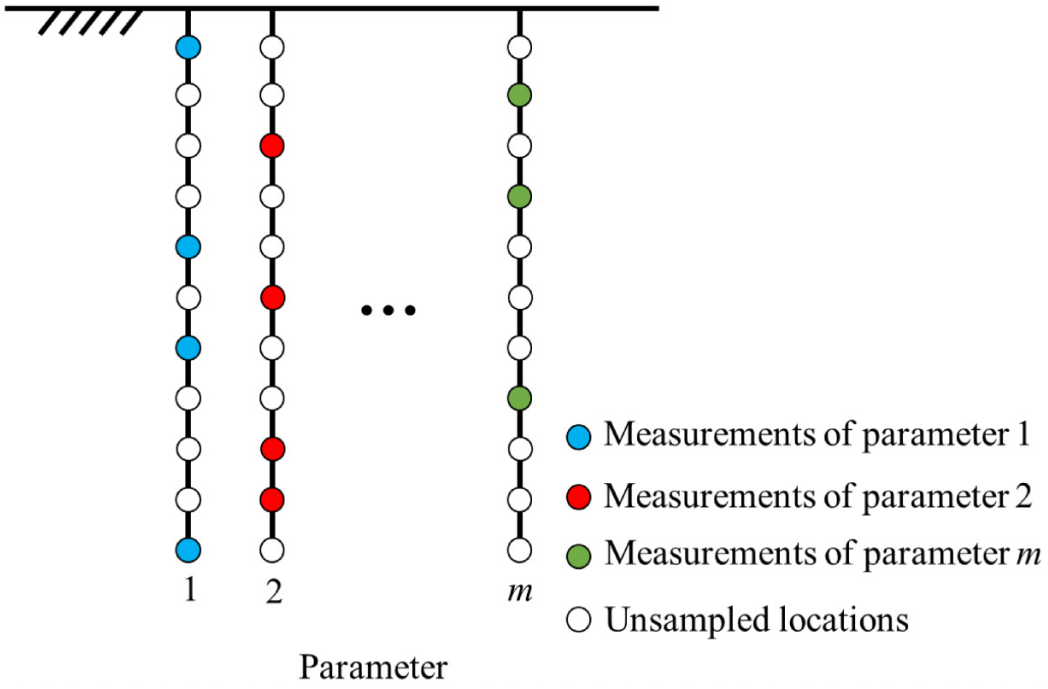


Table 1. Geotechnical site investigation data for a site in Onsøy, Norway (after Lacasse and Lunne 1982).

Depth	LL	PI	LI	σ'_{vo}/P_a	σ'_p/P_a	s_u	S_t	B_q	q_t	q_{tu}
1.00	56.20	20.00	1.54	0.06	0.85	0.12	6	0.16	29.11	25.57
1.90	50.20	18.10	1.82	0.12	0.60	0.11	14	0.24	17.69	14.58
3.50	59.90	30.50	0.93	0.22	0.48	0.11	15	0.30	10.52	8.41
5.20	56.80	22.90	1.07	0.32	0.45	0.12	7	0.35	7.70	6.11
7.60	66.30	31.50	0.87	0.47	0.54	0.11	14	0.47	5.89	4.25
9.50	65.10	29.60	0.97	0.58		0.15	12	0.41	6.19	4.74
10.80	74.40	36.10	0.81	0.65	0.84	0.16	9	0.46	5.93	4.31
13.40	71.40	35.80	0.87	0.81	1.05	0.19		0.47	5.95	4.24
16.30	72.70	34.70	0.76	0.99	0.99	0.24		0.55	6.13	3.88

Note: LL, liquid limit; PI, plasticity index; LI, liquidity index; σ'_{vo}, vertical effective stress; σ'_p, effective pre-consolidation stress; P_a, atmospheric pressure; s_u, undrained shear strength; S_t, sensitivity; B_q, pore pressure ratio; q_t, normalized cone tip resistance; q_{tu}, effective cone tip resistance.

technical constraints, test data from a specific site are often sparse and incomplete. The test data are sparse because different tests are only performed at limited depths/locations. This also leads to an incomplete dataset due to missing values from certain tests at some depths (e.g., Phoon et al. 2024), as illustrated in Fig. 1. Consider, for example, Table 1, which presents geotechnical site investigation data obtained from a site in Onsøy, Norway (Lacasse and Lunne 1982). The columns in Table 1 represent different soil properties (e.g., liquid limit, plasticity index, …, effective cone tip resistance), and the rows represent measurement data from each test at different depths (e.g., 1.0 m, 1.9 m, …, 16.3 m). It can be observed from Table 1 that measurements for some tests (e.g., pre-consolidation stress and sensitivity) are missing at certain depths, as shown by red shadows in the cells. The presence of missing data makes such a limited geotechnical dataset even less informative and more difficult to analyze, posing a significant challenge in developing reliable statistical models (e.g., multivariate probability distribution functions or transformation models) for geotechnical properties at a typical site.

Phoon et al. (2019) described the geotechnical site investigation data as MUSIC-X (multivariate, uncertain and unique, sparse, incomplete, and potentially corrupted with "X" denoting the spatial/temporal dimension). Ching and Phoon (2019) proposed a Gibbs sampler approach to address incomplete geotechnical data by simulating missing values with consideration of the auto- and cross-correlation of different test data. In their method, the auto-correlation of different test data is modeled using the same stationary random field, implying that different geotechnical properties share the same auto-correlation structure. In addition, their approach requires the pre-estimation of site-specific auto-correlation structures as input. However, real geotechnical

data are often non-stationary and non-Gaussian (e.g., exhibiting complex multimodal distributions), and different geotechnical properties may have significantly different autocorrelation structures. Furthermore, it is notoriously challenging to reliably estimate the auto- and cross-correlation parameters from sparse site-specific measurements. The autocorrelation parameters are particularly difficult to estimate in the horizontal direction (Ching et al. 2018). Guan and Wang (2021) proposed a Bayesian compressive sampling-based method for constructing a joint probability distribution of cross-correlated geotechnical properties from sparse site-specific measurements. Although their method is non-parametric, it was developed based on the assumption of independent measurement data along depth, i.e., the autocorrelation of each property is ignored. More recently, Mu et al. (2024) utilized the multivariate Gaussian copula to model the cross-correlation among different geotechnical properties, and employed the nearest correlation matrix and bootstrapping techniques to address the non-positive definiteness problem of the cross-correlation matrix. Despite these advancements, dealing with missing values from sparse site-specific measurements remains a challenging task due to the various assumptions in existing methods that real geotechnical data may not satisfy. Such leap of faith assumptions that are not informed by data should be kept to a minimum.

Phoon et al. (2022*b*) argued that the primary goal of "data-centric geotechnics" is to prioritize a "data-first" agenda in practice. This goal can only be achieved through the development and implementation of purely data-driven methods capable of handling real-world geotechnical data with varying, or even unknown statistical characteristics (e.g., stationary or non-stationary; Gaussian or non-Gaussian). Since different test data are usually cross-correlated and auto-correlated at different depths, multivariate geotechnical datasets may have a simple underlying structure, i.e., with only a limited number of important features/patterns (e.g., Phoon et al. 2022*a*; Guan et al. 2024). This offers a promising alternative to recover the missing value in geotechnical datasets from all available measurements by leveraging the underlying structure of geotechnical datasets. To explore this alternative, this study proposes a novel, data-driven method for recovering missing values from all available measurements of different tests. In the proposed method, the auto- and cross-correlation structures of different test data are exploited using a low-rank modeling approach. In the proposed approach, a multivariate geotechnical dataset (e.g., Table 1) is represented as a low-rank matrix for exploiting its low-rank structure. The low-rank structure of a dataset or matrix means that the original complete data matrix, $\mathbf{M}$, can be well-approximated by a matrix with a much lower rank than $\mathbf{M}$, in which the number of independent rows or columns (i.e., the rank of a matrix) is less than the number of all rows or columns. Leveraging the underlying low-rank structure of geotechnical datasets offers a promising alternative to predict missing data based on the available observations. In this study, missing values are proposed to be recovered from all measurements using a low-rank matrix completion method called the singular value thresholding (SVT) algorithm (Cai et al. 2010). To account for measurement noise in geotechnical datasets, a k-fold cross-validation strategy is proposed to determine the most probable noise level. Additionally, bootstrapping is adopted to quantify uncertainty in the prediction of the missing entries. The proposed approach is non-parametric and data-driven, and it does not require assumptions of Gaussian distribution or stationary random fields, making it directly applicable to real-world geotechnical data.

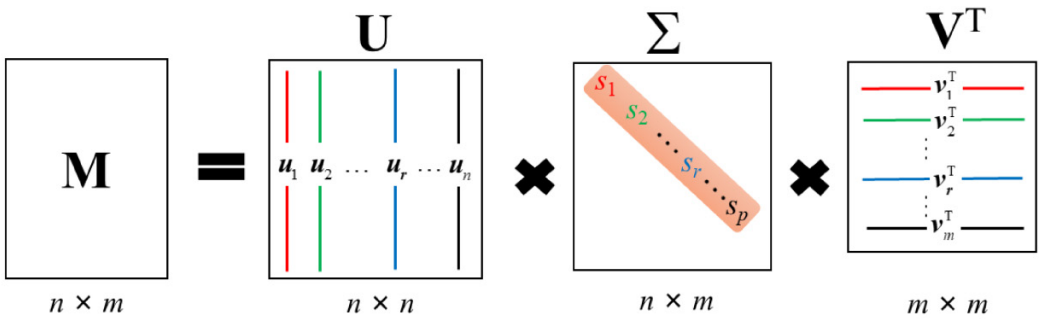
The remainder of this manuscript is organized as follows: Section 2 introduces the low-rank structure of geotechnical datasets. Section 3 proposes a method for recovering missing geotechnical data using the SVT algorithm, followed by a k-fold cross-validation strategy for determining the most probable level of measurement noise in Section 4. Section 5 details the implementation procedure of the proposed method. Quantification of uncertainty in the prediction by bootstrapping is described in Section 6. Section 7 demonstrates the method using a real geotechnical dataset, and concluding remarks are provided in Section 8.

2. Low-rank structure of multivariate geotechnical datasets

It is well-recognized that many real-world multivariate datasets often exhibit a low-rank structure or characteristics (e.g., Lingala et al. 2011; Liu et al. 2013; Tang et al. 2014; Nguyen et al. 2019). For instance, a multivariate data matrix (e.g., Table 1) can frequently be approximated by a matrix with a much lower rank than the original matrix. In other words, only a limited number of independent rows or columns (i.e., important patterns or features of the original datasets) are sufficient to capture the main information of the original datasets. Because rows or columns of the original datasets are usually highly cross-correlated, low-dimensional structures, or patterns, often emerge in the original datasets. For example, images typically have a low-rank structure because the pixel values in an image are often highly correlated, indicating redundant information in the original image. This suggests that only a relatively small number of underlying patterns or features is required to properly represent the original image. Similarly, in geotechnical site investigations, geotechnical properties estimated from different tests are usually cross-correlated among properties and auto-correlated along depths, resulting in a low-rank structure for geotechnical datasets.

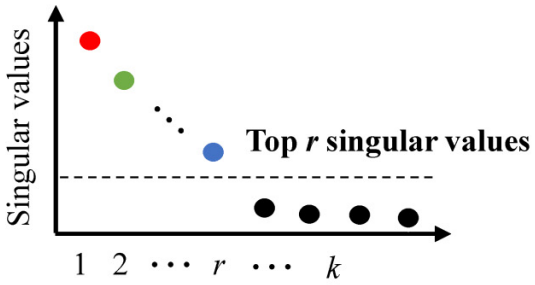
The basic idea of low-rank modeling is to represent a geotechnical data matrix as a summation of a limited number of components, ordered by their importance. Only the most important components are selected and retained to preserve the majority of the information from the original dataset. Singular value decomposition (SVD) is a widely used method for low-rank modeling (e.g., Candès and Tao 2010), as it can capture the essential features of the original matrix and approximate the matrix as a summation of a limited number of rank-one matrices (components), as illustrated in Fig. 2. Let a matrix $\mathbf{M}$ with a dimension of $n \times m$ represent a geotechnical dataset with a number, m, of geotechnical properties, each of which is measured at n different locations or depths. An SVD of the matrix $\mathbf{M}$ provides a unique matrix decomposition,

Fig. 2. Low-rank approximation of a multivariate geotechnical data matrix, (**M**). SVD, singular value decomposition.

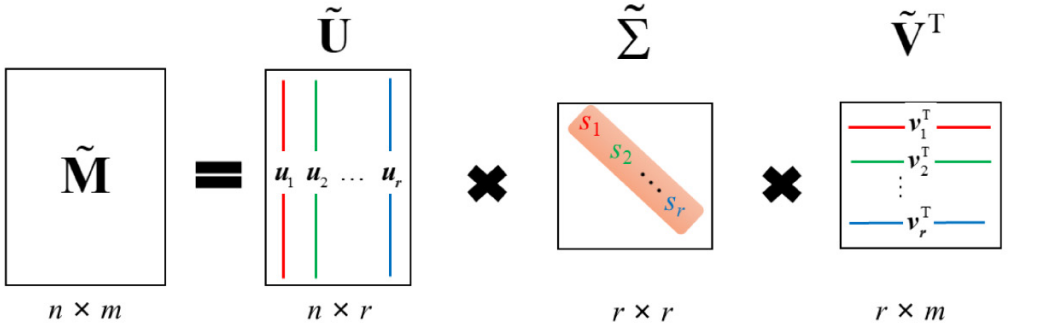


$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \sum_{i=1}^{p} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$$

(a) Singular value decomposition (SVD) for a geotechnical data matrix, **M**



(b) Truncation of the top r singular values



$$\mathbf{M} \approx \tilde{\mathbf{M}} = \tilde{\mathbf{U}}\,\tilde{\Sigma}\,\tilde{\mathbf{V}}^{\mathrm{T}} = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}$$

(c) Low-rank approximation of **M**

Table 2. Geotechnical site investigation data for a site in Beauharnois, Quebec, Canada (after Ching and Phoon 2014).

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	92.83	1.06	51.00	29.00	1.52	70.49	38.23	0.54	35.22
5.76	90.88	91.66	1.01	58.00	35.00	1.51	69.05	31.98	0.46	28.16
6.47	94.78	95.17	1.00	66.00	41.00	1.29	38.05	29.25	0.77	24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56	81.29	44.08	0.54	38.52
7.69	102.59	103.37	1.01	68.00	43.00	1.23	42.98	40.18	0.93	33.40
8.36	106.49	107.66	1.01	68.00	42.00	1.29	93.72	37.45	0.40	31.34
8.89	109.61	111.17	1.01	62.00	36.00	1.31	57.84	44.47	0.77	38.86
9.41	112.73	115.07	1.02	55.00	30.00	0.93	33.71	63.58	1.89	58.14
9.92	116.24	118.19	1.02	57.00	25.00	0.84	22.50	56.56	2.51	53.78
10.36	118.97	121.70	1.02	44.00	22.00	1.23	63.02	58.90	0.93	57.38
10.88	122.48	125.60	1.03	58.00	33.00	1.18	52.17	48.76	0.93	43.58
11.96	130.67	133.01	1.02	51.00	25.00	0.88	32.27	60.85	1.89	57.86

Note: σ'_{vo}, vertical effective stress; σ'_p, effective pre-consolidation stress; OCR, over-consolidation ratio; LL, liquid limit; PI, plasticity index; LI, liquidity index; S_t, sensitivity; $s_{u,VST}$, undrained shear strength obtained from vane shear test; $s_{u(mob)}$, mobilized undrained shear strength.

expressed as (e.g., Stewart 1993):

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\,\mathbf{V}^{\mathrm{T}} \tag{1}$$

where an $n \times n$ matrix $\mathbf{U}$ and an $m \times m$ matrix $\mathbf{V}$ are unitary matrices; $\boldsymbol{\Sigma}$ is an $n \times m$ rectangular diagonal matrix with non-negative real numbers on the diagonal. The columns of $\mathbf{U}$, $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ are called left singular vectors of $\mathbf{M}$, while the columns of $\mathbf{V}$, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ are called right singular vectors of $\mathbf{M}$. The diagonal elements of Σ, $s_1, s_2, \ldots, s_p$ are called singular values, where p is the minimal value of n and m, $p = \min\{n, m\}$. The singular values are listed in descending order (i.e., from the largest to the smallest, $s_1 \geq s_2 \geq \ldots \geq s_p$). In SVD, $\mathbf{M}$ can be represented as a linear combination of p rank-one matrices generated by an outer products of the left and right singular vectors, as shown in Fig. 2*a* (e.g., Stewart 1993):

$$\mathbf{M} = \sum_{i=1}^{p} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the i-th left and right singular vectors of $\mathbf{M}$, respectively; $\boldsymbol{u}_i\boldsymbol{v}_i^{\mathrm{T}}$ represents the i-th rank-one matrix/component; s_i is the corresponding weight coefficient or scaling factor. Because $\mathbf{M}$ is a linear summation of p rank-one matrices, $\mathbf{M}$ has a rank of p.

SVD provides interpretable insights about the auto- and cross-correlation structures of the geotechnical datasets. The left singular vectors capture patterns across rows (i.e., depth in this study) for the dominant latent modes in the data, while the right singular vectors correspond to patterns across columns, i.e., the cross-correlation among different geotechnical variables (e.g., Wall et al. 2003; Jolliffe and Cadima 2016). Essentially, SVD transfers the original geotechnical matrix into a new coordinate system generated by the singular vectors of $\mathbf{M}$. The singular values, $s_1, s_2, \ldots, s_p$ indicate the scaling factors along each dimension of the new coordinate system. The larger the singular value, the more important the corresponding dimension (component) is for representing patterns of the geotechnical dataset. Since different geotechnical properties are usually cross-correlated and measurements of each geotechnical property are auto-correlated along depths, only a limited number of components with significantly large weight coefficients are required to capture the main patterns of the original data matrix. Then, only the first r most important components with significantly large weights are needed for approximating the original data matrix, as shown in Fig. 2*b*. In other words, the data matrix, $\mathbf{M}$ with a rank of p, is approximated using a matrix $\tilde{\mathbf{M}}$ with a much lower rank, r ($r < p$), as illustrated in Fig. 2*c*:

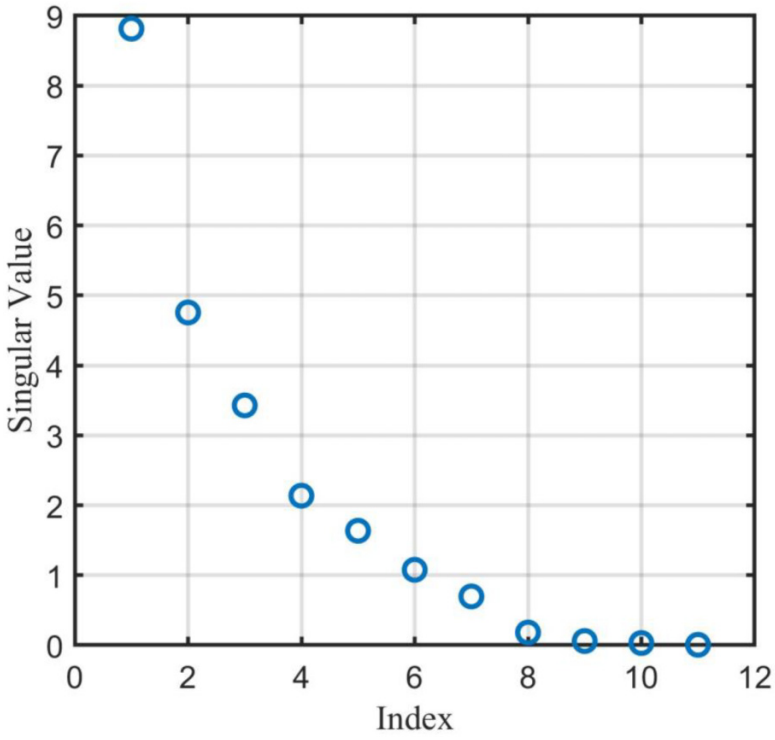
$$\mathbf{M} \approx \tilde{\mathbf{M}} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\left(\tilde{\mathbf{V}}\right)^{\mathrm{T}} = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \tag{3}$$

where an $n \times r$ matrix, $\tilde{\mathbf{U}} = \left[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r\right]$ contains the top r left singular vectors; an $m \times r$ matrix, $\tilde{\mathbf{V}} = \left[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r\right]$ contains the top r right singular vectors; $\tilde{\Sigma}$ is a $r \times r$ square diagonal matrix includes the top r singular values, $s_1, s_2, \ldots, s_r$.

Consider, for example, a complete geotechnical dataset from Beauharnois, Quebec, Canada, obtained from the CLAY/10/7490 database (Ching and Phoon 2014), as shown in Table 2. This dataset can be treated as a matrix with a dimension of 12 × 11. Then, SVD is performed for this matrix. The ordered singular values of this matrix or the weight coefficients of components are shown in Fig. 3. It is observed from this figure that the first several singular values are significantly larger than the remaining ones. In SVD, relative energy is often adopted to evaluate the significance of the singular values in capturing the variability of a data matrix, which is defined as the proportion of the total variability captured by the first q singular values (e.g., Chatterjee 2000):

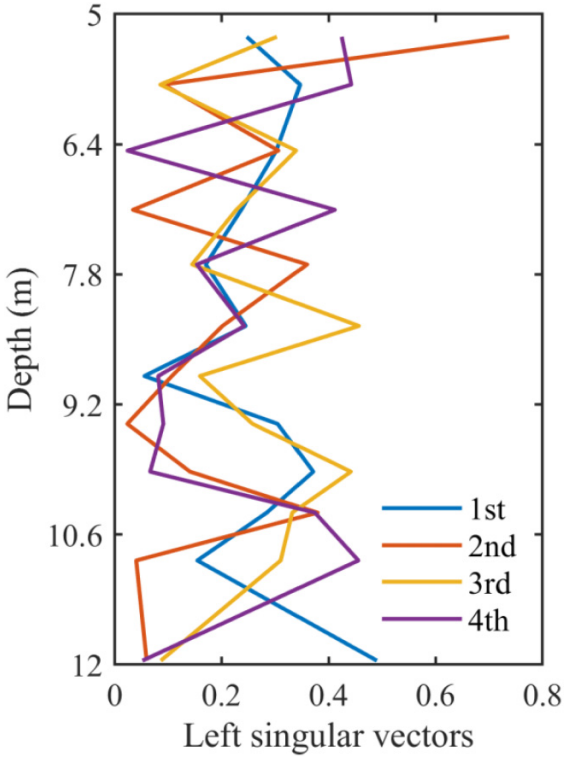
$$E_q = \frac{s_1^2 + s_2^2 + \cdots + s_q^2}{s_1^2 + s_2^2 + \cdots + s_p^2} \tag{4}$$

Fig. 3. Singular values of the geotechnical data matrix for the site in Beauharnois, Quebec, Canada.



Using eq. 4, the relative energy for the first four singular values is calculated as, $E_4 = 96.4\%$. This means that keeping just the first four singular values may capture approximately 96.4% variability of this geotechnical dataset. Therefore, the first four components can be used to effectively approximate the original complete dataset. In other words, the rank of this matrix is about four. The first four dominant left singular vectors, representing key spatial patterns along depth for different properties, are shown in Fig. 4. The low-rank structure of geotechnical datasets suggests a presence of redundant information in the datasets because many data points are dependent or highly correlated. This redundancy allows for an effective prediction of missing data based on the available observations, as described in the following section.

Fig. 4. First four dominant left singular vectors for the geotechnical data matrix.



3. Missing data prediction using low-rank matrix completion method

In geotechnical engineering practice, it is common that certain tests are not conducted at every depth, due to the destructive nature of some tests and some practical constraints such as high cost (e.g., expensive and time-consuming triaxial tests). This results in incomplete geotechnical datasets. This leads to a problem of how to recover the original complete data matrix, **M**, with n rows and m columns, when only a portion of the entries are observed. In general, if each entry in the data matrix is independent or uncorrelated, recovering the missing values is impossible. However, when the matrix has an underlying simple structure, or more specifically, when it is low-rank or approximately low-rank (i.e., the matrix can be represented by only a limited number of important components), it becomes possible to accurately predict the missing data by leveraging the low-rank structure.

One of the well-known applications of matrix completion/recovery is the one-million-dollar Netflix challenge (e.g., Piotte and Chabbert 2009; Feuerverger et al. 2012; Gomez-Uribe and Hunt 2015). In Netflix, users have the opportunity to rate movies, which can be summarized in a matrix form, where the rows represent users, the columns represent movies, and the entries indicate the ratings. However, users typically rate only a limited number of movies, leading to an incomplete matrix with very limited observed entries. The challenge is to recover this incomplete matrix so that Netflix can efficiently recommend movies to individual users. The user ratings matrix typically has an approximately low-rank structure because the variability in ratings can be captured by only a few latent factors or features (e.g., genres, actors, or themes). Similar challenges arise in predicting musical preferences for Apple Inc. and customer preferences for Amazon.com.

The basic idea behind low-rank matrix completion for incomplete geotechnical datasets is to impute the missing values of different tests in a manner that aligns with the low-rank approximation of the observed measurements. Mathematically, the challenge is to determine a complete data matrix with the lowest possible rank, denoted as $\mathbf{X}$, that corresponds to the measurements. This can be formulated as an optimization problem below (e.g., Candès and Recht 2008):

$$\text{(5)} \quad \begin{aligned} &\text{minimize} \quad \text{rank}(\mathbf{X}) \\ &\text{subject to} \quad X_{i,j} = M_{i,j} \quad (i,j) \in \Omega \end{aligned}$$

where $M_{i,j}$ represents the observed measurements at the i-th row and j-th column of the incomplete data matrix; Ω indicates the set of locations corresponding to the observed entries, i.e., $(i, j) \in \Omega$; $X_{i,j}$ is the entry at the i-th row and j-th column of the matrix, $\mathbf{X}$. The physical meaning of eq. 5 is that it finds the simplest underlying structure that can adequately explain the observed data. However, the minimization of matrix rank is nonconvex and generally impossible to solve as its solution usually requires an intractable combinatorial search. Recent developments in matrix completion (e.g., Candès and Recht 2008; Candès and Plan 2010; Candès and Tao 2010) suggest that the solution to the matrix rank minimization problem, as expressed in eq. 5, can be relaxed to a nuclear norm minimization problem. This approach is computationally tractable via convex optimization and typically results in the lowest possible rank:

$$\text{(6)} \quad \begin{aligned} &\text{minimize} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad X_{i,j} = M_{i,j} \quad (i,j) \in \Omega \end{aligned}$$

where $\|\cdot\|_*$ represents the nuclear norm of a matrix, which is the sum of singular values of a matrix.

In geotechnical engineering practice, measurements from different tests, i.e., observed entries in the matrix, often contain noise. Exactly or closely fitting to the significantly noisy geotechnical measurement data can lead to an over-fitting problem and be less generalizable, i.e., the model performs well on the observed entries but poorly on missing or unseen entries. To overcome this issue, entries from a recovered low-rank matrix shall not correspond exactly to the measured data points. Consequently, instead of enforcing a strict equality constraint, a measurement error term is introduced to the eq. 6:

$$\text{(7)} \quad \begin{aligned} &\text{minimize} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad \frac{\sqrt{\sum_{(i,j)\in\Omega} \left(X_{i,j} - M_{i,j}\right)^2}}{\sqrt{\sum_{(i,j)\in\Omega} {M_{i,j}}^2}} \leq \varepsilon \end{aligned}$$

where ε indicates the level of measurement noise. In this study, eq. 7 is solved using a SVT algorithm (Cai et al. 2010).

3.1. Singular value thresholding algorithm

The SVT algorithm is an efficient method for matrix completion. The basic idea of the SVT algorithm is to progressively build up the low-rank structure of the estimated complete data matrix by iteratively adding corrections to the matrix estimate and shrinking the singular values. It starts with an initial iteration matrix and then adaptively applies SVD to the matrix and shrinks its singular values by a certain threshold. The iteration matrix defines how the next iteration is based on the current approximation. At each iteration, the iteration matrix is updated by adding the difference between the observed entries and current estimates of the complete geotechnical data matrix. The thresholding of singular values encourages the solution to have low rank by removing small singular values, which correspond to noise or trivial components. This involves subtracting a fixed value (i.e., a threshold) from each singular value in the matrix by the soft-thresholding operator, $D_\tau(\bullet)$, which performs the SVD for the t-th iteration matrix $\mathbf{Y}_t$, $\mathbf{Y}_t = \mathbf{U\Sigma V}^{\mathrm{T}}$ and modifies the singular values by applying the shrinkage (e.g., Combettes and Wajs 2005):

$$\text{(8)} \quad \begin{aligned} &D_\tau(\mathbf{Y}_t) = \mathbf{U}S_\tau(\mathbf{\Sigma})\mathbf{V}^{\mathrm{T}} \\ &S_\tau(\Sigma_{i,i}) = \begin{cases} \Sigma_{i,i} - \tau & \text{if} \quad \Sigma_{i,i} > \tau \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where τ is the pre-determined threshold parameter; $\Sigma_{i,i}$ is the i-th diagonal entry in $\mathbf{\Sigma}$. The physical meaning of the iteration matrix is to build a balance between the observed data and the assumption of low-rank structure, updated at each step to converge toward the simplest model that still fits the available information. When no prior information is available for the missing entries, it is recommended to start with a zero matrix because it ensures the model starts from an unbiased state (e.g., Cai et al. 2010). For the complete geotechnical data matrix estimate at the t-th iteration, $\mathbf{X}_t$, it can be obtained as:

$$\text{(9)} \quad \mathbf{X}_t = D_\tau(\mathbf{Y}_t)$$

At each iteration, the difference (or residual) is calculated between the observed entries in an incomplete data matrix, $\mathbf{M}'$ and the corresponding entries in the current estimate

(e.g., t-th estimate, $\mathbf{X}_t$). Mathematically, for the observed entries $(i, j) \in \Omega$, the corresponding difference, $R_{i,j}$ can be calculated as:

$$R_{i,j} = M_{i,j} - (\mathbf{X}_t)_{i,j} \tag{10}$$

where $(\mathbf{X}_t)_{i,j}$ indicates the entry at the i-th row and the j-th column of the matrix $\mathbf{X}_t$. This correction is added back to the current iteration matrix so that the next iteration of the geotechnical data matrix estimate will better approximate the observed values (Cai et al. 2010):

$$\mathbf{Y}_{t+1} = \mathbf{Y}_t + \delta P_\Omega(\mathbf{M}' - \mathbf{X}_t) \tag{11}$$

where δ represents the step size that controls the amount of adjustment made to the matrix in each iteration; $P_\Omega(\cdot)$ is the projection operation that retains the observed entries within the measurements domain, Ω and sets the remaining entries to zero:

$$P_\Omega(\mathbf{M}' - \mathbf{X}_t) = \begin{cases} R_{i,j} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

The update step can be viewed as gradually refining the matrix. The residual indicates how well the current estimate $\mathbf{X}_t$ matches the observed entries in $\mathbf{M}'$. Adding this residual back helps to reduce the difference between $\mathbf{X}_t$ and $\mathbf{M}'$ in the subsequent iterations. Since $\mathbf{M}'$ is an incomplete data matrix, updating only at the observed entries ensures the progressive incorporation of the known information, while the unobserved entries are predicted by the low-rank assumption and filled in during SVT. The iteration process continues until the difference, or the error, is smaller than a pre-specified level. Such difference can be quantitatively calculated as, $\| P_\Omega(\mathbf{X}_t - \mathbf{M}') \|_2 / \| P_\Omega(\mathbf{M}') \|_2$, where $\|\cdot\|_2$ indicates the L^2-norm, which is calculated as the square root of the sum of the squared vector or matrix values.

3.2. Data preprocessing

Geotechnical data from different tests typically have varying magnitudes, and the scale of different test data can significantly influence the results of matrix completion, because SVD is highly sensitive to the scale of the data (e.g., Akritas and Malaschonok 2004). If the features in the dataset have remarkably different scales, the singular values and corresponding vectors will be disproportionately influenced by the larger-scale features, even if they are not inherently more important to the remaining data. Therefore, it is critical to preprocess the datasets to convert the different test data into a common scale. In this study, z-score normalization is used for data processing, which is a well-known technique in machine learning for handling datasets with different ranges or units (e.g., Larsen and Marx 2005). By applying z-score normalization, the measurements from different tests in the data matrix are scaled to have a mean of zero and a standard deviation of one.

3.3. SVT implementation procedure

The implementation procedure of SVT for multivariate geotechnical datasets is summarized below:

Step 1: obtain multivariate geotechnical datasets and store these data as an incomplete data matrix, $\mathbf{M}'$ with a dimension of $n \times m$ where some entries are missing.

Step 2: Convert the measurements of each test in $\mathbf{M}'$ to a common scale using the z-score normalization technique.

Step 3: Initialize the iteration matrix, $\mathbf{Y}_0 = \mathbf{0}$ (a zero matrix with a dimension of $n \times m$) and determine the threshold, τ, step size, δ and stopping criteria, ε.

Step 4: Obtain the first iteration matrix, $\mathbf{Y}_1 = \delta P_\Omega(\mathbf{M}' - \mathbf{Y}_0)$ and set the iteration counter, $t = 1$.

Step 5: Obtain the original complete matrix estimate, $\mathbf{X}_t = D_\tau(\mathbf{Y}_t)$

Step 6: Update the iteration matrix, $\mathbf{Y}_{t+1} = \mathbf{Y}_t + \delta P_\Omega(\mathbf{M}' - \mathbf{X}_t)$.

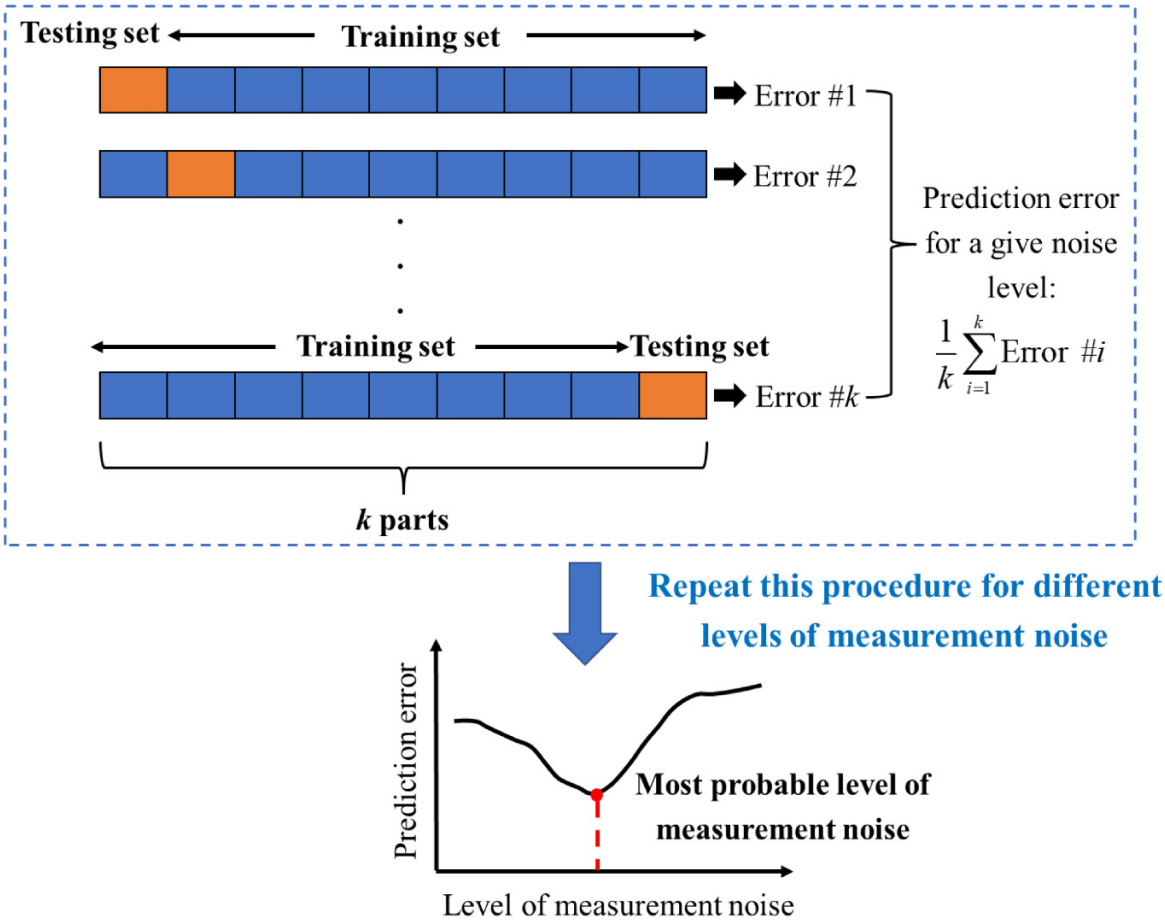
Step 7: Calculate the difference between the observed entries in an incomplete data matrix, $\mathbf{M}'$ and the corresponding entries in the current estimate, $\| P_\Omega(\mathbf{X}_t - \mathbf{M}') \|_2 / \| P_\Omega(\mathbf{M}') \|_2$. If the difference is smaller than the pre-specified measurement level, ε, stop the algorithm. Otherwise, set $t = t + 1$ and return to Step 5.

As suggested by Cai et al. (2010), τ can be set as $\tau = 5\sqrt{n \times m}$ to ensure that the complete matrix estimate has the smallest possible rank, while step size, δ and stopping criteria, ε can be taken as a small value (e.g., $\delta = 0.1$; $\varepsilon = 10^{-4}$). It should be noted that the stopping criteria, ε reflects the level of measurement noise, and geotechnical measurements unavoidably contain significant noise. Therefore, it may not be appropriate to simply adopt a small value for ε as recommended by literature or image processing (e.g., Candes et al. 2013). When substantial prior knowledge is available about the measurement noise, ε can be selected based on engineering experience or judgment. However, in many other cases, such information is usually not available, and thus a k-fold cross-validation strategy is proposed in this study for selecting the most probable measurement noise level, as illustrated in the following section.

4. k-fold cross-validation strategy for determining measurement noise

Geotechnical measurement error may be induced by various factors, such as equipment inaccuracies, procedural inconsistencies, and random testing errors (e.g., Orchant et al. 1988; Phoon and Kulhawy 1999). Equipment-related effects can arise from inaccuracies in measuring devices and systems (e.g., calibration issues), while sampling and procedural errors may stem from disturbed samples or improper handling. Additionally, random testing errors refer to other unaccounted factors due to random factors, such as slight variations in soil samples, uncontrollable environmental factors, or unavoidable noise in the measurement system. When ε is taken as a small value (e.g., $\varepsilon = 10^{-4}$), it implies that the model exactly or closely fits the observed measurements. Since the measurements contain significant noise, the model captures not only the underlying pattern but also the

Fig. 5. *k*-fold cross-validation strategy for determining the most probable level of measurement noise.



measurement noise. As a result, the model becomes overly complex and learns noise-specific patterns that do not generalize to unobserved data, leading to overfitting and poor prediction performance for missing data. To address this issue, a *k*-fold cross-validation strategy is developed in this study to select the most probable level of measurement noise.

The *k*-fold cross-validation strategy is widely used to assess the performance of a model in machine learning (e.g., Bengio et al. 2017). The idea is to split the dataset into k (e.g., $k = 8$) equal parts (folds), and then the model is developed using $k-1$ folds and the remaining fold for validation. This process is repeated k times for each fold, leading to performances of the model for the k different input scenarios. This helps to obtain a more robust prediction of the model's performance than a single input scenario. This strategy is adopted in this study to select the most probable measurement noise level for incomplete geotechnical data matrix completion. The general idea of the proposed method is to systematically train and validate the model using different subsets of the geotechnical data to obtain a robust evaluation of the model performance for different potential levels of measurement noise (i.e., different ε values). After that, the most probable or optimal measurement noise level is the one that minimizes the prediction error because it strikes the best balance between properly fitting the observed measurements and avoiding overfitting to noise. The procedure of the *k*-fold cross-validation for the most probable measurements noise determination is shown in Fig. 5.

5. Implementation procedure of the proposed method

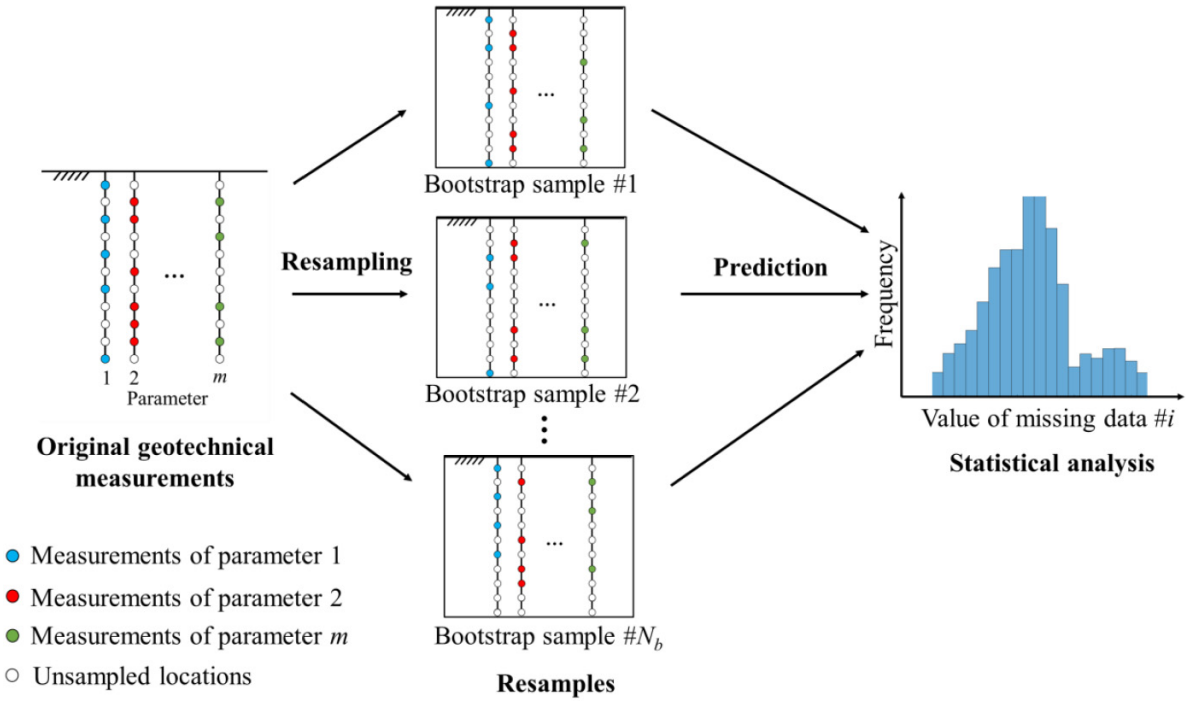
The procedure for implementing the proposed method mainly contains three modules: (1) preprocessing of a geotechnical dataset, (2) selection of the most probable level of measurement noise, and (3) prediction of missing values. Each module is described in detail below.

In Module 1, measurements of different m tests are transformed to a common scale using *z*-score normalization, and geotechnical datasets are stored in an incomplete data matrix, $\mathbf{M}'$ with a dimension of $n \times m$. Consider, for example, the measurement data of the *j*-th geotechnical test at the *i*-th depth, $M_{i,j}$ (i.e., the *i*-th row and the *j*-th column of the data matrix) after the *z*-score normalization:

$$M'_{i,j} = \frac{M_{i,j} - \mu_j}{\sigma_j} \quad (13)$$

where $M'_{i,j}$ represents the normalized measurement of $M_{i,j}$; μ_j, and σ_j are the sample mean and standard deviation of all measurements for the *j*-th geotechnical test.

Fig. 6. Bootstrapping-based quantification of uncertainty in the recovered missing data.



In Module 2, observed measurements are equally divided into k (e.g., $k = 8$) equal parts (folds), and then the performance of the model with different possible levels of measurement noise (e.g., $\varepsilon = 0.01, 0.02, \ldots, 0.5$) is evaluated using k-fold cross-validation following the procedure described in Section 3.3.

In Module 3, the missing entries in the incomplete geotechnical data matrix, $\mathbf{M}'$ are predicted using the SVT algorithm and the selected ε value. Then, the recovered data are converted back to the original data space by applying the inverse transformation, $X_{m,n} = X'_{m,n}\sigma_n + \mu_n$, where $X'_{m,n}$ and $X_{m,n}$ represent the recovered data and corresponding value in the original space at the m-th row and the n-th column of the data matrix, respectively; μ_n and σ_n are the sample mean and standard deviation of measurements for the n-th geotechnical test parameters.

6. Uncertainty quantification by bootstrapping

The uncertainty in the prediction from the proposed method may be quantified by bootstrapping. Bootstrapping is a powerful statistical resampling technique commonly used to quantify uncertainty in an estimate of a property or variable of interest through multiple subsets from a single dataset (e.g., Carey 2004). The basic idea of bootstrapping-based uncertainty quantification is illustrated in Fig. 6. In bootstrapping, a large number of subsets of measurements, i.e., bootstrap samples, are generated by repeatedly sampling from the observed data with replacement. Each subset is then used to predict the missing data using the SVT algorithm. Based on the predictions, the distribution of the estimator can be obtained. In addition, to quantify the uncertainty in the selection of measurement error, ε, the ε value is also determined repeatedly based on the remaining measurements in each bootstrap iteration, rather than using the k-fold cross-validation strategy described above. Specifically, for each bootstrapping iteration, the available measurements are divided into two groups: bootstrap sample (e.g., 90% of the available measurements) used as the input for missing entries prediction and remaining sample (e.g., 10% of the available measurements) used for ε selection. The steps for bootstrapping-based uncertainty quantification are as follows:

Step 1: For a given geotechnical incomplete data matrix, $\mathbf{M}'$, generate a new set of samples drawing M_b data points from $\mathbf{M}'$ (e.g., 90% of the available measurements), leading to a new incomplete data matrix, $\mathbf{D}$.

Step 2: Repeat this process N_b times (e.g., $N_b = 10\,000$) and create N_b bootstrap samples, $\mathbf{D_1}, \mathbf{D_2}, \ldots, \mathbf{D}_{N_b}$

Step 3: Predict the missing entries for each bootstrap sample using the SVT algorithm. In this step, the SVT algorithm is repeatedly used to predict the missing data using various ε values, and the error between predicted and true values for the remaining (e.g., 10%) measurements is calculated for each ε scenario. Then, the ε value that results in the lowest prediction error is selected as the most probable measurement noise level.

Step 4: Repeat the process for each bootstrap sample, leading to N_b estimates for each missing entry.

Step 5: Calculate the statistical properties of interest (e.g., mean, variance, or confidence interval) for each missing entry based on the N_b bootstrap estimates.

In the following section, the proposed method is demonstrated and validated using a real geotechnical dataset.

Table 3. The incomplete geotechnical data matrix with 12 missing values.

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	$M_{1,3}$	1.06	51.00	29.00	1.52	70.49	38.23	0.54	35.22
5.76	90.88	91.66	1.01	58.00	35.00	1.51	69.05	31.98	0.46	28.16
6.47	94.78	95.17	1.00	66.00	41.00	1.29	38.05	29.25	0.77	24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56	81.29	44.08	$M_{4,10}$	38.52
7.69	102.59	103.37	1.01	68.00	43.00	1.23	42.98	40.18	0.93	$M_{5,11}$
8.36	106.49	107.66	1.01	$M_{6,5}$	42.00	1.29	93.72	37.45	0.40	31.34
8.89	109.61	111.17	1.01	62.00	36.00	$M_{7,7}$	57.84	44.47	0.77	38.86
9.41	112.73	115.07	1.02	55.00	30.00	0.93	33.71	63.58	1.89	$M_{8,11}$
9.92	116.24	118.19	1.02	57.00	25.00	0.84	22.50	56.56	2.51	$M_{9,11}$
10.36	118.97	121.70	$M_{10,4}$	44.00	$M_{10,6}$	1.23	63.02	$M_{10,9}$	0.93	$M_{10,11}$
10.88	122.48	125.60	1.03	58.00	33.00	1.18	52.17	$M_{11,9}$	0.93	43.58
11.96	130.67	133.01	1.02	51.00	25.00	0.88	32.27	60.85	1.89	57.86

Note: σ'_{vo}, vertical effective stress; σ'_p, effective pre-consolidation stress; OCR, over-consolidation ratio; LL, liquid limit; PI, plasticity index; LI, liquidity index; S_t, sensitivity; $s_{u,VST}$, undrained shear strength obtained from vane shear test; $s_{u(mob)}$, mobilized undrained shear strength.

7. Real geotechnical data from Beauharnois, Quebec, Canada

The proposed method developed in this study is used to recover the incomplete geotechnical datasets obtained from Beauharnois, Quebec, Canada, as shown in Table 2. This geotechnical dataset includes 10 different geotechnical properties (i.e., vertical effective stress, effective pre-consolidation stress, …, undrained shear strength), and each test is conducted at 12 different depths (i.e., 5.24 m, 5.76 m, …, 11.96 m). Then, this dataset is treated as a matrix with a dimension of 12 × 11 (10 geotechnical properties plus one additional column for depth). To mimic the missing data in engineering practice, $N_m = 12$ measurements from different tests are randomly removed, except for the data on depth and vertical effective stress, which are typically available. This results in the incomplete data matrix shown in Table 3. The missing values are highlighted by red shadings in Table 3.

7.1. Prediction of missing values using the proposed method

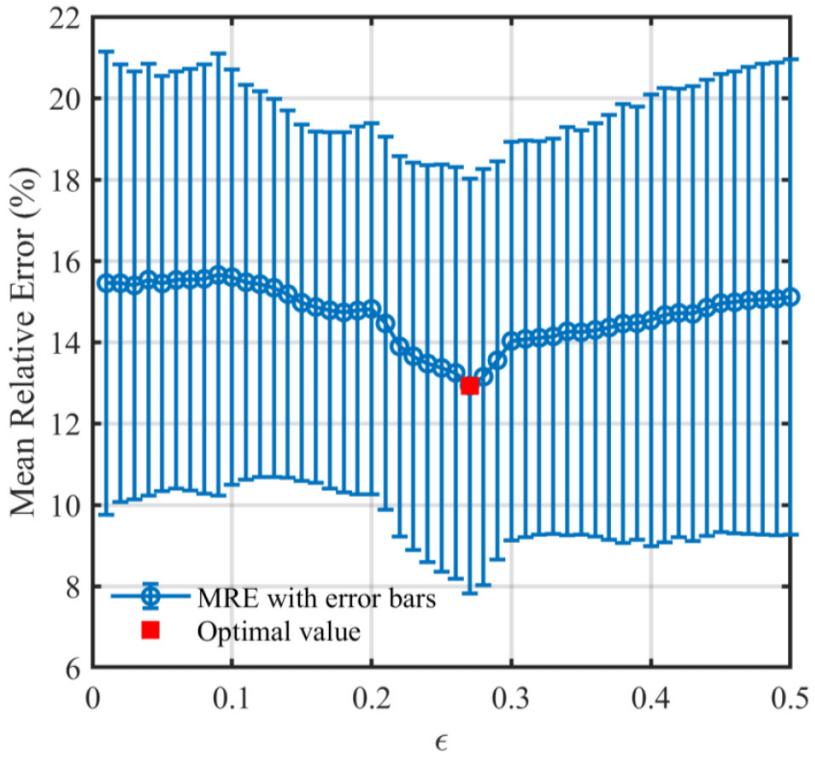
In Module 1, z-score normalization is used to transform each measurement of different geotechnical properties to a common scale. For example, the fifth geotechnical property (i.e., liquid limit (LL)) has a total number of 11 measurements, and the sample mean, μ_5 and standard deviation, σ_5 of these measurements are calculated as 57.5 and 7.0, respectively, and then each measurement of LL is normalized using eq. 13. For instance, the measurement at the depth of 5.24 m, $M_{1,5} = 51$, is normalized to the corresponding value as $M'_{1,5} = (51 - 57.5)/7.0 = -0.93$. This procedure is repeated for each measurement, ensuring that the measurements of each geotechnical property have a mean of zero and a standard deviation of one.

In Module 2, the observed measurements of different geotechnical properties except for the vertical effective stress are equally divided into $k = 8$ equal parts (folds), i.e., each fold contains 12 measurements of different geotechnical properties. Then, the performance of the model is evaluated with different possible levels of measurement noise, $\varepsilon = 0.01, 0.02, \ldots, 0.5$ using k-fold cross-validation. For a given fold i, $i = 1, 2, \ldots, 8$, the model is trained using the remaining data except those from the i-th fold. In other words, the data in the i-th fold are predicted from the remaining measurements using the SVT algorithm for a given ε. The mean relative error (MRE) is used to quantitatively compare the recovered values with the true ones for a given fold. It is defined as the average of the absolute differences between predicted and actual values, divided by the actual values:

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \times 100\% \quad (14)$$

where $\widehat{y}_i$ is the i-th predicted data; y_i is the i-th original data; n is the number of observations for a given fold. This process is repeated for each of the 8 folds, leading to 8 MREs for a given measurement noise level. After that, a similar procedure is repeatedly conducted for different levels of measurement noise, and the results are summarized in Fig. 7. The figure shows MREs with a mean plus/minus one standard deviation error bar as a function of the level of measurement noise, ε. It can be observed from this figure that, in general, MRE firstly decreases with increasing ε and then starts to rise. When ε is small, it indicates that the model exactly or closely fits the training data, leading to the high risk of over-fitting. As ε is slightly increased, the model starts to reduce the impact of noise or overfitting, which initially improves its performance on validation data. However, when ε exceeds 0.27, the MRE begins to increase, indicating that the model has become too simple to capture the underlying pattern of the dataset. The point of $\varepsilon = 0.27$ yields the lowest mean MRE, where the model has the best generalization to new or unseen data (e.g., Giagkiozis and Fleming 2014). At this optimal point, the model is neither too complex (avoiding overfitting) nor too simple (avoiding underfitting). Using the k-fold cross-validation strategy, the most probable level of measurement noise is selected as $\varepsilon = 0.27$.

Fig. 7. Cross-validation estimate of mean relative error (MRE) with mean plus/minus one standard deviation error bar as a function of the level of measurement noise, ε (the optimal value is the one with the minimal mean value of MRE).



In Module 3, the missing entries in the incomplete geotechnical data matrix, $\mathbf{M}'$ are recovered using the SVT algorithm. The input parameters are determined as $\tau = 5\sqrt{n \times m} = 57$, $\delta = 0.1$ and $\varepsilon = 0.27$. Following the procedure described in Section 3.2, the 12 missing entries are recovered, as shown in the third column of Table 4. For comparison, the original data are shown in the second column of Table 4. It can be observed that the predicted values are generally consistent with the true ones, and the MRE of the 12 predicted values is calculated as 13.3%. The prediction process takes only a few seconds on a personal computer equipped with an Intel® Core™ i7-6700 3.4 GHz CPU and 16.0 GB RAM, demonstrating the high computational efficiency of the proposed method. These results indicate that the proposed method can efficiently provide reasonable estimates of missing values from available observations.

In addition, the uncertainty in the predictions from the proposed method is quantified by bootstrapping. In bootstrapping, a large number of ($N_b = 10\,000$) bootstrap samples are generated by repeatedly drawing 108 data points from the original incomplete data matrix (90% of the available measurements), leading to 10,000 new incomplete data matrices, $\mathbf{D_1}$, $\mathbf{D_2}$,…, $\mathbf{D}_{10\,000}$. Then, the original 12 missing entries of each bootstrap sample are predicted using the SVT algorithm, leading to 10 000 estimates for each missing entry. For example, the histogram of bootstrap estimates for the mobilized undrained shear strength, $s_{u(mob)}$ at a depth of 7.69 m is shown in Fig. 8, with a mean value of 30.92 and a standard deviation of 1.98. The mean values, standard deviations and 95% confidence interval of the 12 missing data points are summarized in the Table 4. The MRE of the predicted mean values is calculated as 13.8%. It is observed that five out of twelve true data points (41.6%) fall within the 95% confidence interval, suggesting that the uncertainty may be slightly underestimated.

7.2. Effect of ε

In this subsection, the proposed method is compared with the existing SVT algorithm using a small ε value for geotechnical data matrix completion. In the traditional SVT algorithm, a fixed ε value (e.g., $\varepsilon = 10^{-4}$) is usually adopted for matrix completion, meaning that measurement error is typically ignored. Therefore, the input parameters are determined as $\tau = 57$, $\delta = 0.1$, and $\varepsilon = 10^{-4}$. Following the procedure described in Section 3.2, the 12 missing entries are recovered, as shown in the fourth column of Table 4. It can be observed from Table 4 that the predicted values with $\varepsilon = 0.27$ are generally more accurate than those with $\varepsilon = 10^{-4}$. The MRE calculated for the scenario with $\varepsilon = 10^{-4}$ is 14.6%, which is larger than the MRE of 13.3% when considering the measurement noise level. This result indicates that accounting for measurement noise can improve prediction accuracy.

7.3. Effect of the number of missing data

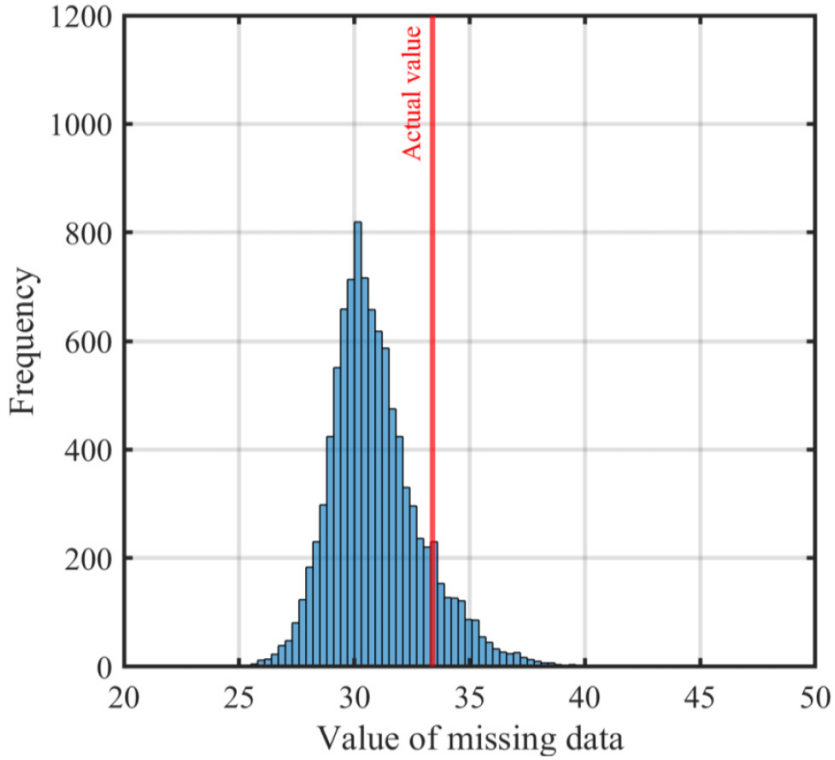
In this subsection, the effect of the number of missing data on the prediction accuracy are investigated. Four more scenarios of the number of missing data, $N_m = 36$ (missing ratio = 27%), $N_m = 48$ (missing ratio = 36%), $N_m = 60$ (missing ratio = 45%), and $N_m = 72$ (missing ratio = 55%), are considered in this subsection. The corresponding incomplete data matrices for these different scenarios are illustrated

Table 4. Reconstructed missing entries using different methods.

Missing entry	Original value	Gaussian process regression	SVT with k-fold cross-validation	SVT with $\varepsilon = 10^{-4}$	Uncertainty quantification using bootstrapping		
					Mean	Standard deviation	95% confidence interval
$M_{1,3}$	92.83	88.76	103.77	110.63	105.79	4.43	[96.76, 114.76]
$M_{4,10}$	0.54	0.93	0.51	0.47	0.58	0.08	[0.50, 0.81]
$M_{5,11}$	33.40	33.44	29.71	29.87	30.92	1.98	[27.66, 35.54]
$M_{6,5}$	68.00	61.28	59.52	58.22	59.39	2.04	[55.29, 63.29]
$M_{7,7}$	1.31	1.16	1.23	1.30	1.24	0.04	[1.17, 1.33]
$M_{8,11}$	58.14	40.02	43.52	43.83	43.70	3.20	[33.87, 50.63]
$M_{9,11}$	53.78	42.47	44.69	44.21	44.84	3.87	[36.67, 48.72]
$M_{10,4}$	1.02	1.02	1.03	1.02	1.03	0.01	[1.02, 1.04]
$M_{10,6}$	22.00	27.74	29.68	29.19	29.81	2.01	[26.29, 34.35]
$M_{10,9}$	58.90	57.72	51.81	48.83	50.21	2.75	[44.27, 55.03]
$M_{10,11}$	57.38	44.56	46.52	42.98	44.26	3.02	[36.67, 48.72]
$M_{11,9}$	48.76	59.25	50.63	49.00	50.24	2.22	[45.55, 54.35]
Mean relative error		18.5%	13.3%	14.6%	13.8%		

Note: SVT, singular value thresholding.

Fig. 8. Histogram of bootstrap estimates for the mobilized undrained shear strength at a depth of 7.69 m.



in Table 5. Following the procedure described in Section 5, the missing data for each scenario are repeatedly recovered using the SVT algorithm. Table 6 illustrates the complete geotechnical data matrix reconstructed with different missing ratios, and the true values are shown in the brackets. It can be observed from Table 6 that the predicted values are generally consistent with the true ones. The MRE calculated for missing ratios of 27%, 36%, 45%, and 55% is 14.3%, 15.2%, 16.0%, and 23.0%, respectively. It is observed that the MRE increases from 13.3% to 23.0% as the number of missing data increases from 12 to 72. When the missing ratio is below 45%, the proposed method can provide reasonable predictions. However, as the missing ratio increases to 55%, the available measurements become too limited, which may lead to biased prediction results.

Furthermore, following the procedure described in Section 6, the uncertainty in predictions for different numbers of missing data can be quantified by bootstrapping. For example, the histograms of bootstrap estimates for $s_{u(mob)}$ at a depth of 7.69 m, corresponding to different numbers of

Table 5. The incomplete geotechnical data matrix with different missing ratios.

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37		1.06	51.00	29.00	1.52		38.23	0.54	
5.76	90.88	91.66	1.01	58.00	35.00	1.51		31.98	0.46	
6.47	94.78	95.17	1.00	66.00	41.00	1.29		29.25		24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56		44.08		38.52
7.69	102.59	103.37		68.00	43.00	1.23		40.18		
8.36	106.49	107.66			42.00	1.29	93.72	37.45		31.34
8.89	109.61		1.01	62.00	36.00		57.84	44.47	0.77	
9.41	112.73		1.02	55.00	30.00	0.93	33.71	63.58		
9.92	116.24	118.19	1.02		25.00		22.50	56.56	2.51	
10.36	118.97	121.70		44.00		1.23	63.02		0.93	
10.88	122.48	125.60	1.03	58.00	33.00	1.18	52.17			
11.96	130.67	133.01			25.00	0.88		60.85		57.86

(a) 27% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37		1.06	51.00					0.54	
5.76	90.88	91.66	1.01	58.00	35.00	1.51		31.98	0.46	
6.47	94.78	95.17	1.00	66.00	41.00	1.29		29.25		24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56		44.08		38.52
7.69	102.59			68.00	43.00			40.18		
8.36	106.49	107.66			42.00	1.29	93.72	37.45		31.34
8.89	109.61		1.01	62.00			57.84	44.47	0.77	
9.41	112.73		1.02	55.00	30.00	0.93	33.71	63.58		
9.92	116.24		1.02		25.00		22.50	56.56	2.51	
10.36	118.97	121.70		44.00			63.02		0.93	
10.88	122.48	125.60				1.18	52.17			
11.96	130.67	133.01				0.88		60.85		57.86

(b) 36% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37		1.06	51.00					0.54	
5.76	90.88	91.66		58.00	35.00			31.98	0.46	
6.47	94.78	95.17		66.00	41.00	1.29		29.25		24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56		44.08		
7.69	102.59				43.00			40.18		
8.36	106.49	107.66					93.72			31.34
8.89	109.61		1.01	62.00			57.84	44.47	0.77	
9.41	112.73				30.00	0.93	33.71	63.58		
9.92	116.24		1.02		25.00			56.56	2.51	
10.36	118.97			44.00			63.02		0.93	
10.88	122.48	125.60				1.18	52.17			
11.96	130.67	133.01				0.88		60.85		57.86

(c) 45% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37		1.06	51.00					0.54	
5.76	90.88	91.66		58.00				31.98	0.46	
6.47	94.78	95.17			41.00	1.29		29.25		
7.11	98.69	99.47	1.01	62.00	36.00	1.56		44.08		
7.69	102.59				43.00			40.18		

Table 5. (*concluded*).

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
8.36	106.49	107.66					93.72			31.34
8.89	109.61		1.01				57.84	44.47	0.77	
9.41	112.73							63.58		
9.92	116.24		1.02		25.00					
10.36	118.97						63.02			
10.88	122.48					1.18	52.17			
11.96	130.67	133.01				0.88		60.85		57.86

(d) 55% missing data

Note: σ'_{vo}, vertical effective stress; σ'_p, effective pre-consolidation stress; OCR, over-consolidation ratio; LL, liquid limit; PI, plasticity index; LI, liquidity index; S_t, sensitivity; $s_{u,VST}$, undrained shear strength obtained from vane shear test; $s_{u(mob)}$, mobilized undrained shear strength.

Table 6. The recovered geotechnical data matrix (the values in the bracket indicate true ones).

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	103.82(92.83)	1.06	51.00	29.00	1.52	56.51(70.49)	38.23	0.54	35.51(35.22)
5.76	90.88	91.66	1.01	58.00	35.00	1.51	59.87(69.05)	31.98	0.46	28.87(28.16)
6.47	94.78	95.17	1.00	66.00	41.00	1.29	64.66(38.05)	29.25	0.69(0.77)	24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56	52.79(81.29)	44.08	0.92(0.54)	38.52
7.69	102.59	103.37	1(1.01)	68.00	43.00	1.23	63.12(42.98)	40.18	0.89(0.93)	31.46(33.4)
8.36	106.49	107.66	1.01(1.01)	60.69(68)	42.00	1.29	93.72	37.45	0.37(0.4)	31.34
8.89	109.61	109.77(111.17)	1.01	62.00	36.00	1.24(1.31)	57.84	44.47	0.77	36.55(38.86)
9.41	112.73	120.53(115.07)	1.02	55.00	30.00	0.93	33.71	63.58	1.75(1.89)	46.85(58.14)
9.92	116.24	118.19	1.02	57.41(57)	25.00	1.05(0.84)	22.50	56.56	2.51	48.65(53.78)
10.36	118.97	121.70	1.04(1.02)	44.00	30.72(22)	1.23	63.02	49.27(58.9)	0.93	44.7(57.38)
10.88	122.48	125.60	1.03	58.00	33.00	1.18	52.17	51.36(48.76)	1.24(0.93)	44.06(43.58)
11.96	130.67	133.01	1.02(1.02)	52.12(51)	25.00	0.88	34.33(32.27)	60.85	1.89(1.89)	57.86

(a) 27% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	101.64(92.83)	1.06	51.00	35.27(29)	1.37(1.52)	60.15(70.49)	40.5(38.23)	0.54	32.55(35.22)
5.76	90.88	91.66	1.01	58.00	35.00	1.51	56.39(69.05)	31.98	0.46	27.97(28.16)
6.47	94.78	95.17	1.00	66.00	41.00	1.29	63.93(38.05)	29.25	0.64(0.77)	24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56	52.53(81.29)	44.08	0.95(0.54)	38.52
7.69	102.59	102.4(103.37)	1(1.01)	68.00	43.00	1.36(1.23)	64.77(42.98)	40.18	0.76(0.93)	31.79(33.4)
8.36	106.49	107.66	1.02(1.01)	59.61(68)	42.00	1.29	93.72	37.45	0.33(0.4)	31.34
8.89	109.61	110.07(111.17)	1.01	62.00	37.86(36)	1.25(1.31)	57.84	44.47	0.77	37.02(38.86)
9.41	112.73	120.49(115.07)	1.02	55.00	30.00	0.93	33.71	63.58	1.73(1.89)	46.94(58.14)
9.92	116.24	118.53(118.19)	1.02	57.14(57)	25.00	1.05(0.84)	22.50	56.56	2.51	48.27(53.78)
10.36	118.97	121.70	1.04(1.02)	44.00	35.69(22)	1.05(1.23)	63.02	52.84(58.9)	0.93	45.48(57.38)
10.88	122.48	125.60	1.02(1.03)	56.48(58)	35.46(33)	1.18	52.17	52.27(48.76)	1.29(0.93)	44.87(43.58)
11.96	130.67	133.01	1.02(1.02)	52.62(51)	31.81(25)	0.88	44.23(32.27)	60.85	1.71(1.89)	57.86

(b) 36% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	94.49(92.83)	1.06	51.00	39.56(29)	1.18(1.52)	65.55(70.49)	37.94(38.23)	0.54	30.32(35.22)
5.76	90.88	91.66	1.04(1.01)	58.00	35.00	1.33(1.51)	70.21(69.05)	31.98	0.46	27.36(28.16)
6.47	94.78	95.17	1.02(1)	66.00	41.00	1.29	73.68(38.05)	29.25	0.56(0.77)	24.66
7.11	98.69	99.47	1.01	62.00	36.00	1.56	63.28(81.29)	44.08	1(0.54)	30.65(38.52)
7.69	102.59	102.01(103.37)	1.03(1.01)	58.27(68)	43.00	1.26(1.23)	72.03(42.98)	40.18	0.61(0.93)	32.44(33.4)
8.36	106.49	107.66	1.03(1.01)	57.7(68)	39.8(42)	1.25(1.29)	93.72	38.67(37.45)	0.59(0.4)	31.34
8.89	109.61	108.19(111.17)	1.01	62.00	35.5(36)	1.23(1.31)	57.84	44.47	0.77	36.72(38.86)

Table 6. (*concluded*).

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
9.41	112.73	118.79(115.07)	1.02(1.02)	53.51(55)	30.00	0.93	33.71	63.58	1.78(1.89)	47.22(58.14)
9.92	116.24	120.58(118.19)	1.02	57.81(57)	25.00	1.02(0.84)	35.88(22.5)	56.56	2.51	47.7(53.78)
10.36	118.97	120.99(121.7)	1.04(1.02)	44.00	33.88(22)	0.98(1.23)	63.02	52.9(58.9)	0.93	45.95(57.38)
10.88	122.48	125.60	1.01(1.03)	55.75(58)	31.41(33)	1.18	52.17	54.87(48.76)	1.43(0.93)	45.99(43.58)
11.96	130.67	133.01	1.01(1.02)	51.21(51)	27.99(25)	0.88	48.88(32.27)	60.85	1.7(1.89)	57.86

(c) 45% missing data

Depth	σ'_{vo}	σ'_p	OCR	LL	PI	LI	S_t	$s_{u,VST}$	Remolded $s_{u,VST}$	$s_{u(mob)}$
5.24	87.37	94.25(92.83)	1.06	51.00	46.05(29)	1.21(1.52)	81.42(70.49)	28.46(38.23)	0.54	36.51(35.22)
5.76	90.88	91.66	1.03(1.01)	58.00	42.01(35)	1.42(1.51)	74.32(69.05)	31.98	0.46	37.16(28.16)
6.47	94.78	95.17	1.04(1)	55.61(66)	41.00	1.29	70.25(38.05)	29.25	0.54(0.77)	40.01(24.66)
7.11	98.69	99.47	1.01	62.00	36.00	1.56	63.14(81.29)	44.08	0.58(0.54)	41.73(38.52)
7.69	102.59	99.62(103.37)	1.03(1.01)	55.69(68)	43.00	1.26(1.23)	68.81(42.98)	40.18	0.56(0.93)	42.38(33.4)
8.36	106.49	107.66	1.04(1.01)	55.4(68)	36.49(42)	1.19(1.29)	93.72	46.44(37.45)	0.51(0.4)	31.34
8.89	109.61	105.46(111.17)	1.01	58.58(62)	34.83(36)	1.26(1.31)	57.84	44.47	0.77	50.45(38.86)
9.41	112.73	113.88(115.07)	1.02(1.02)	57.95(55)	32(30)	1.15(0.93)	66.22(33.71)	63.58	0.62(1.89)	46.98(58.14)
9.92	116.24	114.59(118.19)	1.02	59.31(57)	25.00	1.18(0.84)	62.22(22.5)	55.98(56.56)	0.64(2.51)	48.55(53.78)
10.36	118.97	113.63(121.7)	1.02(1.02)	57.48(44)	32.69(22)	1.13(1.23)	63.02	52.66(58.9)	0.63(0.93)	49.05(57.38)
10.88	122.48	114.56(125.6)	1.01(1.03)	58.34(58)	31.61(33)	1.18	52.17	54.09(48.76)	0.67(0.93)	53.32(43.58)
11.96	130.67	133.01	1.02(1.02)	55.95(51)	28.78(25)	0.88	59.41(32.27)	60.85	0.68(1.89)	57.86

(d) 55% missing data

Note: σ'_{vo}, vertical effective stress; σ'_p, effective pre-consolidation stress; OCR, over-consolidation ratio; LL, liquid limit; PI, plasticity index; LI, liquidity index; S_t, sensitivity; $s_{u,VST}$, undrained shear strength obtained from vane shear test; $s_{u(mob)}$, mobilized undrained shear strength.

missing data, are summarized in Fig. 9. The mean values of the predicted $s_{u(mob)}$ for missing ratios of 27%, 36%, 45%, and 55% are 32.02, 33.09, 33.11, and 42.42, respectively, while the corresponding standard deviations are 2.91, 2.58, 3.72, and 5.85, respectively. It is observed that the standard deviation increases from 1.98 to 5.85 as the number of missing data increases from 12 to 72. This indicates a significant increase in uncertainty as the number of missing data points rises. Moreover, when the missing ratio reaches 55%, the distribution of bootstrap estimates becomes entirely biased compared to other scenarios, suggesting that predictions with such a high missing ratio may not be reliable.

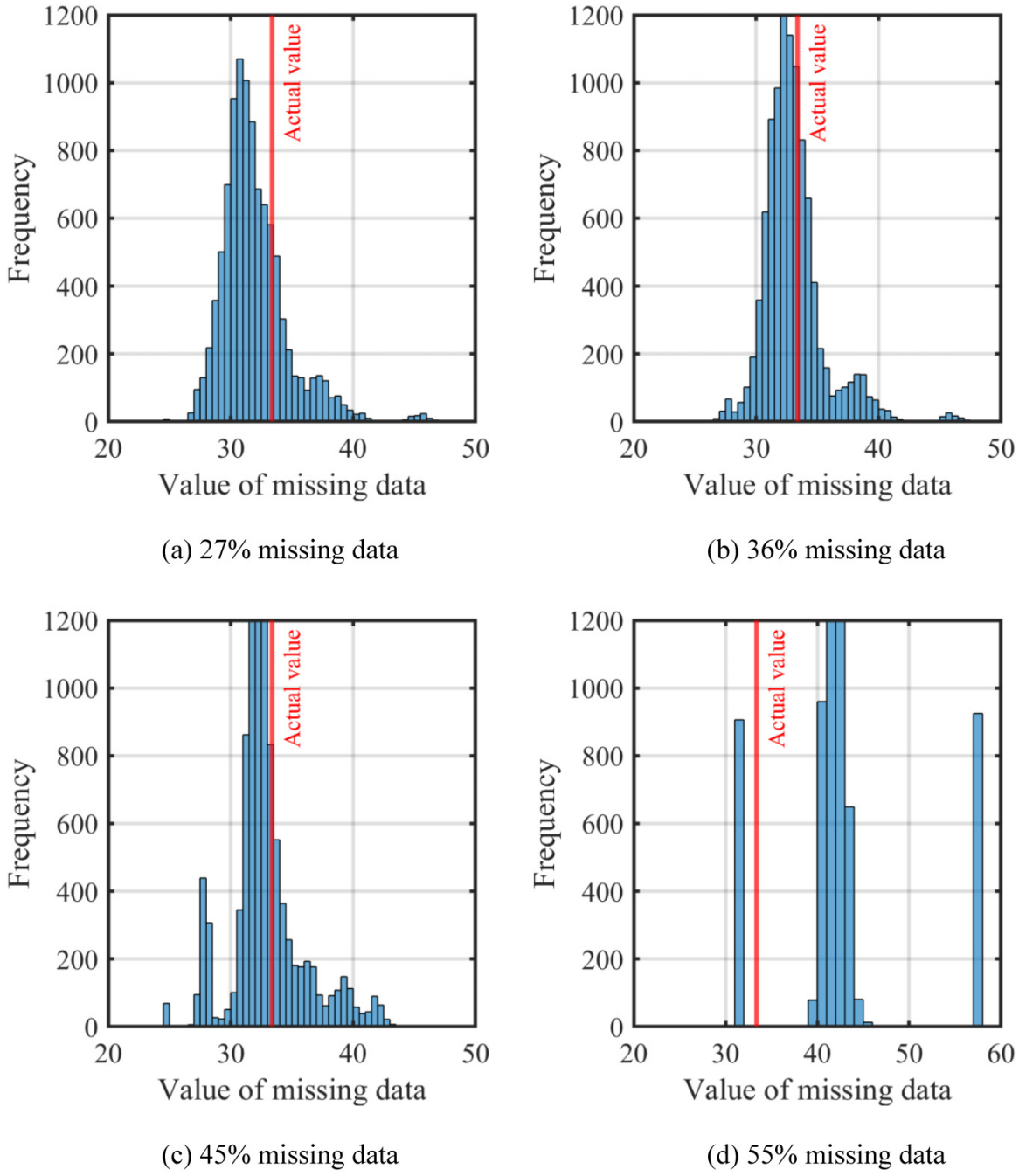
7.4. Comparison with the existing method

The proposed method is compared with Gaussian process regression (GPR), a widely adopted approach for spatial data prediction (e.g., Yoshida et al. 2021; Ching et al. 2023). Following the recommendation by Yoshida et al. (2021), the spatial trend is modeled as a stationary zero-mean Gaussian random field, with a Gaussian covariance function used to capture the auto-correlation along the depth direction. In this comparison, the missing values of each soil property at different depths are predicted independently using GPR. The prediction results are presented in Table 4. The MRE obtained using GPR is 18.5%, which is higher than that of the proposed method, indicating the superior predictive performance of the proposed method. Different geotechnical properties are usually cross-correlated, and the proposed method has a unique capability to leverage the cross-correlations among different geotechnical properties to enhance the prediction accuracy.

8. Synthetic data example

Synthetic geotechnical datasets are used to systematically evaluate the performance of the proposed method under various conditions. These include variations in the number of geotechnical parameters, the degree of auto-correlation along depth, and the level of cross-correlation among different soil properties. Incomplete synthetic datasets involving 20 geotechnical variables are simulated. The mean values of different properties range from 0.1 to 10, and the coefficients of variation span from 0.2 to 0.4. In Scenario #1, a vertical correlation length of 2 m and a cross-correlation coefficient of 0.7 are used to simulate the correlation along depth and across properties. In Scenario #2, the vertical correlation length is decreased to 0.5 m (with the cross-correlation coefficient maintained at 0.7) to represent weak autocorrelation along depth. In Scenario #3, the vertical correlation length remains at 2 m, but the cross-correlation coefficient is reduced to 0.3 to represent weak inter-property correlation. The incomplete datasets for different scenarios are shown in Table 7. The proposed method is then used to recover the missing values for different scenarios, as illustrated in Table 8. The MRE between the original and reconstructed data for Scenario #1, Scenario #2, and Scenario #3 is calculated as 17.4%, 21.7%, and 32.3%. Smaller vertical correlation implies weaker correlation in soil properties along depth, providing reduced

Fig. 9. Histograms of bootstrap estimates for the mobilized undrained shear strength at a depth of 7.69 m corresponding to different amounts of missing data.



prediction accuracy, and thus the MRE of Scenario #2 is larger than that of Scenario #1. On the other hand, when inter-property correlation is weak, the reliability of the model naturally decreases, leading to relatively high prediction error for Scenario #3.

9. Conclusion

This study developed a novel method for recovering missing values in geotechnical datasets with measurements from different tests. The auto- and cross-correlation structures of multivariate geotechnical data are exploited using a low-rank modeling approach in a purely data-driven manner. The only assumption of the proposed method is that the geotechnical datasets have an underlying low-rank structure, which is typically the case for real-world geotechnical data. By leveraging this low-rank structure, missing values can be efficiently recovered from available measurements using SVT algorithms. To account for measurement noise in the dataset, a *k*-fold cross-validation strategy is employed to select the most probable level of noise for matrix completion, which can significantly reduce the risk of overfitting

Table 7. The simulated geotechnical datasets with 30% random missing values.

Depth	Property #1	#2	#3	#4	#5	#6	#7	#8	…	#20
0	0.11		0.19	0.25	0.34	0.40	0.54	0.66	…	
1	0.08		0.14	0.17				0.43	…	5.89
2	0.09	0.11	0.13	0.23			0.43		…	5.51
3	0.11		0.15	0.23			0.40		…	8.13
4	0.09	0.13	0.14			0.28	0.42	0.54	…	
5	0.11	0.16	0.17	0.25		0.35	0.43	0.60	…	11.43
6		0.12			0.29	0.30	0.37		…	9.67
7	0.09	0.14	0.14	0.18	0.29	0.34	0.47	0.59	…	8.08
8	0.08	0.09		0.15	0.22	0.22			…	6.11
9		0.16	0.20	0.21		0.43	0.58	0.73	…	
10	0.11	0.12	0.16			0.40	0.44	0.56	…	

(a) Scenario #1

Depth	Property #1	#2	#3	#4	#5	#6	#7	#8	…	#20
0	0.13		0.24	0.27	0.30	0.54	0.60	0.78	…	
1	0.05		0.12	0.17				0.53	…	7.43
2	0.08	0.10	0.16	0.19			0.45		…	5.70
3	0.12		0.21	0.26			0.47		…	10.78
4	0.10	0.15	0.20			0.39	0.52	0.59	…	
5	0.09	0.10	0.14	0.13		0.27	0.23	0.39	…	5.28
6		0.15			0.26	0.33	0.35		…	14.04
7	0.10	0.15	0.17	0.22	0.28	0.41	0.48	0.53	…	13.08
8	0.12	0.14		0.26	0.34	0.37			…	10.02
9		0.12	0.16	0.23		0.40	0.35	0.51	…	
10	0.10	0.11	0.10			0.28	0.44	0.46	…	

(b) Scenario #2

Depth	Property #1	#2	#3	#4	#5	#6	#7	#8	…	#20
0	0.10		0.15	0.19	0.17	0.34	0.58	0.66	…	
1	0.12		0.18	0.23				0.70	…	5.88
2	0.11	0.07	0.17	0.23			0.43		…	9.90
3	0.09		0.13	0.26			0.23		…	7.75
4	0.10	0.12	0.09			0.32	0.34	0.27	…	
5	0.09	0.13	0.06	0.15		0.29	0.33	0.49	…	10.08
6		0.15			0.17	0.27	0.64		…	9.59
7	0.07	0.09	0.07	0.17	0.20	0.22	0.41	0.67	…	10.45
8	0.09	0.08		0.21	0.22	0.34			…	11.14
9		0.10	0.14	0.17		0.20	0.34	0.77	…	
10	0.10	0.09	0.16			0.12	0.39	0.68	…	

(c) Scenario #3

and improve the generalizability of the model. The effectiveness of the proposed method was demonstrated and validated using real geotechnical datasets. A sensitivity study was also conducted to explore the effects of the number of missing data. The results indicate that the proposed method can provide reasonable predictions even with a high ratio of missing data. In addition, the proposed method can be combined with the bootstrapping technique to probabilistically predict the missing entries. An extremely high missing data ratio might be observed in geotechnical data collected. Therefore, integrating information from existing geotechnical databases is essential for further improving accuracy of the prediction results. Future studies are needed to extend the proposed method to effectively incorporate and leverage such databases (e.g., Ching and Phoon 2014; Otake et al. 2025). The proposed method is directly applicable to

Table 8. The recovered data using the proposed method

	Scenario #1		Scenario #2		Scenario #3	
Missing entry	Original value	Prediction	Original value	Prediction	Original value	Prediction
$M_{7,1}$	1.57	1.51	1.57	1.08	1.15	1.17
$M_{10,1}$	1.28	1.42	1.19	1.29	1.57	1.55
$M_{1,2}$	2.55	1.73	2.89	2.56	0.71	2.14
$M_{2,2}$	0.09	0.11	0.10	0.10	0.10	0.09
$M_{4,2}$	0.11	0.13	0.15	0.14	0.08	0.10
$M_{7,3}$	9.20	7.91	7.45	5.92	5.88	6.30
$M_{9,3}$	0.18	0.21	0.21	0.20	0.20	0.21
$M_{5,4}$	6.16	4.86	7.30	6.82	4.98	3.53
$M_{7,4}$	0.69	0.76	0.83	0.80	1.04	0.54
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$M_{11,20}$	7.48	8.48	9.50	8.23	13.14	4.07
Mean relative error		17.4%		21.7%		32.3%

multilayered ground conditions (i.e., non-stationary data), as its only underlying assumption is the presence of a low-rank structure in the multivariate geotechnical database. This assumption is valid for auto- and cross-correlated geotechnical data typically observed in multilayered subsurface conditions.

Acknowledgements

The work described in this paper was supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region (Project Nos: 11203322 and 11207724). The financial support is gratefully acknowledged.

Article information

History dates

Received: 18 December 2024
Accepted: 17 May 2025
Accepted manuscript online: 27 May 2025
Version of record online: 14 July 2025

Copyright

Data availability

The data are available from the corresponding author on reasonable request.

Author information

Author ORCIDs

Zheng Guan https://orcid.org/0000-0002-1346-7078
Yu Wang https://orcid.org/0000-0003-4635-7059
Kok-Kwang Phoon https://orcid.org/0000-0003-2577-8639

Author notes

Yu Wang served as Editorial Board Member at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by another editorial board member.

Author contributions

Conceptualization: ZG, YW, KKP
Data curation: ZG
Formal analysis: ZG
Funding acquisition: YW
Investigation: ZG, YW
Methodology: ZG, YW, KKP
Project administration: YW
Resources: YW
Supervision: YW
Validation: ZG, YW
Visualization: ZG
Writing – original draft: ZG, YW, KKP
Writing – review & editing: ZG, YW, KKP

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding information

There are no funders to report for this submission.

References

Akritas, A.G., and Malaschonok, G.I. 2004. Applications of singular-value decomposition (SVD). Mathematics and Computers in Simulation, **67**(1-2): 15–31. doi:10.1016/j.matcom.2004.05.005.

Bengio, Y., Goodfellow, I., and Courville, A. 2017. Deep learning. Vol. 1, MIT Press.

Cai, J.F., Candès, E.J., and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, **20**(4): 1956–1982. doi:10.1137/080738970.

Candès, E.J., and Plan, Y. 2010. Matrix completion with noise. Proceedings of the IEEE, **98**(6): 925–936. doi:10.1109/JPROC.2009.2035722.

Candès, E.J., and Recht, B. 2008. Exact low-rank matrix completion via convex optimization. *In* 2008 46th Annual Allerton Conference on Communication, Control, and Computing. IEEE.

Candès, E.J., and Tao, T. 2010. The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory, **56**(5): 2053–2080. doi:10.1109/TIT.2010.2044061.

Candes, E.J., Sing-Long, C.A., and Trzasko, J.D. 2013. Unbiased risk estimates for singular value thresholding and spectral estimators. IEEE Transactions on Signal Processing, **61**(19): 4643–4657. doi:10.1109/TSP.2013.2270464.

Carey, S. 2004. Bootstrapping & the origin of concepts. Daedalus, **133**(1): 59–68. doi:10.1162/001152604772746701.

Chatterjee, A. 2000. An introduction to the proper orthogonal decomposition. Current Science, **78**(7): 808–817.

Ching, J., and Phoon, K.K. 2014. Transformations and correlations among some clay parameters: the global database. Canadian Geotechnical Journal, **51**(6): 663–685. doi:10.1139/cgj-2013-0262.

Ching, J., and Phoon, K.K. 2019. Constructing site-specific multivariate probability distribution model using bayesian machine learning. Journal of Engineering Mechanics, **145**(1): 04018126. doi:10.1061/(ASCE)EM.1943-7889.0001537.

Ching, J., Wu, T.J., Stuedlein, A.W., and Bong, T. 2018. Estimating horizontal scale of fluctuation with limited CPT soundings. Geoscience Frontiers, **9**(6): 1597–1608. doi:10.1016/j.gsf.2017.11.008.

Ching, J., Yoshida, I., and Phoon, K.K. 2023. Comparison of trend models for geotechnical spatial variability: sparse bayesian learning versus Gaussian process regression. Gondwana Research, **123**: 174–183. doi:10.1016/j.gr.2022.07.011.

Combettes, P.L., and Wajs, V.R. 2005. Signal recovery by proximal forward-backward splitting. Multiscale Modeling and Simulation, **4**(4): 1168–1200.

Feuerverger, A., He, Y., and Khatri, S. 2012. Statistical significance of the Netflix challenge. Proceedings of the National Academy of Sciences, **109**(11): 4582–4587.

Giagkiozis, I., and Fleming, P.J. 2014. Pareto front estimation for decision making. Evolutionary Computation, **22**(4): 651–678. doi:10.1162/EVCO_a_00128.

Gomez-Uribe, C.A., and Hunt, N. 2015. The Netflix recommender system: algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS), **6**(4): 1–19.

Guan, Z., and Wang, Y. 2021. Non-parametric construction of site-specific non-Gaussian multivariate joint probability distribution from sparse measurements. Structural Safety, **91**: 102077. doi:10.1016/j.strusafe.2021.102077.

Guan, Z., Wang, Y., and Phoon, K.K. 2024. Fusion of sparse non-co-located measurements from multiple sources for geotechnical site investigation. Canadian Geotechnical Journal, **61**(8): 1574–1592. doi:10.1139/cgj-2023-0289.

Jolliffe, I., and Cadima, J. 2016. Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**(2065): 0150202. doi:10.1098/rsta.2015.0202.

Lacasse, S., and Lunne, T. 1982. Penetration tests in two Norwegian clays. *In* Proceedings of the 2nd European Symposium on Penetration Testing. Amsterdam. pp. 661–670.

Larsen, R., and Marx, M.L. 2005. An introduction to mathematical statistics. Vol. **106**. Prentice Hall, Hoboken, NJ.

Lingala, S.G., Hu, Y., DiBella, E., and Jacob, M. 2011. Accelerated dynamic MRI exploiting sparsity and low-rank structure: kt SLR. IEEE Transactions on Medical Imaging, **30**(5): 1042–1054. doi:10.1109/TMI.2010.2100850.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**(1): 171–184. doi:10.1109/TPAMI.2012.88.

Mu, H.Q., Zhao, Z.T., and Yuen, K.V. 2024. Characterizing multivariate, asymmetric, and multimodal distributions of geotechnical data with dual-stage missing values: BASIC-H. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, **18**(1): 85–106.

Nguyen, L.T., Kim, J., and Shim, B. 2019. Low-rank matrix completion: a contemporary survey. IEEE Access, **7**: 94215–94237. doi:10.1109/ACCESS.2019.2928130.

Orchant, C.J., Kulhawy, F.H., and Trautmann, C.H. 1988. Reliability-based foundation design for transmission line structures: critical evaluation of in situ test methods(Report No. EL-5507[2]). Electric Power Research Institute.

Otake, Y., Ching, J., Saito, T., and Asano, K. 2025. GEOAI benchmark problems BM/AirportSoilProperties/2/2025. Geodata and AI, **2**, 100012. doi:10.1016/j.geoai.2025.100012.

Phoon, K.K., and Kulhawy, F.H. 1999. Characterization of geotechnical variability. Canadian Geotechnical Journal, **36**(4): 612–624. doi:10.1139/t99-038.

Phoon, K.K., Ching, J., and Cao, Z. 2022b. Unpacking data-centric geotechnics. Underground Space, **7**(6): 967–989. doi:10.1016/j.undsp.2022.04.001.

Phoon, K.K., Ching, J., and Shuku, T. 2022a. Challenges in data-driven site characterization. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, **16**(1): 114–126.

Phoon, K.K., Ching, J., and Tao, Y. 2024. Soil and rock parametric uncertainties. Chapter 2 in Uncertainty, Modelling, and Decision Making in Geotechnics. CRC Press, Boca Raton.

Phoon, K.K., Ching, J., and Wang, Y. 2019. Managing risk in geotechnical engineering–from data to digitalization. *In* Proceedings of the 7th International Symposium on Geotechnical Safety and Risk(ISGSR 2019). Taipei, Taiwan. pp. 13–34.

Piotte, M., and Chabbert, M. 2009. The pragmatic theory solution to the Netflix grand prize. Netflix Prize Documentation.

Stewart, G.W. 1993. On the early history of the singular value decomposition. SIAM Review, **35**(4): 551–566. doi:10.1137/1035134.

Tang, K., Liu, R., Su, Z., and Zhang, J. 2014. Structure-constrained low-rank representation. IEEE Transactions on Neural Networks and Learning Systems, **25**(12): 2167–2179. doi:10.1109/TNNLS.2014.2306063.

Wall, M., Rechtsteiner, A., and Rocha, L.M. 2003. Singular value decomposition and principal component analysis. In A practical approach to microarray data analysisVol. Springer US, Boston, MA. pp. 91–109.

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