Development of a database on multivariate soil properties for collapsible loess in Xi'an, China

Xu, Jiabao; Yu, Yongtang; Zheng, Jianguo; Zhang, Lulu; Guan, Zheng; Wang, Yu

DOI
10.1007/s10064-025-04265-4

Publication date
2025

Document Version
Final published version

Published in
Bulletin of Engineering Geology and the Environment

Citation (APA)
Xu, J., Yu, Y., Zheng, J., Zhang, L., Guan, Z., & Wang, Y. (2025). Development of a database on multivariate soil properties for collapsible loess in Xi'an, China. *Bulletin of Engineering Geology and the Environment*, *84*(5), Article 245. https://doi.org/10.1007/s10064-025-04265-4

Bulletin of Engineering Geology and the Environment (2025) 84:245
https://doi.org/10.1007/s10064-025-04265-4

ORIGINAL PAPER

Development of a database on multivariate soil properties for collapsible loess in Xi'an, China

Jiabao Xu[1] · Yongtang Yu[3,4] · Jianguo Zheng[3] · Lulu Zhang[1,5] · Zheng Guan[6] · Yu Wang[2]

Received: 17 October 2023 / Accepted: 13 April 2025 / Published online: 29 April 2025

Abstract

Several global or regional databases for various types of soils have been developed due to their importance in engineering design and analysis. However, a database is not yet available for collapsible loess in which severe geohazards often occur. In this study, a comprehensive loess database with twelve soil parameters is compiled by collecting results of field and laboratory tests on collapsible loess from the city of Xi'an, China. Basic statistics, marginal probability distribution functions (PDFs), and a correlation matrix for loess parameters are estimated from the database. To the best of the authors' knowledge, this is the first collapsible loess database at a municipal level. In addition, existing databases often lack sufficiently complete multivariate measurement data for a proper estimation of statistical correlations among multiple soil properties. In this study, this incomplete multivariate measurement data problem is tackled by Bayesian methods (i.e., Bayesian Gaussian mixture model and Bayesian compressive sampling (BCS) with Karhunen–Loève (KL) expansion, BCS-KL), which are illustrated and validated using the incomplete and complete subsets of the loess database, respectively. Both the Bayesian Gaussian mixture model and BCS-KL are non-parametric, and they offer a flexible way of modeling marginal PDFs and a correlation matrix from incomplete measurements in a realistic manner.

Keywords Collapsible loess database · Marginal distribution · Correlation matrix · Bayesian method

Introduction

It has been increasingly recognized that geotechnical databases compiled from multiple sites can provide valuable supplementary information on soil/rock properties for a typical site with limited local data (Phoon 2018; Ching & Phoon 2020; Bozorgzadeh & Bathurst 2022; Chwała et al. 2023; Phoon & Zhang 2023; Tang & Phoon 2024; Guan et al. 2025). Many rational and scientific methods, such as direct correlation model (Ching et al. 2014; Feng & Vardanega 2019; Asem & Gardoni 2021), quasi-site-specific models (Ching et al. 2021, 2022), and dictionary learning-based methods (Guan et al.

✉ Zheng Guan
zhengguan@tudelft.nl

Jiabao Xu
jiabaoxu2-c@my.cityu.edu.hk

Yongtang Yu
yuyongtang@126.com

Jianguo Zheng
Zhengjg@jk.com.cn

Lulu Zhang
lulu_zhang@sjtu.edu.cn

Yu Wang
wang.yu@ust.hk

[1] State Key Laboratory of Ocean Engineering, Department of Civil Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China

[2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, People's Republic of China

[3] Shaanxi Key Laboratory for the Property and Treatment of Special Soil and Rock, Xi'an, Shaanxi, People's Republic of China

[4] China United Northwest Institute for Engineering Design & Research Co., Ltd., Xi'an, Shaanxi, People's Republic of China

[5] Shanghai Key Laboratory for Digital Maintenance of Buildings and Infrastructure, Shanghai, China

[6] Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

2024; Tian et al. 2025), have been developed for integrating generic soil/rock databases with limited site-specific data to estimate geotechnical properties of interest at a given site. Several global or regional databases including various soil properties have been developed, such as the New Zealand Geotechnical Database (NZGD 2012), CLAY/10/7490 (Ching & Phoon 2014), F-CLAY/7/216 (D'Ignazio et al. 2016), J-CLAY/5/124 (Liu et al. 2016) and SH-CLAY/11/4051 (Zhang et al. 2020). However, to the best of the authors' knowledge, a database on soil properties for collapsible loess is not yet available.

Loess covers an area of approximately 640,000 km^2 in China and large areas worldwide (e.g., Li 2018; Xu et al. 2020). One critical feature of loess is its high collapsibility. Loess with a coefficient of collapsibility (δ_s) equal to or greater than 0.015 is defined as collapsible loess (e.g., PRC MOHURD 2018). Collapsible loess is likely to cause geohazards (e.g., loess landslides and foundation instabilities) because collapsible loess may suddenly collapse under self-weight or surcharge loads after wetting, exhibiting high compressibility and low shear strength (e.g., Jiang et al. 2012; Zhuang et al. 2018). According to Zhou et al. (2002), one-third of landslides in China occurred in loess areas, leading to many casualties and property losses. Xi'an City is a metropolis in northwestern China located on the southern margin of the Loess Plateau. With the development of Xi'an city, many infrastructures and buildings have been constructed on the loess stratification. Developing a Xi'an collapsible loess database of soil parameters has significant implications for improving the characterization of loess parameters and reducing casualties and economic loss caused by loess-related geohazards in Xi'an.

The objective of this study is to develop a comprehensive loess database containing various soil properties and to construct a multivariate probability distribution of these properties based on the database, which characterizes the distinct statistical properties of loess soils. A comprehensive collapsible loess database with twelve soil parameters is compiled using 2266 loess samples with field and laboratory test data from 1764 boreholes in the city of Xi'an, China. The complete multivariate measurement data subset is used to directly compile a loess parameter database. On the other hand, the incomplete multivariate measurement data subset is used as input to the Bayesian Gaussian mixture model and Bayesian compressive sampling (BCS) with Karhunen–Loève (KL) expansion, BCS-KL for non-parametric estimation of marginal probability density functions (PDFs) and a correlation matrix, respectively.

Xi'an collapsible loess database

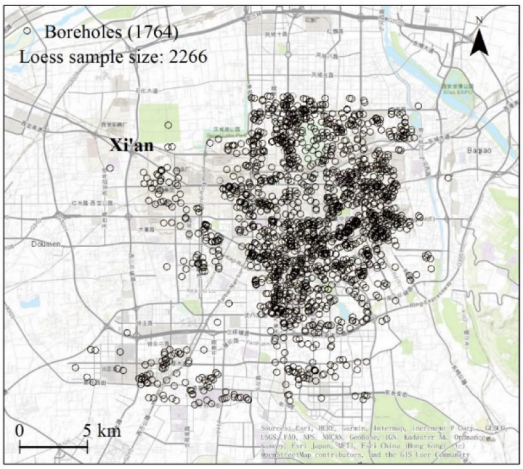
Site Background

Xi'an, a major city located in the northwestern part of China, is situated on the southern margin of the Loess Plateau (Xu et al. 2014). With the development of Xi'an city, many infrastructures and buildings have been constructed on the loess, and plenty of boreholes and geotechnical tests were conducted. In this study, 2266 collapsible loess samples (i.e., with δ_s equal to or greater than 0.015) from the late Pleistocene (Q_3) in Xi'an were collected from 1764 boreholes. As shown in Fig. 1(a), these 1764 boreholes cover an area of approximately 525 km^2. The black circles in Fig. 1(a) represent all the collected boreholes, and the number of boreholes (i.e., 1764) and loess sample size (i.e., 2266) are also labeled in Fig. 1(a). Note that there are sometimes more than one loess sample along different depths in one borehole. Therefore, the number of collapsible loess samples (i.e., 2266) is larger than that of boreholes (i.e., 1764). Xi'an is situated in the Guanzhong Basin and primarily composed of Quaternary loose sedimentary deposits. Based on the available geological and geotechnical data, the soil deposits within the depth range of engineering construction activities are summarized in Table 1.

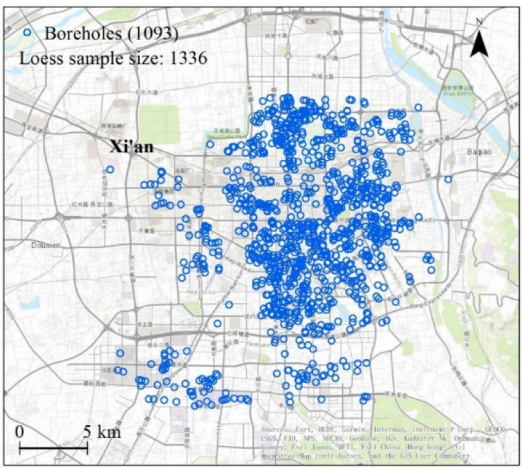
The normalized histogram of δ_s for these 2266 collapsible loess samples is shown in Fig. 2, together with their basic statistics. The normalized histogram of δ_s shows a non-normal distribution, with δ_s values ranging from 0.015 to 0.108. The mean value is 0.041, and a relatively high variability is observed for δ_s with a coefficient of variation (COV) of 0.449.

Twelve loess parameters were measured from these 2266 collapsible loess samples, including seven index parameters: unit weight (γ, kN/m^3), dry unit weight (γ_d, kN/m^3), moisture content (w, %), void ratio (e), liquid limit (LL, %), plastic limit (PL, %), plasticity index (PI, %), and five mechanical parameters: cohesion (c, kPa), internal friction angle (ϕ, °), compressibility coefficient (a_{1-2}, MPa^{-1}), compressive modulus (Es, MPa), and bearing capacity (f_0, kPa), as summarized in Table 2.

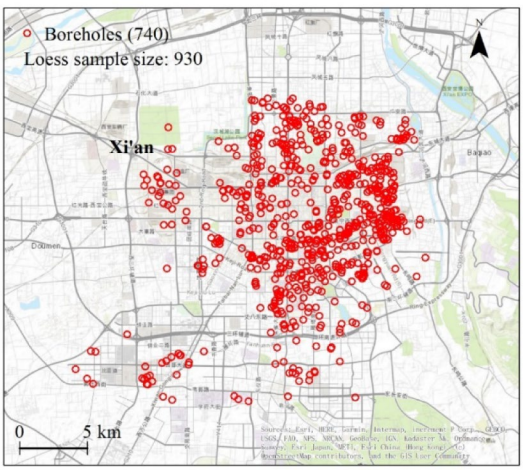
The unit weight γ is and dry unit weight γ_d are determined by specific gravity test. The moisture content w is determined by oven-dry method. The LL and PL are determined by the Atterberg limit test, respectively. The cohesion c and internal friction angle ϕ are usually used in slope stability analysis (Gao et al. 2020; Leng et al. 2021). Most c and ϕ in the database were obtained from direct shear tests, and some were obtained from triaxial tests. This compiled database aims to provide possible ranges of loess parameters, which may be used as prior information for engineering applications in similar geological conditions. Therefore, c and ϕ from different types of tests are combined to provide relatively large ranges of parameters, and the ranges can be used even when the information on the test type is not available. The compressibility coefficient a_{1-2} is a physical quantity that describes the compressibility of loess, which is the slope of the secant at a given range of e-p curves in a compression test (Li et al. 2018). A large a_{1-2} indicates high



(a) locations for boreholes in Xi'an



(b) complete data subset



(c) incomplete data subset

Fig. 1 Site overview and locations for boreholes (**a**) locations for boreholes in Xi'an (**b**) complete data subset (**c**) incomplete data subset

compressibility. The compressive modulus *Es* is another parameter that describes the compressibility of loess, which is the ratio of vertical stress to the strain of loess under the condition of lateral confinement (Li et al. 2018). A small *Es* indicates high compressibility. a_{1-2} and *Es* are usually obtained from oedometer tests, which can be utilized to calculate foundation settlement (Yang & Bai 2015). f_0 is the bearing capacity of loess, which is usually for foundation design (Feng et al. 2015). In this database, f_0 is obtained using a commonly used empirical method for loess in China that relates f_0 with the loess chronology, moisture content, void ratio, and liquid limit (GBJ25-90 1991). This study collects measurement data for these twelve loess parameters and compiles a collapsible loess database for Xi'an.

The data on these twelve loess parameters measured from one loess sample is referred to as one set of multivariate measurement data in this study. Therefore, 2266 sets of multivariate measurement data for these twelve loess parameters are collected, which occupy 2266 rows and 12 columns in Table 2. The first column in Table 2 is the loess sample number (No.), ranging from No. 1 to No. 2266. Measurement data in each row represents one set of multivariate measurement data from one loess sample. As a result of sampling and test limitations, some loess samples were not tested for all 12 loess parameters (see empty cells in Table 2). To develop a complete correlation matrix for all twelve loess parameters, a database without any missing value (i.e., no empty cell in Table 2) is needed. Hence, the 2266 loess sample sets were divided into two data subsets: one with complete multivariate measurements is called the complete data subset (i.e., rows without empty cells in Table 2), and the other with incomplete multivariate measurements is called the incomplete data subset (i.e., rows with some empty cells in Table 2). The incomplete data subset is used to illustrate the effectiveness of these two Bayesian methods (i.e., Bayesian Gaussian mixture model and BCS-KL) in estimating the statistical information of a database based on incomplete multivariate measurement data.

As shown in Fig. 1(b), the complete data subset contains 1336 sets of complete multivariate data (i.e., from 1336 loess samples) with all twelve loess parameters measured from 1093 boreholes. The blue circles represent boreholes with complete multivariate measurement data, and the numbers for boreholes and loess sample size are labeled in Fig. 1(b). The incomplete data subset contains 930 sets of incomplete multivariate measurement data (i.e., from 930 loess samples) from 740 boreholes, as shown in Fig. 1(c). In Fig. 1(c), red circles indicate boreholes with incomplete multivariate measurement data, and the total number of collected boreholes is shown in parentheses. The incomplete data subset has 1827 empty cells. The percentage of incompleteness for the incomplete data subset can be indexed by a ratio of empty cell number over all cell number, i.e., 1827/

Table 1 Geological information for Xi'an

Epoch	Material Produced	Thickness (m)	Soil type	Description
Holocene	Alluvial	2–30	Sand, Gravel	Distributed in river channels and floodplains; lithology varies from fine sand to gravelly sand
	Alluvial	10–20	Loess, Gravel	Occurs in mountain valleys and terraces; lithology is primarily silt with embedded gravel
	Lacustrine	7–30	Clayey Silt, Sand	Found in lacustrine basins; upper layers contain clayey silt, lower layers consist of sand
	Alluvial	10–30	Clayey Silt, Gravel	Deposited in river terraces; predominantly clayey silt in upper layers, with sand and gravel in lower sections
	Alluvial & Floodplain	2–15	Clayey Silt, Sand, Silt	Distributed in deltaic environments; fine-grained silt and sand dominate
Pleistocene	Aeolian	7–25	Loess, Paleosol	Distributed in terraces and upper slopes; composed mainly of loess deposits with paleosol intercalations
	Alluvial & Floodplain	10–30	Clayey Silt, Sand	Deposits formed along river terraces; fine-grained material dominates
	Alluvial	3–5	Sand, Clayey Silt	Formed in alluvial plains; layers of fine sand and silt with interbedded clay
	Glacial	10–30	Sandstone, Breccia, Mudstone	Glacial deposits formed under high-energy environments; characterized by poorly sorted sediments

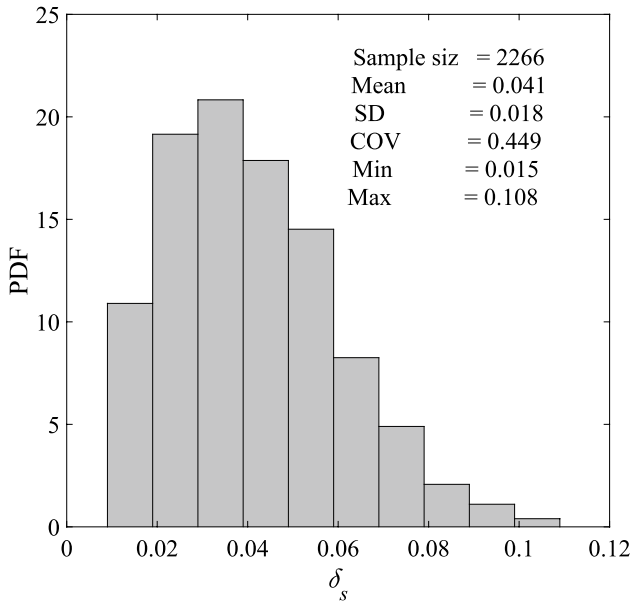


Fig. 2 Normalized histogram for the coefficient of collapsibility

$(930 \times 12) \approx 16.4\%$. Therefore, the percentage of completeness for the incomplete data subset is about 83.6%.

Note that sometimes one borehole contains more than one loess sample, and different samples from the same borehole may belong to complete or incomplete data subsets, respectively. Therefore, the sum (i.e., 1833) of the number of boreholes in the complete data subset (i.e., 1093) and incomplete data subset (i.e., 740) is larger than the total number of all boreholes (i.e., 1764). Statistical characterization of the complete and incomplete data subsets will be described in Subsections "Site Background" and "Statistical description of the complete data subset", respectively.

Statistical description of the complete data subset

Figure 1(b) shows that the boreholes in the complete data subset cover an area similar to the area covered by all boreholes shown in Fig. 1(a). Table 3 summarizes the basic statistics of the complete data subset, and the second column shows the number ($n = 1336$) of measurement data for each loess parameter (i.e., the number of loess samples). The statistics include mean value, standard deviation (SD), COV, minimum value (min), and maximum value (max).

Relatively high variability is observed for c, a_{1-2}, and Es with COV of 0.34, 0.57, and 0.45, respectively, while the rest of the loess parameters show relatively low variability with COV less than 0.15 (Liu et al. 2016). The void ratio e ranges from 0.78 to 1.32, which is relatively large due to the highly porous characteristics of loess (Jiang et al. 2012; Li 2018). Based on the multivariate measurement data in the complete data subset, marginal probability distributions and a correlation matrix for all twelve loess parameters will be developed in Section "Statistical model for the incomplete data subset".

Statistical description of the incomplete data subset

The borehole locations for the incomplete data subset are shown in Fig. 1(c), covering an area similar to the area of all boreholes shown in Fig. 1(a) and (b). Table 4 shows the basic statistics of the incomplete data subset, and the second column displays the number (n) of measurement

data for each loess parameter (i.e., the number of loess samples). The numbers of measurement data points for c and ϕ are quite limited (i.e., 69 and 70, respectively). On the other hand, the numbers of measurement data for the other loess parameters exceed 800, which is reasonably large. The statistics include mean value, standard deviation (SD), coefficient of variation (COV), minimum value (min), and maximum value (max).

The mean values, SD, and COV in Table 4 for the incomplete data subset are similar to those in Table 3 for the complete data subset because these measurements are from loess samples in the same region and geological condition, and the number of measurement data for these two data subsets is similar, except for c and ϕ.

The basic statistics for c and ϕ in Tables 3 and 4 are different because the numbers of measurement data for c and

Table 2 Multivariate measurement data for loess samples in Xi'an

Sample No	γ	γ_d	w	e	LL	PL	PI	c	ϕ	a_{1-2}	Es	f_0
No. 1	13.90	11.80	17.30	1.29	29.90	17.70	12.30	37.00	19.70	0.79	2.85	150.00
No. 2	13.80	11.50	20.00	1.32	32.50	19.70	12.80	-	-	0.79	3.90	120.00
No. 3	14.10	11.90	18.80	1.23	–	–	–	43.00	23.90	0.55	5.20	155.00
No. 4	14.17	12.06	18.05	1.21	29.58	17.95	11.63	18.88	24.74	0.83	3.64	160.00
No. 5	15.90	–	–	1.08	–	–	–	–	–	0.27	8.00	140.00
No. 2266	14.24	12.60	13.07	1.16	29.17	17.90	11.27	19.10	26.00	0.34	10.52	-

γ is unit weight (kN/m^3); γ_d is dry unit weight (kN/m^3); w is moisture content (%); e is void ratio; LL is liquid Limit (%); PL is plastic limit (%); PI is plasticity index (%); c is cohesion (kPa); ϕ is internal friction angle (°); a_{1-2} is compressibility coefficients (MPa^{-1}); Es is compressive modulus (MPa); f_0 is bearing capacity (kPa)

Table 3 Basic statistics of 12 loess parameters from the complete data subset

Parameter	n	Mean	SD	COV	Min	Max
γ (kN/m^3)	1336	16.02	0.91	0.06	13.90	19.70
γ_d (kN/m^3)	1336	13.14	0.64	0.05	11.50	15.30
w (%)	1336	21.81	2.61	0.12	12.30	33.00
e	1336	1.06	0.10	0.09	0.78	1.32
LL (%)	1336	30.64	1.28	0.04	27.50	34.90
PL (%)	1336	18.45	0.64	0.03	16.30	20.70
PI (%)	1336	12.19	0.74	0.06	10.40	14.50
c (kPa)	1336	32.12	10.95	0.34	11.70	80.00
ϕ (°)	1336	22.69	3.10	0.14	12.60	33.90
a_{1-2} (MPa^{-1})	1336	0.43	0.24	0.57	0.09	1.54
Es (MPa)	1336	7.47	3.34	0.45	1.50	23.10
f_0 (kPa)	1336	138.10	18.08	0.13	70.00	220.00

Table 4 Basic statistics of 12 loess parameters from the incomplete data subset

Parameter	n	Mean	SD	COV	Min	Max
γ (kN/m^3)	930	16.09	0.93	0.06	13.60	18.80
γ_d (kN/m^3)	923	13.21	0.66	0.05	11.40	16.00
w (%)	923	21.78	3.00	0.14	12.00	33.80
e	930	1.05	0.10	0.10	0.75	1.40
LL (%)	930	31.11	1.43	0.05	26.80	35.70
PL (%)	927	18.70	0.77	0.04	16.60	22.40
PI (%)	927	12.42	0.77	0.06	10.20	15.00
c (kPa)	69	28.48	8.90	0.31	14.80	56.00
ϕ (°)	70	23.57	2.30	0.10	16.30	27.00
a_{1-2} (MPa^{-1})	930	0.40	0.25	0.61	0.10	1.62
Es (MPa)	926	8.06	3.82	0.47	1.60	24.40
f_0 (kPa)	844	139.84	18.38	0.13	70.00	190.00

ϕ in the incomplete data subset are limited (i.e., 69 and 70, respectively), which may not be sufficient to estimate their marginal probability distributions properly. In addition, no complete multivariate measurement data for all twelve loess parameters were available to estimate a correlation matrix. In engineering practice, incomplete multivariate measurement data are frequently encountered because some soil parameters are not always tested in a soil sample and some tests are destructive (Phoon 2018; Ching & Phoon 2020). Therefore, developing a multivariate statistical model (e.g., a complete correlation matrix or correlation model among soil parameters) is difficult using the incomplete data subset with missing values (Phoon et al. 2022).

Statistical model for the complete data subset

Marginal probability distribution

The marginal probability distribution for each of the twelve loess parameters from the complete data subset is plotted in Fig. 3 using red dashed lines. Figure 3 shows that the normalized histograms of PL and γ_d approximately follow a Gaussian distribution. However, the normalized histograms of LL, PI, c, and e are right-skewed to their mean values, and those of ϕ, w, and f_0 are left-skewed to their mean values, indicating a non-Gaussian pattern. Furthermore, the probability value at around $\gamma = 17$ in the γ histogram of Fig. 3(a) is relatively high. The normalized histograms of γ, a_{1-2}, and Es, do not follow any of the commonly used theoretical PDF types, e.g., the normal or log-normal distribution. Therefore, the traditional methods based on normal or log-normal distribution assumptions might not be applicable to fit or estimate the marginal probability distributions of loess parameters in Xi'an, especially for γ, a_{1-2}, and Es. Therefore, the empirical marginal distributions are directly estimated from the normalized histograms in Fig. 3 without fitting to any theoretical PDF type.

Correlation matrix $\mathbf{C_0}$

A correlation matrix is developed to represent correlation among all loess parameters. The correlation matrix $\mathbf{C}_0$ for these twelve collapsible loess parameters can be calculated directly from the complete multivariate measurement data subset using the Pearson correlation coefficient $r_{i,j}$ (Kendall & Stuart 1963):

$$r_{i,j} = \frac{\sum_{m=1}^{n}(x_{i,m} - \overline{x_i})(x_{j,m})}{\sqrt{\sum_{m=1}^{n}(x_{i,m} - \overline{x_i})^2}\sqrt{\sum_{m=1}^{n}(x_{j,m} - \overline{x_j})^2}} \quad (1)$$

where $x_{i,m}$ and $x_{j,m}$ represent the m-th measurement data of loess parameters i and j in the complete data subset, respectively; $\bar{x}_i$ and $\bar{x}_j$ represent the mean values of all measurement data for loess parameters i and j, respectively; $n = 1336$ is the number of measurement data for each loess parameter in this study. In this study, the correlation matrix $\mathbf{C}_0$, as shown in

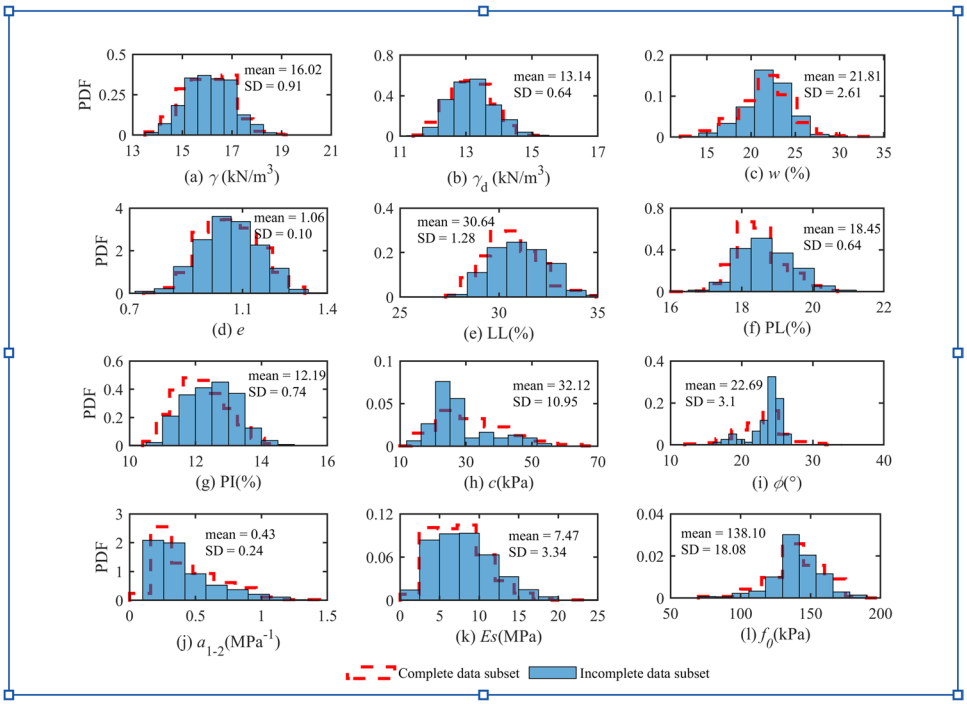


Fig. 3 Comparison of normalized histograms for 12 collapsible loess parameters (γ, γ_d, w, e, LL, PL, PI, c, ϕ, a_{1-2}, E_s, f_0, labeled from (**a**) to (**l**)) between the complete and incomplete data subsets

Table 5, can be obtained directly using the "corr" function in MATLAB with the complete data subset as input.

Table 5 shows that γ shows a positive correlation to γ_d with a correlation coefficient of 0.92. In addition, e shows a strong negative correlation to γ and γ_d with a correlation coefficient of − 0.91 and − 0.97, respectively, because soil with a high void ratio usually shows low unit weight and dry weight. In general, loess with a high void ratio exhibits high compressibility (Kurnaz et al. 2016). Therefore, e is positively and negatively correlated to a_{1-2} and Es with a correlation coefficient of 0.70 and − 0.58, respectively. In addition, γ_d shows a negative and positive correlation to a_{1-2} and Es with a correlation coefficient of − 0.68 and 0.55, respectively, and γ is also negatively and positively correlated to a_{1-2} and Es with a correlation coefficient of − 0.53 and 0.40, respectively, because of the correlations among e, γ and γ_d. a_{1-2} is negatively correlated to Es with a correlation coefficient of − 0.83 because a large a_{1-2} indicates high compressibility, whereas a large Es indicates low compressibility. Loess with lower compressibility often has a high bearing capacity (e.g., Li et al. 2018). Therefore, f_0 is negatively and positively correlated with a_{1-2} and Es with a correlation coefficient of − 0.50 and 0.55, respectively. To illustrate the correlation among some strongly correlated loess parameters, scatter plots are shown by black circles in Fig. 4. Figure 4(a)-(f) show the correlation between parameters γ and e, γ and a_{1-2}, γ_d and a_{1-2}, γ_d and Es, e and a_{1-2}, a_{1-2} and Es, respectively, and the corresponding correlation coefficients are also labeled in each subplot.

Statistical model for the incomplete data subset

In this section, a Bayesian Gaussian mixture model is first used to estimate the marginal probability distribution using the incomplete measurement data in Subsection "Estimation of marginal probability distribution using Bayesian Gaussian mixture model". In Subsection "Estimation of correlation matrix C1 using BCS-KL", a BCS-KL random field generator is used to estimate a correlation matrix using the incomplete data subset. Then, a multivariate joint probability distribution is developed based on the estimated marginal probability distributions and the correlation matrix. Finally, the multivariate joint probability distribution is used to generate multivariate random samples using a KL random field generator to further evaluate and illustrate the results in Subsection "Simulation of multivariate joint distribution for loess parameters".

Estimation of marginal probability distribution using Bayesian Gaussian mixture model

The marginal probability distribution for each of the twelve loess parameters in the incomplete data subset is plotted using blue bins in Fig. 3. Figure 3 shows that most blue bins are consistent with these red dashed lines, except for c and ϕ. This consistency is reasonable because the measurement data in the two data subsets are from the same region with the same geological conditions, and the numbers of measurement data for these two data subsets are similar, except for c and ϕ. The numbers of measurement data for c and ϕ are 69 and 70, respectively, which are insufficient to estimate marginal probability distributions properly. Therefore, the marginal probability distributions of c and ϕ for the incomplete and complete data subsets exhibit significant differences. In addition, the marginal probability distributions for c and ϕ in the complete data subset are right-skewed and left-skewed, respectively. This is further evidence that the frequently used normal or log-normal distributions are not suitable to fit or represent the marginal probability distributions of c and ϕ. In this subsection, the Bayesian Gaussian mixture model (Guan & Wang 2021) is used to estimate the

Table 5 Correlation matrix $\mathbf{C}_0$ for the complete data subset

	γ	γ_d	w	e	LL	PL	PI	c	ϕ	a_{1-2}	Es	f_0
γ	1.00	0.92	0.53	–0.91	0.17	0.15	0.17	0.21	–0.08	–0.53	0.40	0.16
γ_d		1.00	0.19	–0.97	0.06	0.06	0.06	0.24	0.02	–0.68	0.55	0.28
w			1.00	–0.22	0.32	0.28	0.31	0.01	–0.28	0.18	–0.22	–0.22
e				1.00	–0.08	–0.06	–0.09	–0.31	0.05	0.70	–0.58	–0.34
LL					1.00	0.91	0.94	0.32	–0.05	–0.08	0.16	0.27
PL						1.00	0.73	0.22	0.03	–0.07	0.14	0.15
PI		Symmetry					1.00	0.36	–0.11	–0.07	0.16	0.33
c								1.00	–0.16	–0.34	0.45	0.50
ϕ									1.00	–0.11	0.12	–0.06
a_{1-2}										1.00	− 0.83	− 0.50
Es											1.00	0.55
f_0												1.00

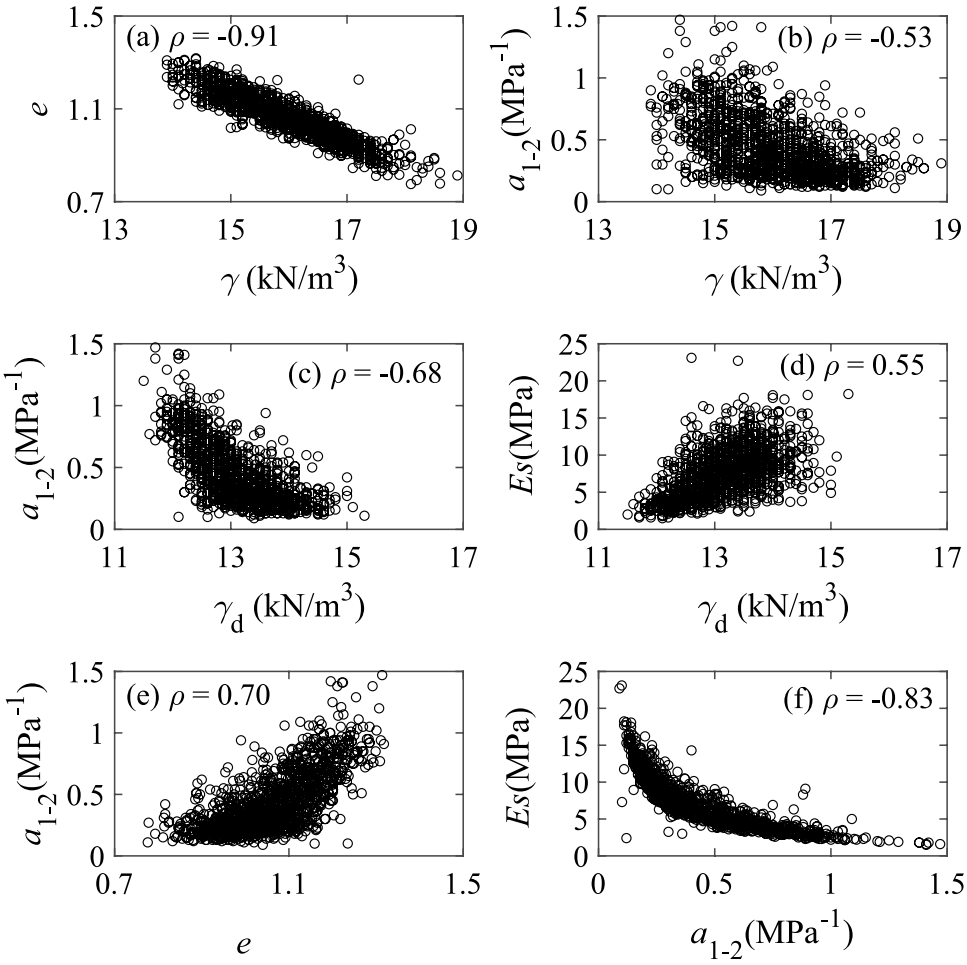


Fig. 4 Strong correlation between collapsible loess parameters from the complete data subset: (**a**) γ and e; (**b**) γ and a_{1-2}; (**c**) γ_d and a_{1-2}; (**d**) γ_d and Es; (**e**) e and a_{1-2}; (**f**) a_{1-2} and Es

marginal probability distribution of c and ϕ using limited measurement data.

The Bayesian Gaussian mixture model can estimate a site-specific marginal probability distribution for a random variable from limited site-specific measurement data (Wang & Cao 2013; Deng et al. 2022). In this method, the PDF $p(x)$ of a random variable x (e.g., a soil property) is represented as follows (e.g., McLachlan & Peel 2000; Rasmussen 1999; Wang et al. 2015):

$$p(x)=\int_{\mu,\sigma} p(x|\mu,\sigma)p(\mu,\sigma|\text{data})d\mu d\sigma \tag{2}$$

where $p(x|\mu, \sigma)$ represents a component of the mixture model (e.g., a Gaussian PDF with mean μ and standard deviation σ); $p(\mu, \sigma|\text{data})$ represents the corresponding weight which can be formulated as the posterior PDF of μ and σ under a Bayesian framework. $p(\mu, \sigma|\text{data})$ can be estimated by combining the measurement data with prior information from engineer judgment/experience under a Bayesian framework as follows (, Wang & Cao 2013; Ibsen 2019):

$$p(\mu,\sigma|\text{data}) = Kp(\text{data}|\mu,\sigma)p(\mu,\sigma) \tag{3}$$

where K represents a normalizing constant; $p(\text{data}|\mu, \sigma)$ and $p(\mu, \sigma)$ are the likelihood function and prior distribution, respectively. When no informative prior knowledge is available, the prior information can be taken to follow a uniform distribution (Wang & Cao 2013; Guan & Wang 2021). Only the possible range of μ and σ are needed to define a uniform prior distribution (e.g., reasonable ranges of a loess parameter concerned) (Wang et al. 2016). Combining Eqs. (2) and (3), the PDF of a variable can be expressed as (Guan & Wang 2021):

$$p(x|\text{data}) = K\int_{\mu,\sigma} p(x|\mu,\sigma)p(\text{data}|\mu,\sigma)p(\mu,\sigma)d\mu d\sigma \tag{4}$$

Equation (4) represents the marginal probability density function $p(x)$ of a random variable x as an integral over the product of the conditional probability density function $p(x|\mu, \sigma)$, which describes the distribution of x given parameters μ and, the likelihood $p(\text{data}|\mu, \sigma)$, and the prior distribution

$p(\mu, \sigma)$. The term $p(x|\mu, \sigma)$ represents a Gaussian probability density function, and the equation is effectively a weighted summation of these Gaussian PDFs, where the weights are determined by the posterior distribution $p(\mu, \sigma|\text{data})$ and the prior $p(\mu, \sigma)$. Equation (4) is usually complicated and hard to obtain an analytical resolution. Therefore, Markov chain Monte Carlo (MCMC) simulation is often used to generate many x samples sequentially, and the PDF of x is obtained directly from the generated samples (Guan & Wang 2021; Xu et al. 2021a).In this study, uniform distributions with respective minimum and maximum values of μ and σ for c and ϕ are adopted (Li 2018), as shown in Table 6. By combining prior information and limited measurement data, 15,000 MCMC samples for c and ϕ are generated from MCMC simulation, respectively. The initial 5,000 samples are discarded to consider the burn-in period, leaving 10,000 effective MCMC samples for c and ϕ, respectively. Figure 5 represents the estimated PDF of c and ϕ from MCMC samples using yellow bins. For comparison, Fig. 5 also includes the PDFs of c and ϕ from the complete data subset using red dashed lines. Figure 5 shows that the PDFs estimated from the Bayesian Gaussian mixture model are similar to those obtained from the complete data subset for both c and ϕ. Therefore, the Bayesian Gaussian mixture model provides a reasonable marginal PDF estimation for loess parameters when the measurement data are limited. Unlike a conventional parametric model (e.g., multivariate normal distribution model), the Bayesian Gaussian mixture model is non-parametric and does not pre-specify the PDF function types.

Table 6 Prior information for c and ϕ

Parameter	μ_{min}	μ_{max}	σ_{min}	σ_{max}
c (kPa)	0.01	100	0.01	20
ϕ (°)	0.01	100	0.01	20

Estimation of Correlation Matrix $\mathbf{C}_1$ Using BCS-KL

The correlation matrix $\mathbf{C}_1$ for the incomplete data subset cannot be estimated directly from incomplete multivariate measurement data using the Pearson correlation coefficients. Guan & Wang (2021) developed a novel random field-based approach to estimate a correlation matrix from incomplete multivariate measurement data. The method combines Bayesian compressive sampling (BCS) with Karhunen–Loève (KL) expansion to estimate a correlation matrix.

Using BCS-KL method for estimating a correlation matrix is based on an idea that the cross-correlation among different variables can be modeled similarly to the auto-correlation of a non-stationary random process when the variables are arranged in a specific sequence (Guan & Wang 2021). In other words, when various soil properties are listed in a particular sequence, which is treated as a virtual spatial coordinate of a latent dimension, the cross-correlation among these properties is similar to the auto-correlation along this virtual coordinate.

Then, the BCS-KL generator can be used to generate many sets of complete random field samples from incomplete multivariate measurement data in each row because the BCS-KL generator can directly simulate non-Gaussian and non-stationary random field samples from sparse measurements (Guan & Wang 2021; Wang et al. 2022). After incomplete multivariate measurement data in all rows are used by BCS-KL, a correlation matrix for all variables is estimated directly from these simulated complete random field samples using the Pearson correlation coefficients. In this study, BCS-KL is used to simulate 1,000 sets of complete random field samples from each set (or each row in Table 2) of incomplete multivariate measurements. The implementation procedure of BCS-KL is briefly described in the Appendix. Because the incomplete data subset contains 930 rows, as described

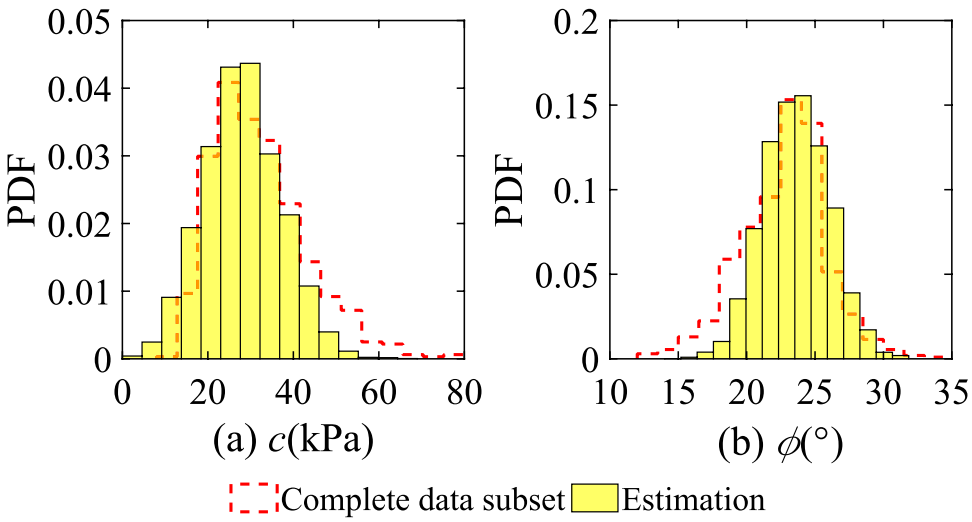


Fig. 5 Comparison of normalized histograms between the Bayesian Gaussian mixture model results using incomplete data subset and direct estimation from complete data subset: (**a**) cohesion c; (**b**) internal friction angle ϕ

in Subsection "Statistical description of the complete data subset", a total of 930 × 1000 = 930,000 sets of complete random field samples are generated subsequently.

Using the generated random field samples, the correlation matrix $\mathbf{C}_1$ for all twelve loess parameters can be estimated directly using Eq. (1), as shown in Table 7. The absolute differences between the estimated correlation matrix $\mathbf{C}_1$ in Table 7 and C0 from the complete data subset in Table 5 are also calculated and are shown in Table 8. It is found that each element in Table 8 ranges from 0.01 to 0.31. The absolute differences are relatively small, especially for the strong correlation coefficients. For example, the absolute differences of correlation coefficients between LL and PL, LL and PI, and PI and PL in $\mathbf{C}_{01}$ are 0.01, 0.04, and 0.01, respectively, which are all close to zero. In addition, absolute differences for the correlation coefficients between γ and γ_d, γ and e, γ_d and e in $\mathbf{C}_{01}$ are 0.04, 0.05 and 0.05, which are also close to zero. Note that most absolute difference values are less than 0.20. Seven (out of 66) absolute differences are greater than 0.2, including the correlation coefficients between e and LL, and e and PL, c and LL, c and PL, c and PI, ϕ and LL, ϕ and PL in $\mathbf{C}_{01}$. These relatively large differences are primarily due to the relatively small values of the seven corresponding correlation coefficients in $\mathbf{C}_0$ (i.e., − 0.08, − 0.06, 0.32, 0.22, 0.36, − 0.05, and 0.03). The weak correlation among loess parameters is similar to a small auto-correlation of a non-stationary random process, which is very challenging to simulate accurately.

To further compare the correlation matrices $\mathbf{C}_1$ and $\mathbf{C}_0$, eigen-decomposition is performed on each correlation matrix to obtain their eigenvalues and eigenvectors. Because both $\mathbf{C}_1$ and $\mathbf{C}_0$ have a dimension of 12 × 12, the number of both eigenvectors and eigenvalues is 12 for $\mathbf{C}_1$ and $\mathbf{C}_0$, as shown in Fig. 6. Figure 6(a) plots the eigenvalues in a descending order of the eigenvalues. All 12 eigenvalues and eigenvectors are similar for $\mathbf{C}_1$ and $\mathbf{C}_0$. The summation of all 12 eigenvalues for both $\mathbf{C}_1$ and $\mathbf{C}_0$ is 12.0. Figure 6(b) plots the eigenvectors in a descending order of the eigenvalues. The 1 st, 2nd, 3rd, 6 th, 7 th, 8 th, 9 th and 12 th eigenvectors are almost identical for C_1 and C_0, and the other four eigenvectors also exhibit similar patterns. The summation of corresponding eigenvalues for these eight identical eigenvectors for C_1 is 10.40, which occupies a large proportion

Table 7 Correlation matrix $\mathbf{C}_1$ estimated from the incomplete data subset

	γ	γ_d	w	e	LL	PL	PI	c	ϕ	a_{1-2}	Es	f_0
γ	1.00	0.87	0.49	–0.86	0.04	–0.02	0.08	0.25	–0.23	–0.42	0.32	0.21
γ_d		1.00	0.12	–0.92	–0.04	–0.09	–0.03	0.34	–0.09	–0.58	0.50	0.37
w			1.00	–0.13	0.27	0.21	0.31	–0.09	–0.45	0.24	–0.31	–0.28
e				1.00	0.13	0.19	0.07	–0.33	0.07	0.62	–0.50	–0.41
LL					1.00	0.91	0.90	0.03	–0.29	–0.02	0.11	0.08
PL						1.00	0.72	–0.02	–0.24	0.01	0.09	–0.02
PI		Symmetry					1.00	0.05	–0.30	0.00	0.09	0.14
c								1.00	0.04	–0.42	0.34	0.40
ϕ									1.00	–0.11	0.10	0.04
a_{1-2}										1.00	–0.79	–0.57
Es											1.00	0.57
f_0												1.00

Table 8 Absolute difference $\mathbf{C}_{01}$ between correlation matrix $\mathbf{C}_1$ and $\mathbf{C}_0$

	γ	γ_d	w	e	LL	PL	PI	c	ϕ	a_{1-2}	Es	f_0
γ	1.00	0.04	0.03	0.05	0.13	0.18	0.08	0.05	0.15	0.10	0.08	0.05
γ_d		1.00	0.07	0.05	0.11	0.15	0.08	0.10	0.12	0.10	0.05	0.09
w			1.00	0.09	0.04	0.07	0.00	0.10	0.17	0.06	0.09	0.06
e				1.00	0.21	0.25	0.16	0.02	0.02	0.08	0.07	0.07
LL					1.00	0.01	0.04	0.29	0.24	0.06	0.05	0.19
PL						1.00	0.01	0.24	0.27	0.08	0.05	0.17
PI							1.00	0.31	0.19	0.07	0.07	0.19
c								1.00	0.20	0.08	0.12	0.10
ϕ									1.00	0.00	0.02	0.10
a_{1-2}										1.00	0.04	0.07
Es											1.00	0.02
f_0												1.00

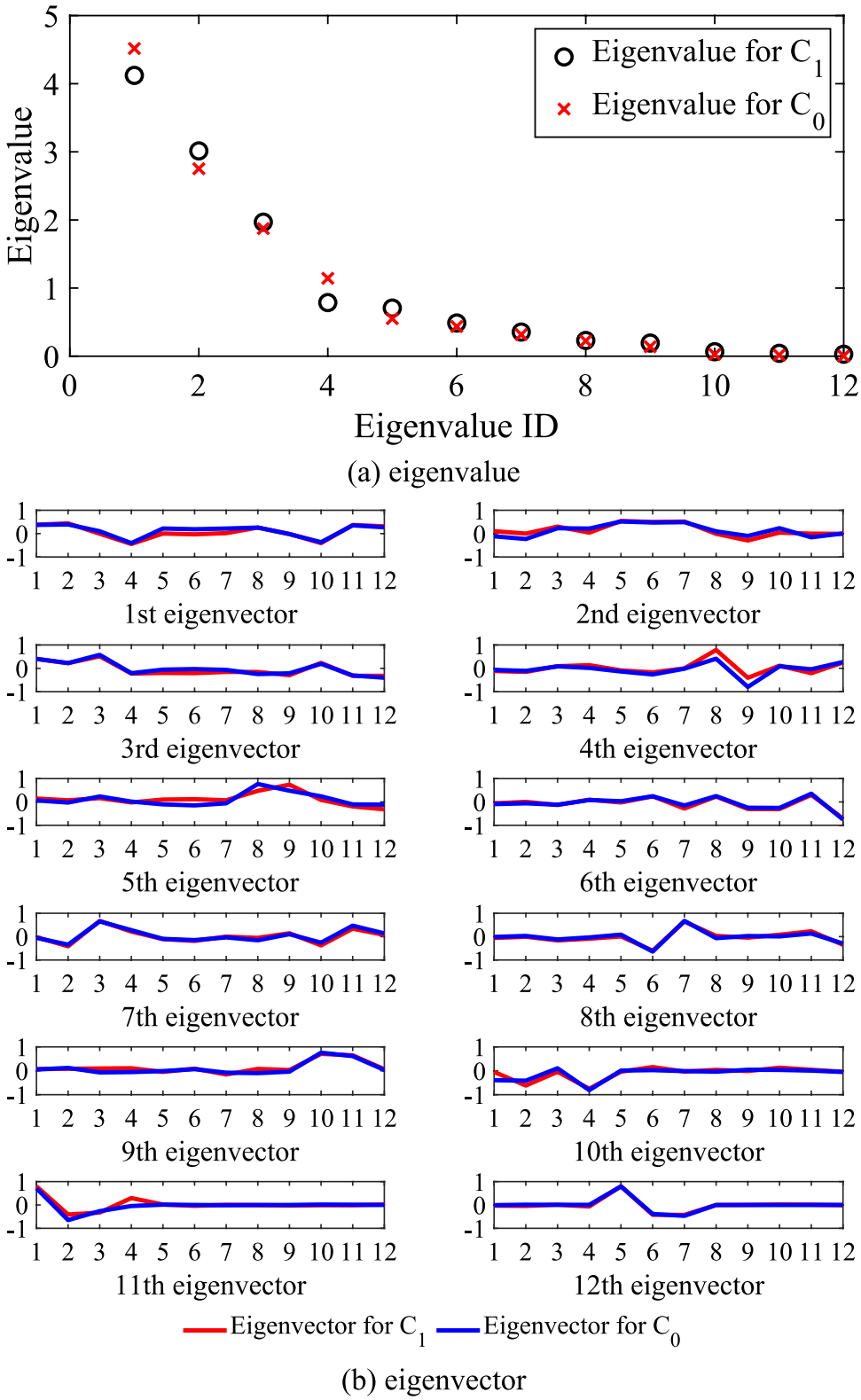


Fig. 6 Comparison between $\mathbf{C}_1$ and $\mathbf{C}_0$ for: (**a**) eigenvalues; (**b**) eigenvectors

for all the eigenvalues for C_1 with a ratio of = 10.40/12.0 ≈ 86.67%. These eight identical eigenvectors can approximately represent a majority of all the 12 eigenvectors. Therefore, the correlation matrix $\mathbf{C}_1$ obtained from BCS-KL with the incomplete data subset properly represents the correlation matrix $\mathbf{C}_0$ obtained from the complete data subset, based on specific correlation coefficient values and eigendecomposition results.

Simulation of multivariate joint distribution for loess parameters

After determining the marginal probability distributions and correlation matrix $\mathbf{C}_1$ from the incomplete data subset, a multivariate joint probability distribution for loess parameters is specified. To further illustrate and validate the multivariate joint probability distribution, multivariate random samples are generated simultaneously using KL expansion and the estimated multivariate joint probability distribution (e.g., Phoon et al. 2002, 2005). In this study, 10, 000 sets of simulated data for these twelve loess parameters are generated, resulting in a simulated data matrix with a dimension of $10,000 \times 12$.

Scatter plots among some strongly correlated loess parameters are represented in Fig. 7 using blue triangles from the 10,000 sets of simulated data. Figure 7(a)-(f) show the correlations between γ and e, γ and a_{1-2}, γ_d and a_{1-2}, γ_d and Es, e and a_{1-2}, a_{1-2} and Es, respectively. To evaluate the accuracy of the simulated data, the scatter plots from the complete data subset are also plotted in Fig. 7, as shown by the black circles. Figure 7 shows that these black circles are bounded by these blue triangles, which means that the scatter plots from the simulated data are consistent with those from the complete data subset. This suggests that the multivariate joint probability distribution estimated from the Bayesian methods using incomplete multivariate measurement data provides consistent results with those obtained from the complete data subset.

The results in Subsections "Estimation of marginal probability distribution using Bayesian Gaussian mixture model"-"Simulation of multivariate joint distribution for loess parameters" show that the Bayesian Gaussian mixture model and the BCS-KL method can effectively estimate marginal distributions and a correlation matrix from incomplete multivariate measurements, a challenge frequently encountered in engineering practice. In addition, these two Bayesian methods are non-parametric. The Bayesian Gaussian mixture model does not pre-specify the PDF function types, and hence, it is applicable and appealing for geotechnical parameters which might follow unknown, or even empirical, types of the marginal distribution. The Bayesian Gaussian mixture model and BCS-KL method offer a flexible method

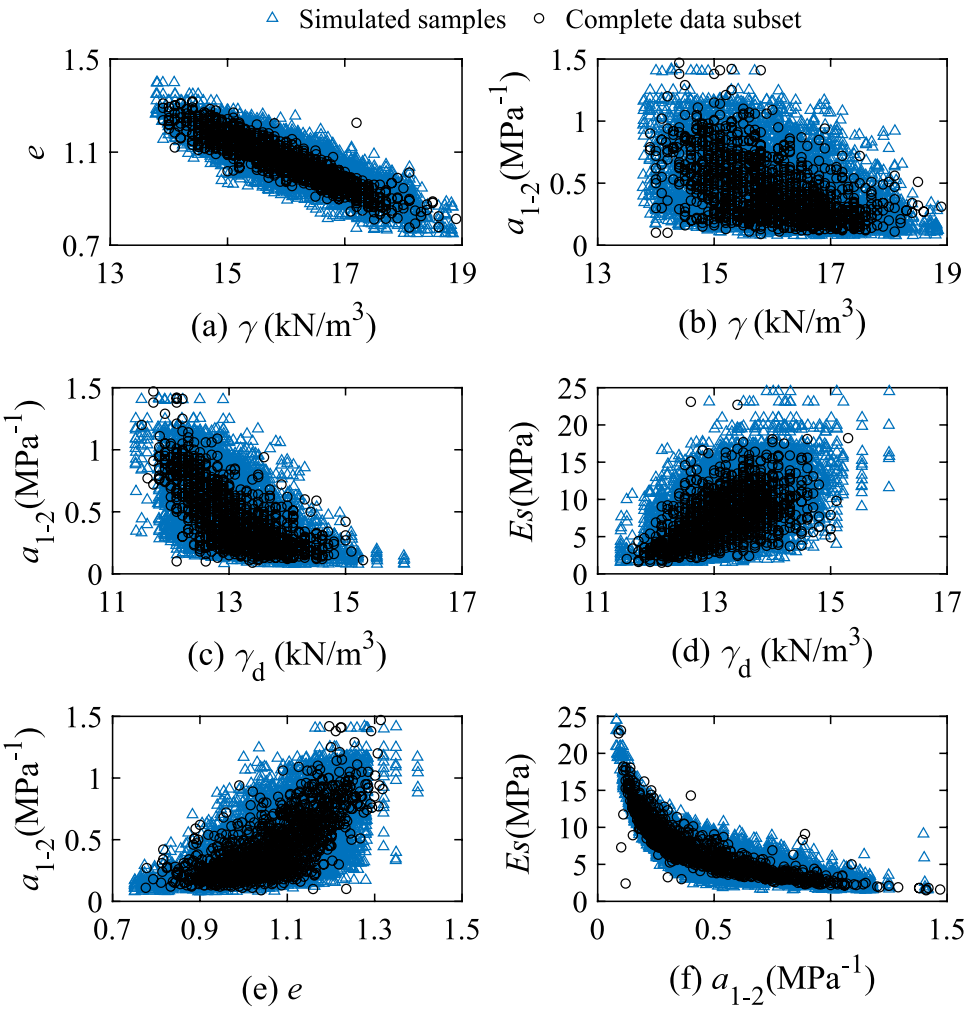


Fig. 7 Comparison of strongly correlated loess parameters between simulated samples and complete data subset: (**a**) γ and e; (**b**) γ and a_{1-2}; (**c**) γ_d and a_{1-2}; (**d**) γ_d and Es; (**e**) e and a_{1-2}; (**f**) a_{1-2} and Es

for modeling marginal PDFs and the correlation matrix from incomplete measurements in a realistic manner.

Conclusions

A comprehensive collapsible loess database was compiled in this study by collecting 2266 loess samples from 1764 boreholes with field and laboratory test data in the City of Xi'an, which is the first collapsible loess database at a municipal level.

- The developed database provides valuable prior information that can be used, together with limited site-specific data often encountered in engineering practice, to estimate the soil properties of interest at a given site.
- The Bayesian Gaussian mixture model and BCS-KL method have been demonstrated to be effective in estimating marginal probability distributions and correlation matrices from incomplete multivariate measurement data. The results obtained from these methods were consistent with those from the complete data subset, validating their reliability.
- These Bayesian methods offer a flexible, non-parametric approach to modeling marginal probability density functions (PDFs) and correlation matrices. They are particularly advantageous for handling incomplete measurements or missing values, a common challenge in the development of soil parameter databases in engineering practice.

Appendix

Under the framework of BCS, a signal (e.g., a column vector $\boldsymbol{f}$ with a length of N) can be expressed as a weighted summation of a series of basis functions (Candès et al. 2006; Donoho 2006; Tropp & Gilbert 2007):

$$\boldsymbol{f} = \mathbf{B}\boldsymbol{\omega} \tag{A1}$$

where $\mathbf{B}$ is an $N \times N$ orthonormal matrix; each column of $\mathbf{B}$ represents a pre-specified basis function, such as the cosine function (e.g., Salomon 2004); $\boldsymbol{\omega}$ is a weight vector with a length of N, and most elements of $\boldsymbol{\omega}$ are very small or almost zero, except for a few non-trivial ones due to the compressibility of $\boldsymbol{f}$ (Wang & Zhao 2016).

When non-trivial components in $\boldsymbol{\omega}$ are estimated from measurement data on $\boldsymbol{f}$, denoted by a row vector $\boldsymbol{y}$ with M elements ($M < N$), $\boldsymbol{f}$ can be approximately reconstructed using non-trivial coefficients, denoted as $\boldsymbol{\omega}_s$, and corresponding basis functions (Ji et al. 2008; Xu et al. 2021b, 2022). When the Bayesian method is used to estimate $\boldsymbol{\omega}_s$, $\boldsymbol{\omega}_s$ follows a multivariate Gaussian distribution with a mean of $\boldsymbol{\mu}_{\boldsymbol{\omega}_s}$ and covariance matrix $\mathbf{COV}_{\boldsymbol{\omega}_s}$. Then, the mean and covariance of the reconstructed $\boldsymbol{f}$ are expressed as (Wang & Zhao 2017):

$$\begin{aligned} \boldsymbol{\mu}_{\hat{f}} &= \mathbf{B}\,\boldsymbol{\mu}_{\boldsymbol{\omega}_s} \\ \mathbf{COV}_{\hat{f}} &= \mathbf{B}\,\mathbf{COV}_{\boldsymbol{\omega}_s}\,\mathbf{B}^{\mathrm{T}} \end{aligned} \tag{A2}$$

BCS results can be used together with KL expansion to generate random field samples directly from sparse measurements $\boldsymbol{y}$, which can be expressed as (Wang et al. 2018):

$$\boldsymbol{f} = \boldsymbol{\mu}_f + \sum_{i=1}^{s} \sqrt{\lambda_{\hat{f}_i}}\,\boldsymbol{V}_i\,z_i \tag{A3}$$

where S is the number of non-trivial approximation coefficients in $\boldsymbol{\omega}_s$; $\lambda_{\hat{f}_i}$ and $\mathbf{V}_i$ represent the i-th eigenvalue and eigenvector of $\mathbf{COV}_{\hat{f}}$, respectively; $\mathbf{z}_i$ ($i = 1,2,\ldots, S$) is a set of uncorrelated random variables with zero-mean and unit-variance. In this study, $\mathbf{z}_i$ ($i = 1, 2,\ldots, S$) is taken as a standard Gaussian random variable (Wang et al. 2019). The BCS-KL can generate many complete random field samples directly from sparse measurements (i.e., incomplete multivariate measurement data in each row in this study). The implementation procedure of BCS-KL is shown below.

Algorithm: Bayesian compressive sampling (BCS) with Karhunen–Loève (KL)

Input: Measurement vector $\boldsymbol{y}$, measurement matrix $\boldsymbol{\Psi}$, basis function $\mathbf{B}$.

Output: Random field samples of $\hat{\boldsymbol{f}}$

1. Compute the matrix $\mathbf{A} = \boldsymbol{\Psi}\mathbf{B}$.
2. Initialize hyperparameters: α = small positive values.
3. Set initial covariance matrix $\mathbf{D} = \mathrm{diag}(\alpha)$.
4. Compute initial mean and covariance:

$$\boldsymbol{\mu}_{\boldsymbol{\omega}_s} = \mathbf{H}\mathbf{A}^{\mathrm{T}}\boldsymbol{y}$$

$$\mathbf{COV}_{\boldsymbol{\omega}_s} = \frac{\mathrm{d_n}\mathbf{H}}{\mathrm{c_n} - 1}$$

$$\mathrm{c_n} = \mathrm{M}/2 + \mathrm{c};\ \mathrm{d_n} = \mathrm{d} + (\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{\omega}_s}{}^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\mu}_{\boldsymbol{\omega}_s})/2;\ \mathrm{c} = \mathrm{d} = 10^{-4}$$

$$\mathbf{H} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{D})^{-1}$$

5. Iterate until convergence:

a. Update $\mathbf{D}$ using maximum likelihood estimation.
b. Recompute $\mathbf{H}$, $\boldsymbol{\mu}_{\boldsymbol{\omega}_s}$, and $\mathbf{COV}_{\boldsymbol{\omega}_s}$
c. Check convergence condition.

6. Compute reconstructed statistics:

$$\boldsymbol{\mu}_{\hat{f}} = \mathbf{B}\boldsymbol{\mu}_{\omega_s}$$
$$\mathbf{COV}_{\hat{f}} = \mathbf{B}\mathbf{COV}_{\omega_s}\mathbf{B}^{\mathrm{T}}$$

7. Generate random field samples:

a. Compute eigenvalues/eigenvectors of $\mathbf{COV}_{\hat{f}}$

b. Apply KL expansion: $\hat{f} = \mu_{\hat{f}} + \sum_{i=1}^{s} \sqrt{\lambda_{\hat{f}_i}}\,\mathbf{V}_i\,\mathbf{z}_i$

8. Return $\hat{f}$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10064-025-04265-4.

Acknowledgements The work described in this paper was supported by a grant from the National Natural Science Foundation of China (grant No: 52130805). The financial support is gratefully acknowledged.

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

Declarations

Conflict of interest 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.

References

Asem P, Gardoni P (2021) A generalized Bayesian approach for prediction of strength and elastic properties of rock. Eng Geol 289:106187

Bozorgzadeh N, Bathurst RJ (2022) Hierarchical Bayesian approaches to statistical modelling of geotechnical data. Georisk: Assess Manag Risk Eng Syst Geohazards 16(3):452–469

Candès EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 59(8):1207–1223

Ching JY, Phoon KK (2014) Transformations and correlations among some clay parameters—the global database. Can Geotech J 51(6):663–685

Ching JY, Phoon KK (2020) Constructing a site-specific multivariate joint probability distribution using sparse, incomplete, and spatially variable (MUSIC-X) data. J Eng Mech 146(7):04020061

Ching JY, Phoon KK, Yang ZY, Stuedlein AW (2022) Quasi-site-specific multivariate probability distribution model for sparse, incomplete, and three-dimensional spatially varying soil data. Georisk 16(1):53–76

Ching JY, Wu S, Phoon KK (2014) Transformations and correlations among some clay parameters—the global database. Can Geotech J 51(6):663–685

Ching JY, Wu S, Phoon KK (2021) Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model. J Eng Mech 147(10):04021069

Chwała M, Phoon KK, Uzielli M, Zhang J, Zhang L, Ching J (2023) Time capsule for geotechnical risk and reliability. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 17(3):439–466. https://doi.org/10.1080/17499518.2022.2136717

Deng QX, He J, Cao ZJ, Papaioannou I, Li DQ, Phoon KK (2022) Bayesian learning of Gaussian mixture model for calculating debris flow exceedance probability. Georisk 16(1):154–177

D'Ignazio M, Phoon KK, Tan SA, Lansivaara T (2016) Correlations for undrained shear strength of Finnish soft clays. Can Geotech J 53(10):1628–1645

Donoho DL (2006) Compressed Sensing IEEE Trans Inf Theory 52(4):1289–1306

Feng S, Vardanega PJ (2019) Correlation of the hydraulic conductivity of fine-grained soils with water content ratio using a database. Environ Geotechnics 6(5):253–268

Feng SJ, Du FL, Shi ZM, Shui WH, Ke T (2015) Field study on the reinforcement of collapsible loess using dynamic compaction. Eng Geol 185:105–115

Gao JL, Wang JD, Wei Y, Jiao ST, Jing M, Gao JL, Wang C (2020) Study on treatment of loess landslide based on nanosilica and fly ash composite stabilizer filling fissures. Adv Civil Eng 2020:8884981

GBJ25-90 (1991) Code for building construction in collapsible loess regions. China Planning Press (in Chinese), Beijing

Guan Z, Wang Y (2021) Non-parametric construction of site-specific non-Gaussian multivariate joint probability distribution from sparse measurements. Struct Saf 91:102077

Guan Z, Wang Y, Phoon KK (2024) Dictionary learning of spatial variability at a specific site using data from other sites. J Geotech Geoenv Eng 150(9):04024072

Guan Z, Wang Y, Phoon KK (2025) Data-driven geotechnical site recognition using machine learning and sparse representation. Eng Geol 346:107893

Ibsen CC (2019) On the use of Bayesian networks as a meta-modelling approach to analyze uncertainties in slope stability analysis. Georisk 13(1):53–65

Ji SH, Xue Y, Carin L (2008) Bayesian compressive sensing. IEEE Trans Signal Process 56(6):2346–2356

Jiang MJ, Hu HJ, Liu F (2012) Summary of collapsible behavior of artificially structured loess in oedometer and triaxial wetting tests. Can Geotech J 49(10):1147–1157

Kendall MG, Stuart A (1963) The Advanced Theory of Statistics. Oxford University Press, New York

Kurnaz TF, Dagdeviren U, Yildiz M, Ozkan O (2016) Prediction of compressibility parameters of the soils using artificial neural network. Springerplus 5:1801

Leng XL, Wang C, Zhang J, Sheng Q, Cao SL, Chen J (2021) Deformation development mechanism in a loess slope with seepage fissures subjected to rainfall and traffic load. Front Earth Sci 9:769257

Li YR (2018) A review of shear and tensile strengths of the Malan Loess in China. Eng Geol 236:4–10

Li YR, Zhao JG, Li B (2018) Loess and Loess Geohazards in China. Taylor & Francis Group, London, UK
Liu SY, Zou HF, Cai GJ, Bheemasetti BV, Puppala AJ, Lin J (2016) Multivariate correlation among resilient modulus and cone penetration test parameters of cohesive subgrade soils. Eng Geol 209:128–142
McLachlan G, Peel D (2000) Mixtures of Factor Analyzers. Oxford University Press, New York
New Zealand Geotechnical Database (2012) NZGD. NZGD-2012
Phoon KK (2018) Editorial for special collection on probabilistic site characterization. ASCE-ASME J Risk Uncertainty Eng Syst Part A Civ Eng 4(4):02018002
Phoon KK, Ching JY, Shuku T (2022) Challenges in data-driven site characterization. Georisk 16(1):114–126
Phoon KK, Huang HW, Quek ST (2002) Simulation of second-order processes using Karhunen-Loeve expansion. Comput Struct 80(12):1049–1060
Phoon KK, Huang HW, Quek ST (2005) Simulation of strongly non-Gaussian processes using Karhunen-Loeve expansion. Probabilistic Eng Mech 20(2):188–198
Phoon KK, Zhang W (2023) Future of machine learning in geotechnics. Georisk: Assess Manag Risk Eng Syst Geohazards 17(1):7–22
PRC MOHURD (2018) National Standard GB50025–2018. Standard for building construction in collapsible loess regions. Beijing, Ministry of Housing and Urban-Rural Development of the People's Republic of China (in Chinese)
Rasmussen CE (1999) The infinite Gaussian mixture model. Proc 12th Int Conf Neural Inf Process Syst MIT Press, Cambridge 554–560
Salomon D (2004) Data Compression: The Complete Reference. Springer Science and Business Media, New York
Tang C, Phoon KK (2024) Databases for Data-Centric Geotechnics: Geotechnical Structures. CRC Press
Tian HM, Wang Y, Phoon KK (2025) Construction of Quasi-Site-Specific Geotechnical Transformation Models Using Bayesian Sparse Dictionary Learning. J Geotech Geoenv Eng 151(1):04024147
Tropp JA, Gilbert AC (2007) Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans Inf Theory 53(12):4655–4666
Wang Y, Akeju OV, Cao ZJ (2016) Bayesian Equivalent Sample Toolkit (BEST): an Excel VBA program for probabilistic characterization of geotechnical properties from limited observation data. Georisk 10(4):251–268
Wang Y, Cao ZJ (2013) Probabilistic characterization of Young's modulus of soil using equivalent samples. Eng Geol 159:106–118
Wang Y, Hu Y, Phoon KK (2022) Non-parametric modeling and simulation of spatiotemporally varying geo-data. Georisk 16(1):77–97
Wang Y, Zhao TY (2016) Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. Can Geotech J 53(9):1547–1559
Wang Y, Zhao TY (2017) Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. Géotechnique 67(6):523–536
Wang Y, Zhao TY, Cao ZJ (2015) Site-specific probability distribution of geotechnical properties. Comput Geotech 70:159–168
Wang Y, Zhao TY, Hu Y, Phoon KK (2019) Simulation of random fields with trend from sparse measurements without detrending. J Eng Mech 145(2):04018130
Wang Y, Zhao TY, Phoon KK (2018) Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. Can Geotech J 55(6):862–880
Xu JB, Wang Y, Zhang LL (2021a) Interpolation of extremely sparse geo-data by data fusion and collaborative Bayesian compressive sampling. Comput Geotech 134:104098
Xu JB, Zhang LL, Li JH, Cao ZJ, Yang HQ, Chen XY (2021b) Probabilistic estimation of variogram parameters of geotechnical properties with a trend based on Bayesian inference using Markov chain Monte Carlo simulation. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards 15(2):83–97. https://doi.org/10.1080/17499518.2020.1757720
Xu JB, Wang Y, Zhang LL (2022) Fusion of geotechnical and geophysical data for 2D subsurface site characterization using multi-sources Bayesian compressive sampling. Can Geotech J 59:1756–1773
Xu JB, Zhang LL, Wang Y, Wang CH, Zheng JG, Yu YT (2020) Probabilistic estimation of cross-variogram based on Bayesian inference. Eng Geol 277:105813
Xu L, Dai F, Tu X, Tham LG, Zhou Y, Iqbal J (2014) Landslides in a loess platform. North-West China Landslides 11(6):993–1005 (in Chinese)
Yang J, Bai XH (2015) Nonlinear compression stress-strain relationship of compacted loess and its application to calculation of foundation settlement. Rock Soil Mech 36(4):1002–1008 (in Chinese)
Zhang DM, Zhou YL, Phoon KK, Huang HW (2020) Multivariate joint probability distribution of Shanghai clay properties. Eng Geol 273:105675
Zhou JX, Zhu C, Zheng JM, Wang XH, Liu ZH (2002) Landslide disaster in the loess area of China. J Forest Res 13(2):157–161
Zhuang JQ, Peng JB, Wang GH, Javed I, Wang Y, Li W (2018) Distribution and characteristics of landslide in loess plateau: a case study in Shaanxi province. Eng Geol 236:89–96