This paper was downloaded from the Online Library of the International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE). The library is available here:

https://www.issmge.org/publications/online-library

This is an open-access database that archives thousands of papers published under the Auspices of the ISSMGE and maintained by the Innovation and Development Committee of ISSMGE.

The paper was published in the proceedings of the 20th International Conference on Soil Mechanics and Geotechnical Engineering and was edited by Mizanur Rahman and Mark Jaksa. The conference was held from May 1st to May 5th 2022 in Sydney, Australia.
Data-driven subsurface stratigraphy from prior knowledge and sparse site-specific measurements using multiple point statistics

Chao Shi & Yu Wang
Department of Architecture and Civil Engineering, City University of Hong Kong, China, chaoshi6-c@my.cityu.edu.hk

ABSTRACT: Delineation of subsurface stratigraphy is an essential task for geotechnical site characterization. It is usual practice to use straight lines to connect stratum boundaries from adjacent measurements for development of subsurface stratigraphy. This usual practice is acceptable for a small or medium-sized site with simple geology, but difficulties are encountered for interpretation of complex stratigraphic relationships, particularly when only limited site-specific measurements are available. On the other hand, valuable prior knowledge of local geology from previous sites with similar geological settings has not been fully utilized. In this study, a novel non-parametric and data-driven Bayesian machine learning method based on Multiple Point Statistics (MPS) is proposed to interpolate subsurface soil stratigraphy from limited site-specific measurements and an ensemble training image. Both a simulation example and real geological profiles from a reclamation project in Hong Kong are used to illustrate the proposed approach. It is found that the proposed method accurately interprets subsurface stratigraphy, and areas of large interpretation uncertainties are explicitly quantified.

1 INTRODUCTION

Subsurface soil classification and stratification are an essential task for geotechnical site characterization. The major tasks involve determination of soil stratigraphic boundaries, soil types, and distribution of lenses, transitions, and other features (e.g., inclusions) (e.g., Hegazy and Mayne 2002). For practical site characterization, only a small portion of the target ground is carefully investigated (e.g., Clayton 2001). This may result in significant uncertainty and ambiguity in interpretation of soil properties as well as subsurface stratigraphic profiles.

There are mainly three kinds of approaches for delineation of subsurface stratigraphic boundaries of soils: (i) linear interpolation of soil boundaries between site-specific measurements based on visual examination and prior engineering judgement (e.g., Houshsy and Houshsy 2013); (ii) prior determination of spatially varying soil properties followed by soil classification using empirical soil behavior charts (e.g., Robertson 2010) and (iii) inference of spatial distribution of categorical variables using advanced probabilistic or geostatistical models (e.g., Gómez-Hernández and Wen 1998; Wang et al. 2017; Wang 2020).

Linear interpolation of stratigraphic soil boundaries based on soil categories revealed from site-specific measurements (e.g., boreholes or cone penetration tests (CPTs)) is most popular among all the approaches. As geological soil boundaries are predominantly horizontal, the laminar or stratified geological profiles interpreted from the simplified method are acceptable for projects with simple geology. However, great challenges are encountered when complex soil deposits exist and only sparse and limited site-specific measurements are available.

It is imperative to have an effective spatial interpolation method for delineation of subsurface stratigraphy from sparse and limited measurement data. In this study, a data-driven Bayesian machine learning approach based on multiple point statistics (MPS) is proposed to develop non-stationary subsurface stratigraphy from sparse site-specific measurements. Note that it is the first time that MPS is applied to geotechnical engineering. All the potential stratigraphic relationships of the soil deposits are adaptively learned from an ensemble training image, which reflects all the prior geological knowledge of the investigated ground. MPS interpolates subsurface stratigraphy by learning potential stratigraphic features from the training image while conditioning on all the sparse measurements. The application of training image for subsurface stratigraphy effectively tackles the intrinsic dilemma associated with sparse data in geotechnical site characterization. More importantly, the proposed method not only interpolates the most probable subsurface stratigraphy but also quantifies the associated interpolation uncertainty. The performance of the proposed method is validated using a simulation example and real data from a real reclamation project in Hong Kong.

The reminder of this study is organized as follows. In the second section, the rationale of MPS for spatial interpolation of categorical variable is introduced, followed by detailed implementation procedures for delineation of subsurface stratigraphy. Subsequently, a heterogeneous geological profile is simulated to illustrate the performance of the proposed method.
In the fifth section, a real geological profile from a recent reclamation project in Hong Kong is collected to further demonstrate the proposed method. Finally, conclusions are drawn.

2 BASIC IDEA OF MULTIPLE POINT STATISTICS

In geosciences, indicator kriging is popular for spatial interpolation of categorical variable, and it primarily relies on two-point geostatistical relationships. In order to capture curvilinear features of geobodies, Guardiano and Srivastava (1993) extended the concept of indicator kriging and employed multiple point statistics, e.g., quadrivariate, for the determination of spatially varying geological patterns.

The typical process of direct MPS simulation (e.g., ENESIM) is shown in Figure 1. All the prior geological knowledge about the target ground is concisely reflected within an ensemble training image. In addition, all the site-specific measurements are stored within a single simulation image, where the colored cells (i.e., yellow, green and blue) represent measured soil types. The purpose of MPS simulation is spatial interpolation of soil types at those blank cells within the simulation image while conditioning on both the training image and available site-specific data in the simulation image.

A random simulation path, marked as red dashed line, is generated to visit each blank cell. Note that MPS simulation relies on an important concept called data event $d_n[Z(x), u]$, which is defined as a cluster of cells. The cluster consists of the central blank cell and its $n$ closest neighboring colored cells (i.e., available measurements) and associated spatial distance vector $u$. For instance, the three colored cells and the central blank cell in the simulation image in Figure 1 represent the data event $d_3[Z(x), u]$.

MPS simulation operates by searching similar data events, e.g., clusters with the same number and color of neighboring cells and the same spatial distance vector $u$, within the training image. The similarity between data events is gauged by a distance measure, which compares the proportion of nonmatching neighboring cells within a data event (Mariethoz and Renard 2010). A pair of data events with completely identical neighboring cells have a distance of zero. For instance, all the three conditioning cells of example 1 have the same color and spatial distance with respect to the central blank cell, and the associated distance is 0. In contrast, only two out of the three neighboring cells of example 2 are identical to those of the target data event in the simulation image, resulting in a distance value of $2/3$. A threshold distance value, $t$, for regulating the similarity between data events should be specified, and the default value is 0. That means only data events with a distance less than $t$ are included. Note that when more than one candidate data events are identified within the training image, a random sample is drawn from a collection of the qualified data events and, the associated central cell is pasted to the target data event in the simulation image.

A weighting scheme can also be defined for drawing a random sample based on its distance to the target data event. Subsequently, the interpolated cell is taken as a new measurement for subsequent conditional simulation. The above process repeats until all blank cells within the simulation image are interpolated.

3 PROPOSED NONPARAMETRIC AND DATA-DRIVEN METHOD

MPS can be viewed as a Bayesian supervised machine learning method (Shi and Wang 2020a, 2020b). All the valuable prior geological knowledge of a similar ground is concisely represented within a single training image. MPS integrates sparse and limited site-specific measurements with the training image for stochastic simulation. The use of prior geological knowledge for spatial interpolation effectively tackles the intrinsic difficulty associated with sparse measurement data in geotechnical engineering (e.g., Wang and Cao 2013; Wang et al. 2016).

3.1 Bayesian formulation and simulation

MPS performs conditional simulation by sampling similar data events from the training image while anchoring stochastic simulation to all available site-specific measurements. When
interpreted from the perspective of Bayesian formulation, the occurrence probability of a given data event \( d_n(Z(x), u) \) with soil category \( Z(x) = z \) at the location of central blank cell \( x \) of a simulation image is expressed as \( P(Z(x) = z|d_n[Z(x), u]) \). Based on the Bayesian theorem, the occurrence probability can be calculated as follows:

\[
P(Z(x) = z|d_n[Z(x), u]) = \frac{P(d_n[Z(x), u] \text{ and } Z(x) = z)}{P(d_n[Z(x), u])} \quad x \in \mathbb{R}^2
\]  

(1)

where \( x \) represents 2D coordinate in the simulation image; \( z \) is a categorical variable (i.e., soil type) and can take a set of \( N_c \) discrete states \( \{z_1, z_2, \ldots, z_{N_c}\} \); The numerator \( P(d_n[Z(x), u] \text{ and } Z(x) = z) \) represents the probability that both the data event \( d_n[Z(x), u] \) and the central blank cell with a categorical \( z \) occur at the location \( x \) in the simulation image; the denominator \( P(d_n[Z(x), u]) \) is the occurrence probability, or evidence in Bayesian formulation, of data event \( d_n[Z(x), u] \).

As only sparse measurements are available in the simulation image, all probability terms on the right-hand side of eq.1 cannot be obtained from the measurements, but indirectly estimated from a training image based on the assumption that training image and simulation image share the similar geological settings. All probability terms on the right-hand side of eq.1 cannot be directly borrowed from a nearby site with similar geological settings.

3.2 Distance metric of data events

The similarity between a simulation data event \( d_n[Z(x), u] \) and a training data event \( d_n[Z(y), u] \) relies on a distance measure, which calculates the proportion of nonmatching points within a data event using the following equation (Mariethoz and Renard 2010):

\[
P(d_n[Z(x), u], d_n[Z(y), u]) = \frac{\sum_{i=1}^{N_c} u_{i} I(c_i)}{\sum_{i=1}^{N_c} u_{i}}
\]

(6)

where \( a_i = \{0, 1\} \) if \( Z(x) = Z(y) \), \( a_i = \{1\} \) if \( Z(x) \neq Z(y) \), \( I(c_i) \) is an indicator function and has a value of 1 if the data event with the curly bracket occurs within the training image; \( N_c \) is the total cell number in the training image. Note that the distance calculated from above eq.6 does not consider direction dependence. Shi and Wang (2020b) explicitly take spatial anisotropy into consideration by incorporating a Mahalanobis distance (Mahalanobis 1936).

\[
I(c_i) = \|x - x_i\|_M = \sqrt{(x - x_i)^{T}M(x - x_i)} \quad x, x_i \in \mathbb{R}^2
\]

(7)

\[
M = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}
\]

(8)

where subscript \( M \) is Mahalanobis, superscript \( T \) is transpose operation; \( M \) is covariance matrix and its first entry \( a \) in the diagonal direction denotes anisotropic ratio.

3.3 Training image

As revealed by the Bayesian formulation in eq. 4, all the probability terms are directly estimated from the ensemble training image, which should exhaust all the potential subsurface stratigraphic connectivities of the target field. Note MPS cannot generate new stratigraphic pattern and directly learn soil features from the training image. Therefore, the performance of MPS can be affected if the training image only represents a small portion of potential stratigraphic features in the target field. In practical geotechnical site characterization, the training image may be directly borrowed from a nearby site with similar geological settings.

3.4 Quantification of uncertainty associated with interpolation

A completed simulation image, or a realization in MPS terminology, is obtained when all the blank cells in the simulation image are interpolated. The associated interpolation uncertainty can be explicitly quantified by statistical analysis of multiple realizations, which are generated using the same set of site-specific measurements and the same training image.

After multiple realizations are generated, the most probable interpolation can be derived by assigning each cell in the simulation image with the soil type of the highest frequency among all realizations. Accordingly, interpolation uncertainty can be quantified by evaluating the deviation of multiple realizations from the most probable interpolation using the following dispersion.

\[
P_D(x) = \frac{\sum_{k=1}^{N_r} |Z(x) \neq Z_{mp}(x)|}{N_r}
\]

(9)

where \( P_D(x) \) denotes the proportion of nonmatching categories between a realization and the most probable interpolation results at location \( x \) in the simulation image; \( Z(x) \) is the categorical variable of \( r \)-th realization at location \( x \); \( Z_{mp}(x) \) represents the most probable categorical value at location \( x \); \( N_r \) denotes the total number of realizations. A 2D dispersion colormap can be obtained by assembling dispersion at each simulation point.

The number of realizations is determined when the most probable interpolation does not change with the addition of additional realizations. A relative percentage change \( P_r \) in the most probable interpolation with every additional \( k \) realizations

5129
is defined in this study for determining the adequacy of realizations.

\[ P_c = \frac{\sum_{i=1}^{N_r} \left[ I(Z_{mp, Nr}(x_i)) \times I(Z_{mp, N_{r-k}}(x_i)) \right]}{N_i \times N_r} \]  

where \( Z_{mp, Nr}(x_i) \) and \( Z_{mp, N_{r-k}}(x_i) \) denote categorical variable of the \( i \)-th cell in the most probable interpolation results derived from \( N_r - k \) and \( N_r \) realizations, respectively. Here, \( k \) is taken to be 10 and a stopping criterion of 0.05\% for \( P_c \) is specified. Similarly, the number of realizations can also be determined based on the convergence of interpolation uncertainty or dispersion. Moreover, prediction accuracy, i.e., Acc, defined for comparing the interpolation results with the underlying true geological profile for the illustrative examples in the following sections.

\[ \text{Acc} = \frac{\sum_{i=1}^{N_{test}} I[Z_T(x_i) = Z_{mp}(x_i)]}{N_i \times N_{test}} \]

where \( I(\cdot) \) is an indicator function and equals to 1 when the \( i \)-th point in test image, i.e., \( Z_T(x_i) \), has the same categorical value (e.g., soil type) as that from the most probable interpolation, i.e., \( Z_{mp}(x_i) \), in the simulation image. It is worthwhile to mention that in practical geotechnical site characterization, the underlying true test geological profile is not available and is used for validation purpose only. The detailed implementation procedure of the proposed method is referred to Shi and Wang (2020b).

4 ILLUSTRATIVE EXAMPLE

![Color maps of (a) training and (b) test images used in illustrative examples.](image)

In this section, a pair of training and test image are generated to illustrate the proposed method. A 30m (length) × 15m (thickness) 2D geological profile with five different soil layers is first simulated here. The horizontal length and vertical depth are discretized using a resolution of 0.1m, resulting in a total of 300 × 150 points. The stratigraphic boundaries of the five soil types (see the labels I, II, III, IV, and V in Figure 2) are delineated using quadratic line with associated coefficients following Gaussian distributions.

\[ v = Ah^2 + Bh + C \]

where \( v \) is depth and \( h \) denotes horizontal distance; \( A, B, \) and \( C \) are coefficients. Table 1 summarizes the mean values and associated standard deviations of the three coefficients. The simulated training and test images are shown in Figure 2. As an illustration, only 5 vertical line profiles of equal horizontal spacing are taken from the test image as site-specific measurements.

Table 1. Coefficients for the function \( v = Ah^2 + Bh + C \)

<table>
<thead>
<tr>
<th>Function</th>
<th>( A )</th>
<th>( B )</th>
<th>( C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-0.02</td>
<td>0.08</td>
<td>0.5</td>
</tr>
<tr>
<td>II</td>
<td>-0.04</td>
<td>-0.2</td>
<td>2</td>
</tr>
<tr>
<td>III</td>
<td>1.0</td>
<td>0.3</td>
<td>12.5</td>
</tr>
<tr>
<td>IV</td>
<td>-</td>
<td>0.5</td>
<td>12</td>
</tr>
<tr>
<td>V</td>
<td>-2.0</td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>

Note: *Std.* denotes standard deviation.

5 SIMULATION RESULTS FROM PROPOSED METHOD

Multiple realizations conditioning on the training image in Figure 2a and 5 line measurements in Figure 2b are generated until the percentage change in the most probable interpolation is less than 0.05\% with every additional 10 realizations. As a result, 200 realizations in total are produced. Corresponding most probable interpolation derived from the 200 realizations is calculated and shown in Figure 3a. It is clear that the most probable interpolation can reasonably capture the overall stratigraphic relationships between different soils with an overall accuracy of 95.4\%. For better comparison, the true soil boundaries extracted from the test image are also superimposed on the most probable interpolation. It is observed that the largest discrepancies mainly lie at the boundaries along soil1 and soil5. This can be explained by the fact that the stratigraphic boundaries around soil1 has a different dipping direction as compared with that in the training image. In addition, the stratigraphic patterns around soil5 in the training image are quite limited, leading to inadequate training data events in the training image. It should be noted that the combination of training image and site-specific data for spatial interpolation is established based on the assumption that similar spatial patterns repeat in areas with similar geological settings. Therefore, if a dissimilar training image is adopted, the spatial interpolation performance can be compromised.

The dispersion of multiple realizations from the most probable interpolation is shown in Figure 3b. It is evident that bands of large dispersion mainly cluster around boundaries separating different soil layers. Dispersion values reduce to 0 at the locations of line measurement and gradually increase as the distance from available measurements increases. As the true test image is available from this illustrative example, the prediction accuracy, i.e., Acc, can be evaluated using eq. 11. Similar to the dispersion plot in Figure 3b, areas of large interpolation errors
Figure 3. MPS interpolation results: (a) most probable interpolation; (b) dispersion plot; (c) accuracy plot.

It should be noted that in real geotechnical practice, the test image is the target of spatial interpolation and is not available. However, based on the good correlation between the dispersion and accuracy plots, the dispersion alone can be used as an indication of spatial interpolation accuracy when the accuracy plot is not available.

6 REAL DATA EXAMPLE

Two geological profiles obtained from a recent reclamation project in Hong Kong are used to further validate the proposed method. Both 2D cross sections successively consist of Fill, disturbed marine deposit (DMD), marine deposit (MD), alluvial sand (Alls), alluvial clay (Allc), completely decomposed granite (CDG) and completely decomposed metasiltstone (CDM). The two cross-sections are parallel and have a horizontal separation of about 25m. Soil classification and associated subsurface geological profile were interpreted from over 600 in-situ measurements (e.g., cone penetration tests and boreholes) by experienced geologists.

Figure 4 shows the digitized training and test images. The geological cross-sections in the training and test images are in parallel and have a horizontal separation of approximately 25m. Each image has a total horizontal length and a total depth of 100m and 60m, respectively. Using a resolution of 1.0m and 0.5m for horizontal and vertical direction, each cross section is discretized into a grid of 100 (length) × 120 (depth). As an illustration, 5 line profiles with an equal spacing of 25m are extracted from the test image as site-specific measurements. In total, 200 realizations conditioning on 5 line measurements and the training image in Figure 4a are simulated. The derived most probable interpolation is shown in Figure 5a. It can be seen that the spatial stratigraphic patterns are reasonably captured by the proposed method, particularly the interbedded Alls/Allc. An overall accuracy of 90.0% using eq.11 is obtained. The deviation of multiple realizations from the most probable interpolation calculated using eq.9 and associated dispersion plot is shown in Figure 5b. It can be seen that areas of large dispersion mainly cluster at around boundaries separating different soil deposits. Similar spatial patterns are also reflected by the accuracy plot in Figure 5c.

Effects of measurement data number on MPS performance are also investigated. Another two cases where the measurement data number changes to 3 and 11 are also carried out. The total thickness of fine-grained materials, namely, MD and Allc, are calculated accordingly. For better comparison, the true total thickness calculated from the test image in Figure 4b is also included in Figure 6. When three measurements are available, equivalent to a horizontal spacing of 50m, the largest difference between predicted and actual thickness of MD and Allc is about...
machine learning method based on Multiple Point Statistics (MPS) is proposed to interpolate subsurface soil stratigraphy from limited site-specific measurements and an ensemble training image. The utilization of the training image enables the integration of existing engineering knowledge on local geology for spatial interpolation and effectively tackles the intrinsic difficulty associated with sparse measurements in geotechnical site characterization. Both a simulation example and real geological profiles from a reclamation project in Hong Kong are collected to illustrate the proposed approach. It is found that the proposed method not only infers subsurface stratigraphy accurately from sparse measurements but also quantifies interpolation uncertainty.

8 ACKNOWLEDGEMENTS

The work escribed in this paper was supported by grants from the Research Grants Council of Hong Kong Special Administrative Region, China (Project no. CityU 11213119 and CityU 11202121). The financial support is gratefully acknowledged.

9 REFERENCES


