Geoscience Frontiers 12 (2021) 339–350

Contents lists available at ScienceDirect

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Research Paper

Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties

Chao Shi, Yu Wang *

Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
Spatial interpolation
Multiple point statistics
Bayesian compressive sampling
Compressive sensing
Sparse measurement

ABSTRACT

Spatial interpolation has been frequently encountered in earth sciences and engineering. A reasonable appraisal of subsurface heterogeneity plays a significant role in planning, risk assessment and decision making for geotechnical practice. Geostatistics is commonly used to interpolate spatially varying properties at un-sampled locations from scatter measurements. However, successful application of classic geostatistical models requires prior characterization of spatial auto-correlation structures, which poses a great challenge for unexperienced engineers, particularly when only limited measurements are available. Data-driven machine learning methods, such as radial basis function network (RBFN), require minimal human intervention and provide effective alternatives for spatial interpolation of non-stationary and non-Gaussian data, particularly when measurements are sparse. Conventional RBFN, however, is direction independent (i.e. isotropic) and cannot quantify prediction uncertainty in spatial interpolation. In this study, an ensemble RBFN method is proposed that not only allows geotechnical anisotropy to be properly incorporated, but also quantifies uncertainty in spatial interpolation. The proposed method is illustrated using numerical examples of cone penetration test (CPT) data, which involve interpolation of a 2D CPT cross-section from limited continuous 1D CPT soundings in the vertical direction. In addition, a comparative study is performed to benchmark the proposed ensemble RBFN with two other non-parametric data-driven approaches, namely, Multiple Point Statistics (MPS) and Bayesian Compressive Sensing (BCS). The results reveal that the proposed ensemble RBFN provides a better estimation of spatial patterns and associated prediction uncertainty at un-sampled locations when a reasonable amount of data is available as input. Moreover, the prediction accuracy of all the three methods improves as the number of measurements increases, and vice versa. It is also found that BCS prediction is less sensitive to the number of measurement data and outperforms RBFN and MPS when only limited point observations are available.

1. Introduction

Spatial interpolation of field attributes from scatter measurements is an essential task for many disciplines, e.g., assessment of soil contamination in environmental engineering (e.g., Ersoy et al., 2004), reservoir facies modelling in oil and gas industry (Mariethoz and Caers, 2014), subsurface heterogeneity investigation in mining engineering (Carlson, 1991), prediction of temperatures and rainfall in meteorological engineering (Thornton et al., 1997), and mapping soil nutrient levels in agricultural engineering (Zhang et al., 2007). An accurate interpolation of field properties is a prerequisite for understanding the spatial structures and performing follow-up analysis. Geostatistics provides a powerful tool for assessing spatial variability and heterogeneity. Of all the statistical models, kriging has been popular as it can provide a best linear unbiased estimate (BLUE) of the attribute as well as the associated interpolation uncertainty at the un-sampled locations (e.g., Li and Heap, 2008). However, kriging is established on the assumption of a stationary Gaussian random field (e.g., Webster and Oliver, 2007) and a site- or data-specific autocorrelation structure (i.e., semi-variogram) between spatial measurements is required for a successful spatial interpolation (Oliver and Webster, 2014). The determination of appropriate semi-variogram function and associated parameters (i.e., nugget, still and range) requires extensive measurements, which are frequently unavailable in engineering practice of site investigation. As a result, an unreliable autocorrelation structure derived from limited measurements can lead to inaccurate estimation of spatially varying geotechnical properties.

* Corresponding author.
E-mail addresses: chaoshi6-c@my.cityu.edu.hk (C. Shi), yuwang@cityu.edu.hk (Y. Wang).
Peer-review under responsibility of China University of Geosciences (Beijing).

https://doi.org/10.1016/j.gsf.2020.01.011
Received 13 October 2019; Received in revised form 19 November 2019; Accepted 10 January 2020
Available online 10 February 2020

More importantly, there is a high risk that the reconstructed field tends to reflect the mechanism of two-point parametric variogram model rather than the true natural process (Breiman, 2001), which is often non-stationary and non-Gaussian.

Spatial interpolation of non-stationary and non-Gaussian natural processes has been a long-standing problem for practitioners (e.g., Risser, 2016). Parametric statistical models (e.g., Kriging) in spatial interpolation mainly deal with stationary statistics and oversimplifies natural processes and phenomena, which are often non-stationary, non-Gaussian and non-heteroscedastic (Mariethoz and Caers, 2014). As pointed out by Mariethoz et al. (2009), an ideal model for estimating and reproducing statistical properties of natural phenomena should be non-parametric and data-driven. Emerging machine learning methods provide alternative spatial interpolators (Li and Heap, 2008; Li et al., 2011) for modelling those non-stationary and non-Gaussian processes. Machine learning is a popular soft computing technique, which aims to solve practical problems by progressively and adaptively exploiting imprecision, uncertainty and partial truth (Devendra, 2008). Compared with geostatistical and other deterministic spatial models (e.g.,. inverse distance weighting), machine learning methods are more flexible and less assumption dependent (e.g., assumption of a pre-selected parametric function form is not needed for machine learning methods). The superior spatial predictive performance of machine learning approaches have been proven in various disciplines, such as environmental and meteorological science (Appelhans et al., 2015; Li, 2019).

Among various machine learning approaches, network-based models are appealing to engineering practitioners (e.g. Zhang and Goh, 2014; Zhang et al., 2019) as they can adaptively learn the complex and nonlinear relationships from the measured points. The prediction capacity of neural network depends on the adopted activation function and number of layers and hidden neurons. It is demonstrated by Park and Sandberg (1991) that radial basis function network (RBFN) with one hidden layer can universally approximate any forms of functions. RBFN is a supervised machine learning algorithm and has been successfully applied to the interpolation of spatially varying non-stationary and non-Gaussian variables (Powell, 1987; Lin and Chen, 2004; Rusu and Rusu, 2006). However, conventional radial basis functions (e.g., multiquadric and inverse multiquadric) depend only on radial distance and cannot model anisotropic data. On the other hand, it is well-recognized that geotechnical data are anisotropic due to geological histories. In addition, uncertainty in the predictions cannot be quantified for RBFN. In this study, the conventional RBFN is improved to incorporate geotechnical anisotropy and integrated with ensemble learning method to explicitly quantify prediction uncertainty in spatial interpolation. A comparative study is also performed to benchmark the proposed RBFN method with two other novel non-parametric data-driven approaches, namely Multiple Point Statistics (MPS) and Bayesian Compressive Sensing (BCS). The accuracy in replicating the best estimate and the ability to quantify interpolation uncertainty are explicitly compared.

The reminder of this study is organized as follows. In the second section, the ensemble RBFN method is proposed and the improvements and modifications to the conventional RBFN are highlighted. Numerical examples of a 2D nonstationary non-Gaussian random field are simulated in the third section, which are used as both illustrative example for the ensemble RBFN method and the comparative study. Results of applying the proposed ensemble RBFN to the illustrative example are then presented in the fourth section. Sections five and six compare performances of both MPS and BCS with ensemble RBFN in reconstructing spatially varying non-stationary and non-Gaussian properties. Subsequently, effects of measurement data number on the reconstructed fields by different methods are discussed. Finally, conclusions regarding the capacity of different models in estimating spatially varying geotechnical properties are drawn.

2. Ensemble radial basis function network (RBFN)

Radial basis function network, RBFN, was introduced by Hardy (1971) for multivariate interpolation and has been a popular method for solving scatter point interpolation problems. For RBFN, the interpolated values at un-sampled locations are expressed as a linear combination of radial basis functions with centroids at point observations. In mathematical terms, radial basis function is a real-value radial basis function φ, whose value solely depends on the relative distance r to another point c, called a centroid. Any functions satisfying the property $\varphi(r) = \varphi(\| x - c \|)$ can be used as a radial basis function. For instance, given N points $[x_1, x_2, x_3 \ldots x_n]$ and corresponding responsesy $= [y_1, y_2, y_3 \ldots y_n]^T$, a RBFN interpolator can be established for any point x:

$$y(x) = \sum_{i=1}^{n} \omega_i \varphi(\| x - x_i \|), x \in R^m \tag{1}$$

where x_i represents measurement point in m dimension space and $\| \cdot \|$ denotes Euclidean distance; ω_i is the weight coefficient associated with a radial function, whose center is at x_i. By assuming x coincides with each point in $[x_1, x_2, x_3 \ldots x_n]$, the radial basis matrix $\Phi_{i,j} = \varphi(\| x_i - x_j \|)$, $i, j = 1, 2, \ldots, N$ can be derived. Accordingly, the prediction y_{N+1} at any point x_{N+1} can be solved if Φ is invertible.

$$\omega = \Phi^{-1} y \tag{2}$$

$$y_{N+1} = \phi \Phi^{-1} y \tag{3}$$

where $\phi = [\varphi(\| x_{N+1} - x_1 \|), \varphi(\| x_{N+1} - x_2 \|), \ldots, \varphi(\| x_{N+1} - x_N \|)]$. Ideally, each centroid should associate with a problem-specific shape factor (Lin and Chen, 2004). For brevity, a single shape factor is applied to all the centroids in this study. In addition, the number of hidden neurons has a major impact on the performance of a network. The selection of optimal hidden neuron number relies on the modeler's experience. For practical geotechnical engineering, measurements are normally limited in numbers and locations. Each observation is therefore assumed to carry certain information and is valuable in reconstructing the original field. In this study, the centroids are taken to coincide with all the available point observations. In other words, the number of hidden neurons equals to the number of measurements. Graphically, the calculation process can be viewed as a three-layer feedforward neural network, as shown in Fig. 1.

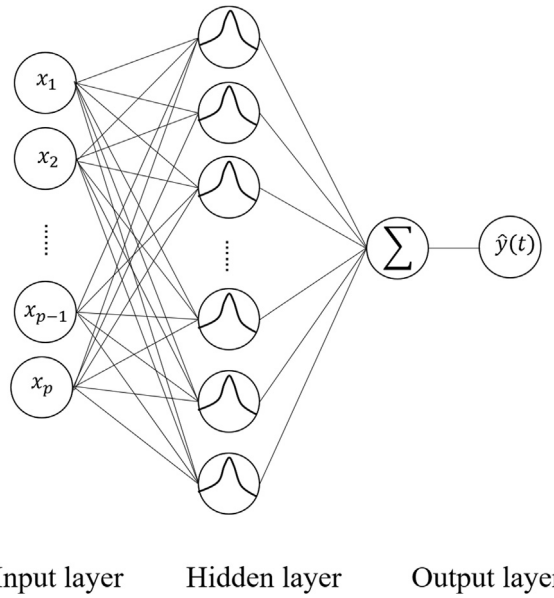

Fig. 1. Illustration of radial basis function network.

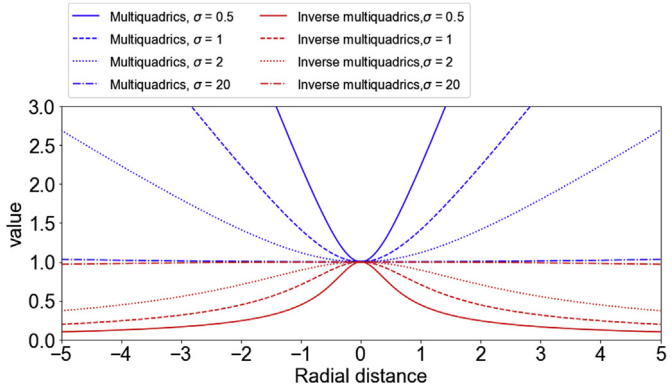

Fig. 2. Illustration of commonly used radial basis function with different shape factor.

2.1. Generalized Euclidean distance for anisotropic data

Radial basis function can take different forms such as spline functions (Zhang and Goh, 2016). For spatial interpolation, commonly used radial basis functions include multiquadric and inverse multiquadratic (e.g., Hardy, 1971; Du Toit, 2008):

$$\text{Multiquadric } \varphi(\parallel x - x_i \parallel) = \sqrt{1 + \frac{(x-x_i)^T (x-x_i)}{\sigma^2}},\ x \in R^m \tag{4}$$

$$\text{Inverse multiquadric } \varphi(\parallel x - x_i \parallel) = \frac{1}{\sqrt{1 + \frac{(x-x_i)^T (x-x_i)}{\sigma^2}}},\ x \in R^m \tag{5}$$

where σ represents shape factor, which governs the decay of a basis with radial distance. Fig. 2 illustrates variation of multiquadric and inverse multiquadric functions with different shape factors. Both radial basis functions exhibit steep changes in values when shape factor is small. In this study, both multiquadric and inverse multiquadric functions are used for spatial interpolation of non-stationary and non-Gaussian natural processes.

It is worth pointing out that conventional radial basis functions use uninformed Euclidean distance to calculate relative distance to a centroid and that anisotropy of the spatial distribution cannot be modelled. However, natural processes normally exhibit statistical anisotropy and have preferred orientations for spatial auto-correlation. This problem can be effectively overcome by incorporating a generalized Euclidean distance, i.e., Mahalanobis distance (Mahalanobis, 1936), for radial functions (Xu et al., 2012):

$$r_i = \parallel x - x_i \parallel_M = \sqrt{(x-x_i)^T \boldsymbol{M}\ (x-x_i)} \tag{6}$$

where r_i denotes the Mahalanobis distance between two points; $\boldsymbol{M}$ represents covariance matrix, which can be adjusted to reflect unequal decay rates at different directions. Mahalanobis distance reduces to the conventional Euclidean distance if an identity matrix is used as the covariance matrix. In this study, a 2 by 2 diagonal matrix is proposed for a 2D problem, as shown by Eq. (7). Specifically, the first element of the diagonal matrix represents anisotropic ratio, a, and the second element of the diagonal matrix is fixed at 1.

$$\boldsymbol{M} = \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \tag{7}$$

Therefore, the interpolant y at a target location x is a linear combination of translates of a radial basis function,

$$y(x) = \sum_{i=1}^{n} \omega_i \sqrt{\frac{(x-x_i)^T \boldsymbol{M}\ (x-x_i)}{\sigma^2} + 1}, x \in R^m \quad \text{for multiquadric function} \tag{8}$$

$$y(x) = \sum_{i=1}^{n} \omega_i \frac{1}{\sqrt{\frac{(x-x_i)^T \boldsymbol{M}\ (x-x_i)}{\sigma^2} + 1}},\ x \in R^m \quad \text{for inverse multiquadric function} \tag{9}$$

Eqs. (8) and (9) are straightforward and involve only two unknown parameters, namely shape factor σ and anisotropic ratio a for diagonal matrix $\boldsymbol{M}$. Both parameters can be determined via grid search within typical ranges, and the detailed determination process is explained in subsection 2.4.

2.2. Characterization of interpolation uncertainty

Spatial interpolation using Eqs. (8) and (9) strongly relies on the choice of shape factor and anisotropic ratio. Once both parameters are determined, deterministic responses at an un-sampled location can be calculated. Aside from best estimate of spatial interpolation, it is also worthwhile to quantify the associated prediction uncertainty. Li et al. (2010) fitted a RBFN with Gaussian radial basis function and a single shape factor from Leave-One-Out Cross-Validation (LOOCV) and considered each deterministic prediction as a realization of Gaussian stochastic processes. The prediction uncertainty was then quantified by statistical analysis of multiple realizations. The same principle is applied in this study to learn the optimal combination of anisotropic ratio and shape factor and its associated prediction uncertainty for a given radial basis function from scatter measurements at hand. All measurements are randomly divided into two groups, namely, training and test data, with a split ratio of 50:50. The training data is used to construct a radial basis network and derive corresponding weight coefficients. Subsequently, the test data is fed into the network to generate predictions using different combinations of anisotropic ratio and shape factor. The pair of anisotropic ratio and shape factor is determined via grid search within given typical ranges. The optimal combination is determined when a minimum test error is obtained. The above procedure is aligned with the essence of cross-validation (Domingos, 2012), which is a typical procedure in machine learning practice to avoid overfitting and trapping in local maxima. A series of realizations are generated by repeating the above splitting procedures, and associated prediction uncertainty can be obtained from statistical analysis of multiple realizations.

It is also worthwhile to characterize prediction uncertainty from the above shuffle process. Essentially, interpolation or prediction error (i.e. $\varepsilon(f^*)$) can be decomposed into three components, namely, approximation error, sampling error (or estimation error) and Bayes error (Hastie et al., 2005; Cucker and Zhou, 2007).

$$\varepsilon(f^*) = [\varepsilon(f^*) - \varepsilon(f_H)] + [\varepsilon(f_H) - \varepsilon_N(\widehat{f}_D)] + \varepsilon_N(\widehat{f}_D) \tag{10}$$

where $\varepsilon(f^*)$ denotes the overall interpolation error; $\varepsilon(f^*) - \varepsilon(f_H)$ represents approximation error, which is characterized by the complexity of hypothesis space (e.g., linear, spline, kernel machine or neural network); $\varepsilon(f_H) - \varepsilon_N(\widehat{f}_D)$ stands for estimation error, which measures the distance between the best approximation function in the hypothesis space and the learned function from limited measurements, and it vanishes when data point goes to infinity; $\varepsilon_N(\widehat{f}_D)$ refers to Bayes error, which represents noise intrinsic to measurements and is irreducible.

Because RBFN with one hidden layer can universally approximate any forms of functions (Park and Sandberg, 1991), its approximation error should be minimal. On the other hand, Bayes error is associated with measurements and cannot be reduced. Therefore, key component in the prediction error is the second term in Eq. (10), estimation error. The estimation error can be effectively reduced when the best approximation

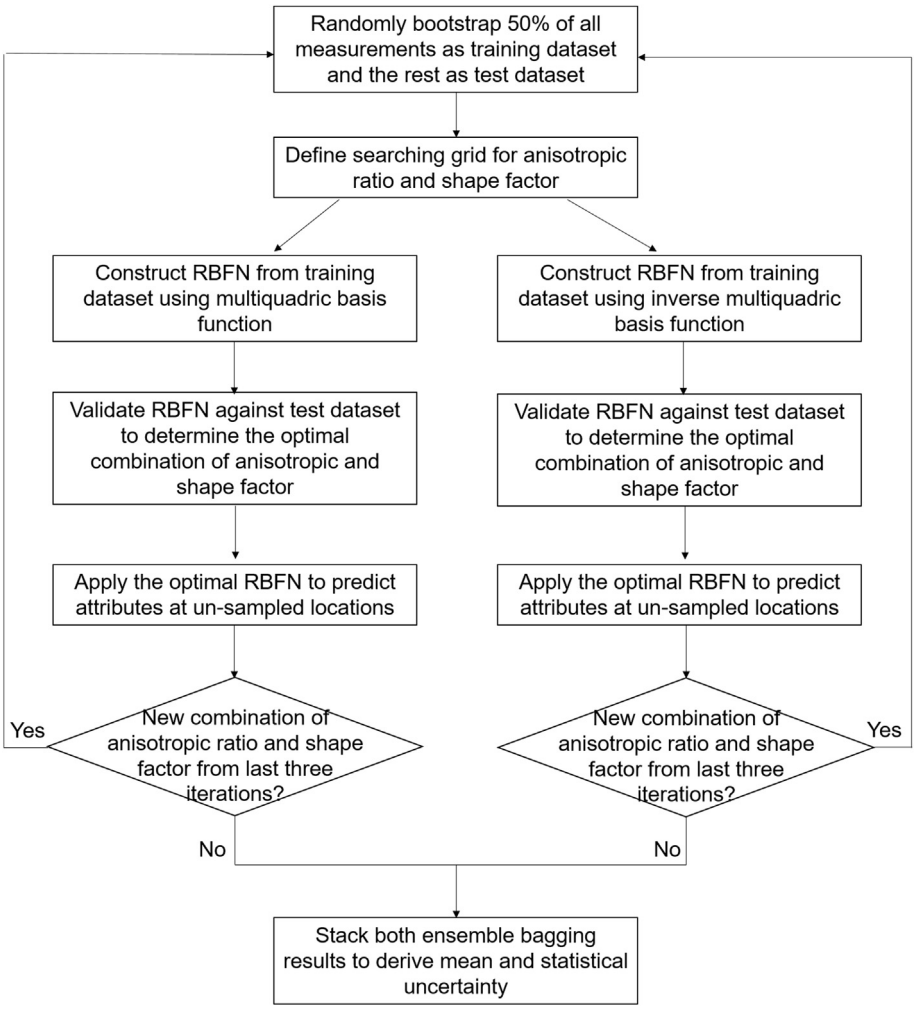

Fig. 3. Flowchart of the proposed ensemble RBFN method.

model in the hypothesis space (H) is obtained. Selection of the most suitable radial basis function for a given set of non-stationary and non-Gaussian data is not trivial, and it requires a sound understanding of the intrinsic spatial patterns of the concerned geotechnical properties. Instead of searching for the most appropriate radial basis function, ensemble learning (Zhou, 2012) is used in this study to capture "model uncertainty" by averaging predictions over multiple models consistent with the training data, as described in the following subsection.

2.3. *Ensemble learning for interpolation uncertainty quantification*

Ensemble learning is a common paradigm in machine learning to combine several learners for solving the same problem, as single ordinary machine learning approach frequently only learns a sub-optimal model from a given hypothesis space (Zhang and Ma, 2012). The integration of multiple learners enables the approximation to the true hypothesis with reduced model errors. It is also conceivable that when the adopted multiple learners (e.g., RBFN in this study) reasonably occupy the whole hypothesis space, the predication uncertainty associated with the ensemble learning can be explicitly quantified by statistical analysis of the results from those multiple learners.

In this study, multiquadric and inverse multiquadric functions are integrated within an ensemble learning framework to quantify estimation error associated with the proposed RBFN. The adopted ensemble learning can be regarded as a modified bagging algorithm, which stacks two different bagging results with equal weights for deriving the final mean and prediction uncertainty. When we examine multiquadric and inverse multiquadric basis functions in Fig. 2, it is evident that both radial basis functions are "bell-shaped" and exhibit opposite trends at enlarged radial distance. For instance, multiquadrics exhibit an increasing trend with an increase in radial distance, while a decreasing trend is observed for inverse multiquadrics. This pair of radial basis functions can be used together to reasonably approximate a value at arbitrary radial distance, filling up the whole space. Moreover, both radial basis functions converge to the same horizontal line when shape factor tends to be infinite. The prediction uncertainty associated with the proposed RBFN can also be explicitly quantified by statistical analysis of the results from the ensemble learning.

2.4. *Implementation procedure*

For clear references, key steps of the proposed ensemble RBFN method to interpolate spatially varying non-stationary and non-Gaussian data are illustrated in Fig. 3 and summarized below.

(i) Obtain measurement data y and associated spatial coordinates x and randomly split all measurements into training and test dataset with a ratio of 50:50, define possible ranges for anisotropic ratio and shape factor

(ii) Calculate corresponding Mahalanobis distance of training dataset using Eq. (6) and radial basis matrix, Φ, using Eqs. (4) and (5) for multiquadric and inverse multiquadric functions, respectively. Obtain weight coefficient, ω , using Eq. (2) and predict properties of interest at spatial coordinates of test dataset using Eq. (3).

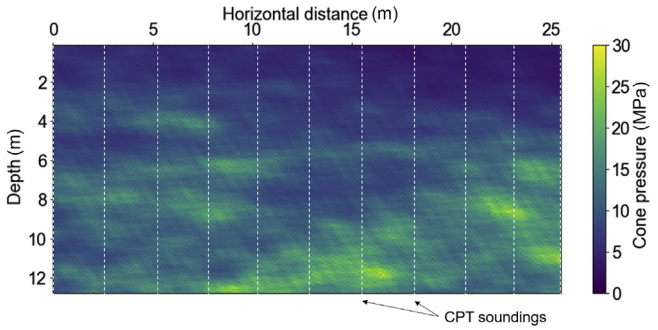

Fig. 4. A numerical example simulated from a non-stationary and non-Gaussian 2D random field.

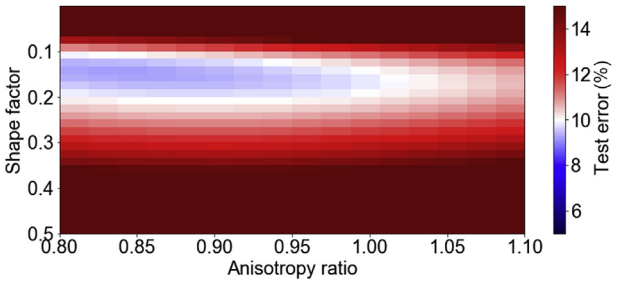

Fig. 5. A typical colormap of test error versus anisotropic ratio and shape factor during grid searching for one simulation.

(iii) Determine the optimal combination of anisotropic ratio and shape factor based on a minimal prediction error for test dataset. Predict the spatially varying geotechnical properties at un-sampled locations using Eqs. (8) and (9).

(iv) Repeat the above steps separately for multiquadric and inverse multiquadric functions until no new result is obtained in the last 3 repetitions. Combine results from all simulations and perform statistical analysis to derive mean and 90% Confidence Interval (CI).

Note that the proposed approach is mathematically simple and can be implemented in a rather straightforward manner by modifying conventional RBFN algorithm in the standard Scipy (Jones et al., 2001). Interpolate package for Python 3.7.

3. Numerical example

In this section, a process of generating a vertical 2D random field for representing cone pressure q_c from CPT soundings in soil is explained. It should be noted only the q_c profile is simulated here as its magnitude is considered representative of the 'point' property of soil without local averaging (Fenton, 1999). The size of the cross-section is 12.8 m in depth and 25.6 m in horizontal distance. A resolution of 0.1 m is used for both directions, resulting in a total of 32,768 (i.e., 128 × 256) data points. Previous statistical analysis carried out by Fenton (1999) indicated that q_c values after logarithmic transformation exhibited a mild trend with depth.

$$lnq_c = a + b \times z + \varepsilon \tag{11}$$

where a and b represents mean value and slope of lnq_c, taken to be 1.5 and 0.1, respectively; z is the depth; ε stands for zero-mean Gaussian random component with a standard deviation of 0.15. Note that the simulated 2D field exhibits a non-linear increasing trend of q_c with depth, which represents a typical profile for a sandy layer. The correlation between ε at any two points (x_i, y_i) and (x_j, y_j) in the random field can be estimated by the following equation (e.g., Shen et al., 2016):

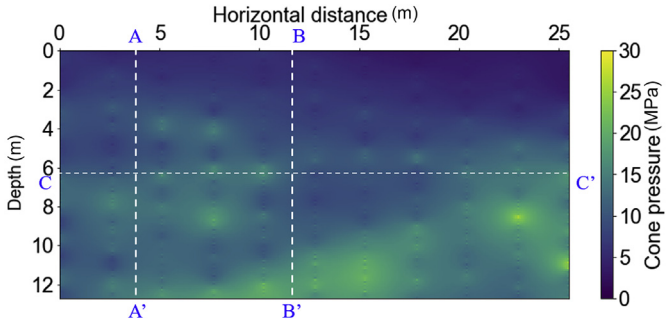

Fig. 6. Colormap of reconstructed 2D field by ensemble RBFN.

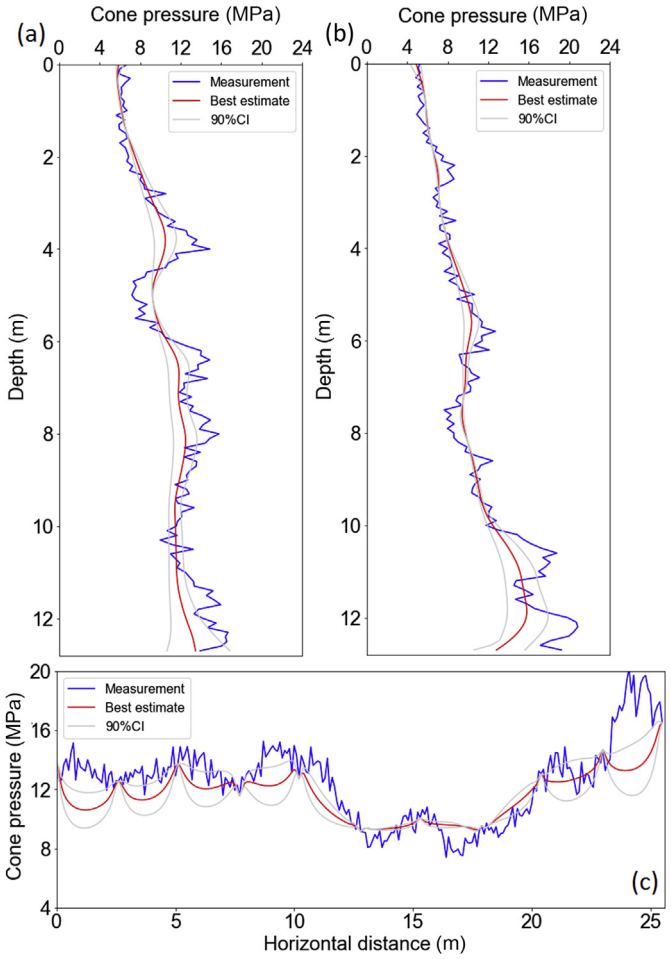

Fig. 7. Comparison between actual and predicted cone pressure profiles by RBFN along three typical planes: (a) A–A′; (b) B–B′; (c) C–C′.

$$\rho = exp\left(-2\sqrt{\frac{(\Delta h)^2}{\lambda_h^2} + \frac{(\Delta v)^2}{\lambda_v^2}}\right) \tag{12}$$

where Δh and Δv denote the relative horizontal and vertical distance between any two points in the field. λ_h and λ_v stand for the horizontal and vertical correlation length, taking to be 15 m and 6 m, respectively. Based on the above parameters, a non-Gaussian and non-stationary 2D random field can be generated and shown in Fig. 4. As an illustration, 11 CPT soundings are taken as sampled profiles. These CPTs are uniformly distributed with an equal horizontal spacing of 2.56 m. In total, there are 1408 (i.e., 11 × 128) sample points, accounting for 4.3% of all the data (i.e., 256 × 128).

4. Results of ensemble radial basis function network

The proposed ensemble RBFN is applied to the example for interpolating spatially varying non-stationary and non-Gaussian geotechnical properties from 11 CPT soundings. Fig. 5 shows a typical colormap of test error with different combinations of anisotropic ratio and shape factor during training process. A minimal test error occurs with an anisotropic ratio of around 0.85, demonstrating the necessity of incorporating anisotropy for spatial interpolation. Fig. 6 shows the colormap of reconstructed 2D cone pressure field by ensemble RBFN. The interpolated field from 11 CPT soundings is quite similar to the original profile in Fig. 4. The percentage difference between the prediction and the actual profile is around 9.4%. Moreover, three representative sections are selected for further comparing prediction performance. Specifically, A–A′ and B–B′ are vertical sections at a horizontal distance of 4.4 m and 12.2 m, respectively, from the origin. C–C′ is a horizontal plane at the mid-depth of 6.4 m. It is worthwhile to point out that the two vertical profiles are farthest away from available CPT soundings and represent the most uninformed locations.

Fig. 7 shows the comparison of best estimate and original q_c profiles along the three selected sections. Note that the blue line represents the actual soil profile and the red line stands for the best estimate from RBFN. 90% confidence interval (CI) deduced from predicted profiles has also been superimposed as grey lines for comparison. Specifically, RBFN prediction and original profile along section A–A′ are shown in Fig. 7a. Essentially, the best estimate follows the trend of the original data. Large deviations occur at the depth between 4 m and 7 m when the original profile exhibits a sudden change. It is worth pointing out that when there is a large deviation, the corresponding 90%CI is also enlarged. While most of the original profile can be well enclosed by the 90%CI. Similar observations are also noted for sections B–B′ and C–C′ in Fig. 7b and c, respectively. As RBFN algorithm is exact interpolation and fully honors measurement points, the predicted and actual profiles along section C–C′ intersect at measurement points, resulting in a caterpillar CI envelope. It is also noted that the 90%CI are rather wide at horizontal distance close to the boundaries (i.e., 0–10 m and 20–25 m) and exhibit a reducing trend when away from boundaries. As discussed in subsection 2.2, estimation error of the ensemble RBFN is primarily originated from the uncertainty of anisotropic ratio and shape factor when only sparse measurements are available. By incorporating ensemble learning, estimation error can be effectively reduced and quantified.

5. Comparative study

With the recent surge and development in machine learning methods, many non-parametric data-driven methods have been developed for spatial interpolation of non-stationary and non-Gaussian data, such as multiple point statistics (MPS) and Bayesian Compressive Sensing (BCS). However, there is no comparative study for benchmarking different non-parametric methods and providing guidelines on selection of the most suitable method for a given non-stationary and non-Gaussian dataset. In this study, both MPS and BCS are applied to the above example and compared with the predictions from ensemble RBFN.

Forecasting error measures can be based on absolute, percentage, symmetric, relative, scaled and other errors (Shcherbakov et al., 2013). Each performance measure evaluates forecasting error from a different aspect. In this study, the absolute difference between predicted and actual cone pressure has direct engineering implication and can be taken as the leading comparison measure. To remove scale dependency, the absolute error is also transformed into absolute percentage error. Therefore, two measures, namely, mean absolute error (MAE) and mean absolute percentage error (MAPE), are adopted for comparison. The detailed mathematical formulations are as follows:

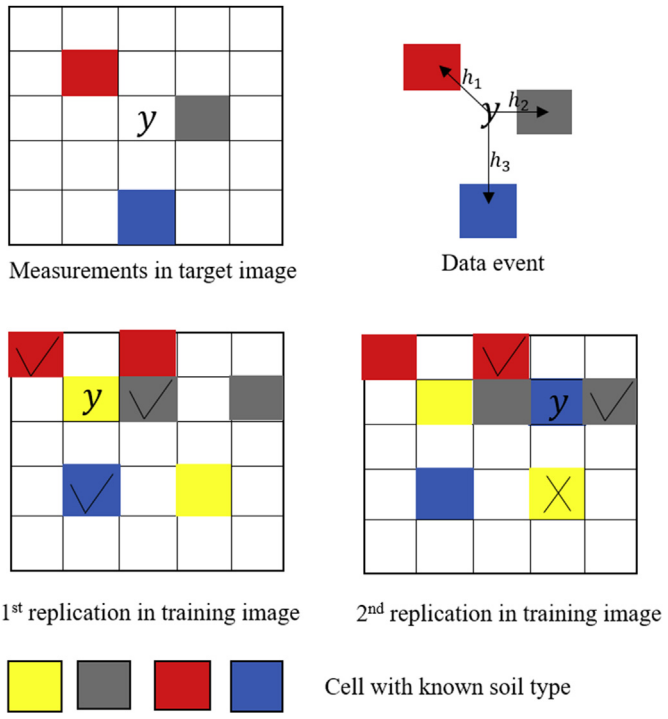

Fig. 8. Illustration of direct sampling algorithm for MPS.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_t - y_t| \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{14}$$

where $\hat{y}_t$ and y_t stand for predicted and actual cone pressure value at t-th point.

6. Results from multiple point statistics

Multiple point statistics (MPS) is a non-parametric and data-driven geostatistical method, which was first proposed by Guardiano and Srivastava (1993). Unlike traditional geostatistical methods, a MPS simulation does not require the explicit specification of a random field and directly infers the empirical multivariate cumulative distribution function from a training image (Strebelle, 2002; Mariethoz et al., 2009). All the potential modelling assumptions have been implicitly embedded within the training image. It is worth pointing out that a training image is normally unavailable for practical geotechnical engineering, in particularly for a small or medium-size project.

MPS has the capability of generating stochastic spatial realizations for both categorical and continuous variables. For continuous variables, Direct Sampling (DS) performs conditional resampling of similar data events from training image and directly determines the value from the closest data event. The above resampling process is illustrated in Fig. 8. The similarity between target data event and potential replicate in the training image is regulated by the weighted Euclidian distance (d) (Mariethoz and Renard, 2010). A smaller value of d implies more resemblance of the target data event to the training event. When the first closest replicate along a random search path in the training image is located, the value at the unknown position is determined. The assigned value is then treated as a known value. The target image is sequentially updated, and the above procedures are repeated until all the unknown locations are filled. A completed target image is also called a realization. Multiple stochastic realizations can be generated by following a random search path in the training image, and the associated uncertainties for

Table 1
Summary of key input parameters for MPS.

Input parameter	Value
Maximum number of counts for conditional probability density function	1
Max number of conditional points, *n*	10
Minimum Euclidean distance, d	0
Shuffle training image path [0: sequential, 1: random]	1
Number of realizations	181[a]

Note.
[a] The number is determined when the last three realizations consecutively show minimal mean error of less than 0.1 kPa from previous simulations.

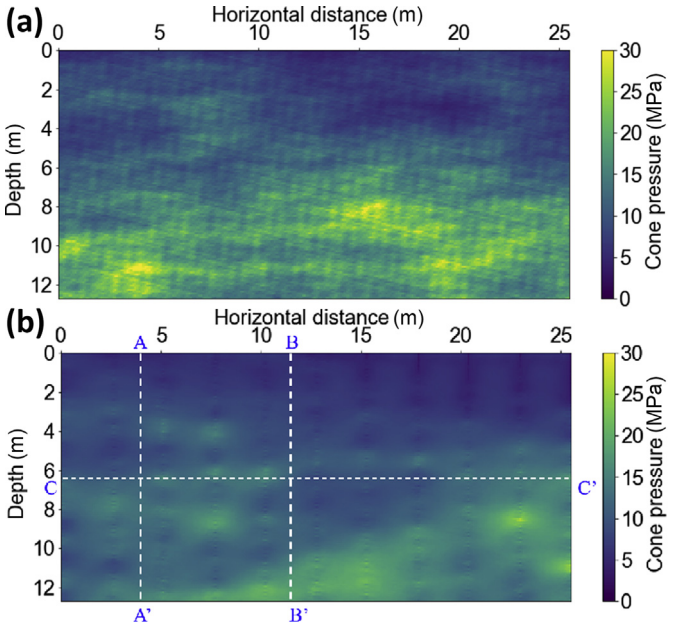

Fig. 9. (a) A training image for MPS. (b) Colormap of reconstructed 2D field by MPS.

spatial interpolation can be quantified via statistical analysis of those realizations.

6.1. Implementation procedure for direction sampling of MPS

The direct sampling algorithm was programmed in C++ program by Hansen and Bach (2016) and can be accessed via both Matlab and Python interfaces. A parameter file containing all the key inputs is required for running the algorithm. Meerschman et al. (2013) provided a practical guidance for the specification of various governing parameters for direct sampling algorithm. All the available measurements should be stored in two data files. One data file containing all the simulation details (i.e., coordinates and q_c values) of the training image should be prepared, the other file lists all the measurement point information (i.e., 11 sampled CPT soundings). The simulation terminates when last three realizations consecutively show that the minimal mean error is less than a small value, i.e., 0.1 kPa, in the previous numerical example. All the key parameters for this study are summarized in Table 1.

6.2. Numerical results

For this study, a training image (see Fig. 9a) is generated from the random field with the same random field parameters (i.e., mean, variance and autocorrelation function) stated in section 3. As both training and test images have been generated following the same process, it is conceivable that multiple replicates of test data events can be found within the training image.

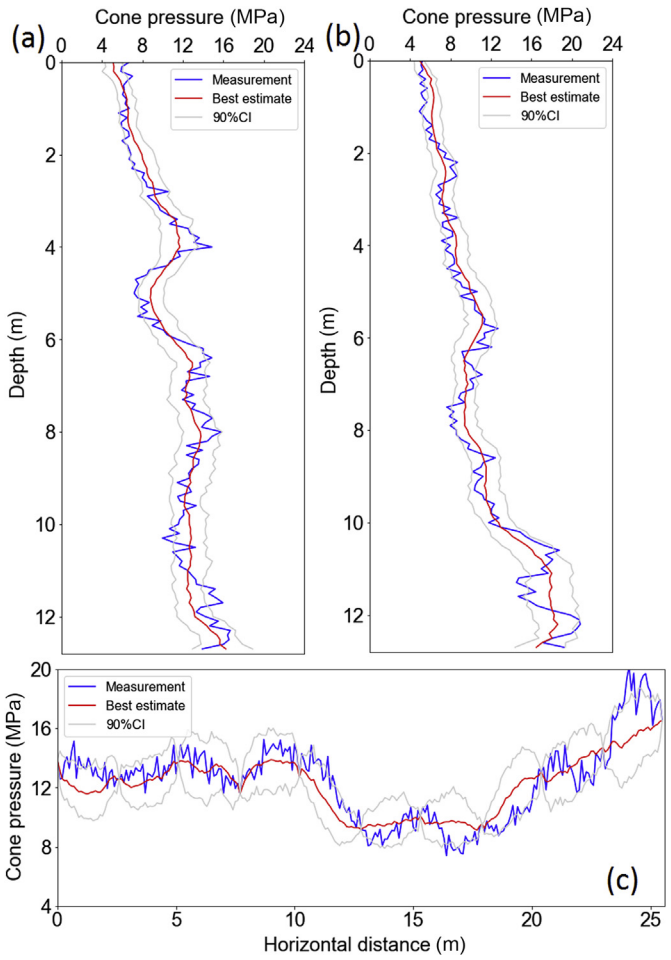

Fig. 10. Comparison between actual and predicted cone pressure profiles by MPS along three typical planes: (a) A–A′; (b) B–B′; (c) C–C′.

Fig. 9b shows the mean q_c plot derived from the 181 stochastic simulations. Essentially, the predicted q_c image is consistent with the test image in Fig. 4 with MAE and MAPE values of 1.0 MPa and 10.4%, respectively. The 1.0 MPa mean difference can be considered small for a field with cone pressure values up to 30 MPa. Both error metrics are quite close to predictions (i.e., 1.1 MPa and 9.4% for MAE and MAPE) from ensemble RBFN. Similarly, three representative sections are selected for further comparing the prediction performance.

Fig. 10 shows the comparison of best estimate and original q_c profiles along the three selected sections. As for section A–A′ in Fig. 10a, the best estimate follows the trend of the original data and the local variations of the original profile can be well enclosed by 90%CI deduced from the 181 simulated profiles. Similar observations are also noted for sections B–B′ and C–C′ in Fig. 10b and c, respectively. MPS algorithm is also an exact interpolation and fully honors measurement points, and the predicted and actual profiles along section C–C′ intersect at measurement points. The upper and lower limits of the predicted envelope exhibit larger deviations from the actual profile at points farther from the measurements (e.g., mid-point between adjacent measurement points). This implies that values at points closed to the measurement points can be estimated with small uncertainty. In fact, the ability to replicate complex spatial connection and variability via multiple simulations is a prominent advantage of MPS (Mariethoz and Caers, 2014). It should be noted that the performance of MPS relies heavily on the quality of the training image. In this example, the training image has been generated using the same set of random field parameters, and the spatial connection and local variation have been retained in both images.

It is also worthwhile to appraise the prediction error from the

Table 2
Implementation procedure for BCS algorithm.

BCS Algorithm	
Step 1	Obtain CPT sounding data and discretize over horizontal and vertical directions
Step 2	Obtain the sizes of 2D field (e.g. 25.6 m × 12.8 m) and specify the resolution (e.g. 0.1 m)
Step 3	Construct orthogonal 2D basis
Step 4	Calculate the non-trivial coefficients, for the 2D basis
Step 5	Output the reconstructed 2D spatial cone pressure profile, and associated statistical uncertainty

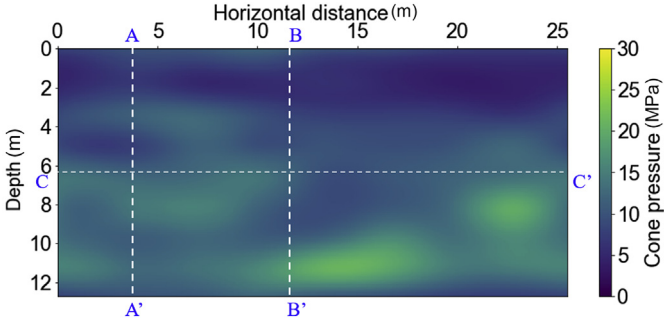

Fig. 11. Colormap of reconstructed 2D field by BCS.

perspective of machine learning. The three errors (i.e., approximation error, sampling error and Bayes error) can also be explicitly identified. Specifically, the approximation error originates from the difference between the training image and the test image and can be effectively reduced once a high-quality training image can be constructed. In fact, the selection of an appropriate training image requires a prior understanding about the variability of the natural process. While the estimation error is mainly contributed by the available number of measurement points as well as the quantity of points for defining a data event. When more points are included in making up a data event, the chance of finding a similar data event in the training image reduces but the accuracy can be guaranteed if such a data event exists. MPS algorithm explicitly quantifies the uncertainty associated with sampling error by generating multiple realizations via randomizing the search grid path and performing statistical analysis (Mariethoz and Caers, 2014). Regarding Bayes error, it is intrinsic to the hard points and cannot be quantified in MPS.

7. Results from Bayesian Compressive Sensing

BCS is another novel non-parametric data-driven method for interpolating spatially varying non-stationary and non-Gaussian geotechnical data from sparse measurements (Wang et al., 2017; Zhao et al., 2018). It is established on the fact that most geotechnical processes exhibit compressibility (e.g., having trends or patterns). This implies that a complete signal can be represented by a weighted summation of a limited number of basic functions (Wang and Zhao, 2016a) and recovered by a remarkably few measurements (Candes et al., 2006; Donoho, 2006). The process of determining the most relevant basis and associated weights is data-driven. More importantly, BCS can explicitly quantify statistical uncertainty of the interpolated profile (Wang and Zhao, 2016b). The BCS method has been successfully applied to deal with different engineering problems, e.g., reconstruction of spatial variables, soil classification as well as simulation of random fields (Zhao et al., 2018; Hu et al., 2019; Wang et al., 2020). The detailed mathematical formulation of BCS for interpolating 2D geodata is referred to Zhao et al. (2018). Only key equations are extracted as below.

The best estimate $\widehat{\boldsymbol{F}}$ of the original signal reconstructed from limited measurements is expressed as follows:

$$\widehat{\boldsymbol{F}} = \sum_{t=1}^{N_h \times N_v} \boldsymbol{B}_t^{2D} \widehat{\omega}_t^{2D} \tag{15}$$

where $\boldsymbol{B}_t^{2D}$ stands for the t-th 2D basis; $\widehat{\omega}_t^{2D}$ denotes the estimated coefficient for the t-th 2D basis; N_h and N_v represent the total number of points for reconstructed $\widehat{\boldsymbol{F}}$ in the horizontal and vertical direction, respectively. The associated mean and variance of the reconstructed signal can be estimated by the following equations:

$$\mu_{\widehat{F}} = E(\widehat{\boldsymbol{F}}) = \sum_{t=1}^{N_h \times N_v} \boldsymbol{B}_t^{2D} \mu_{\widehat{\omega}_t^{2D}} \tag{16}$$

$$Var(\widehat{\boldsymbol{F}} = E\left[(\widehat{\boldsymbol{F}} - \mu_{\widehat{F}})(\widehat{\boldsymbol{F}} - \mu_{\widehat{F}})^T\right] \tag{17}$$

where $\mu_{\widehat{\omega}_t^{2D}}$ represents mean of the estimated coefficient $\widehat{\omega}_t^{2D}$.

7.1. Implementation procedure

The mathematical formulations for 2D BCS appears complicated. However, the implementation and calculation procedures are quite straightforward. Zhao et al. (2018) have coded and packaged the algorithm into a Matlab function. The key input parameters and calculation steps are tabulated in Table 2. The only inputs required for running the program are the measurements (i.e., coordinates and cone pressure values) and dimensions of the reconstructed field (i.e., resolution, horizontal and vertical lengths).

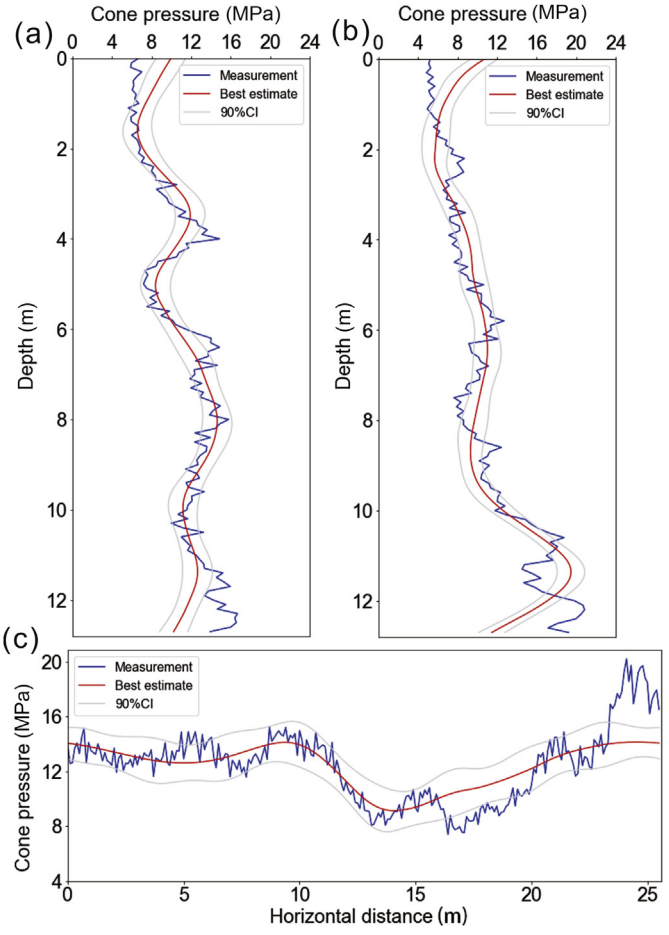

Fig. 12. Comparison between actual and predicted cone pressure profiles by BCS along three typical planes: (a) A–A′; (b) B–B′; (c) C–C′.

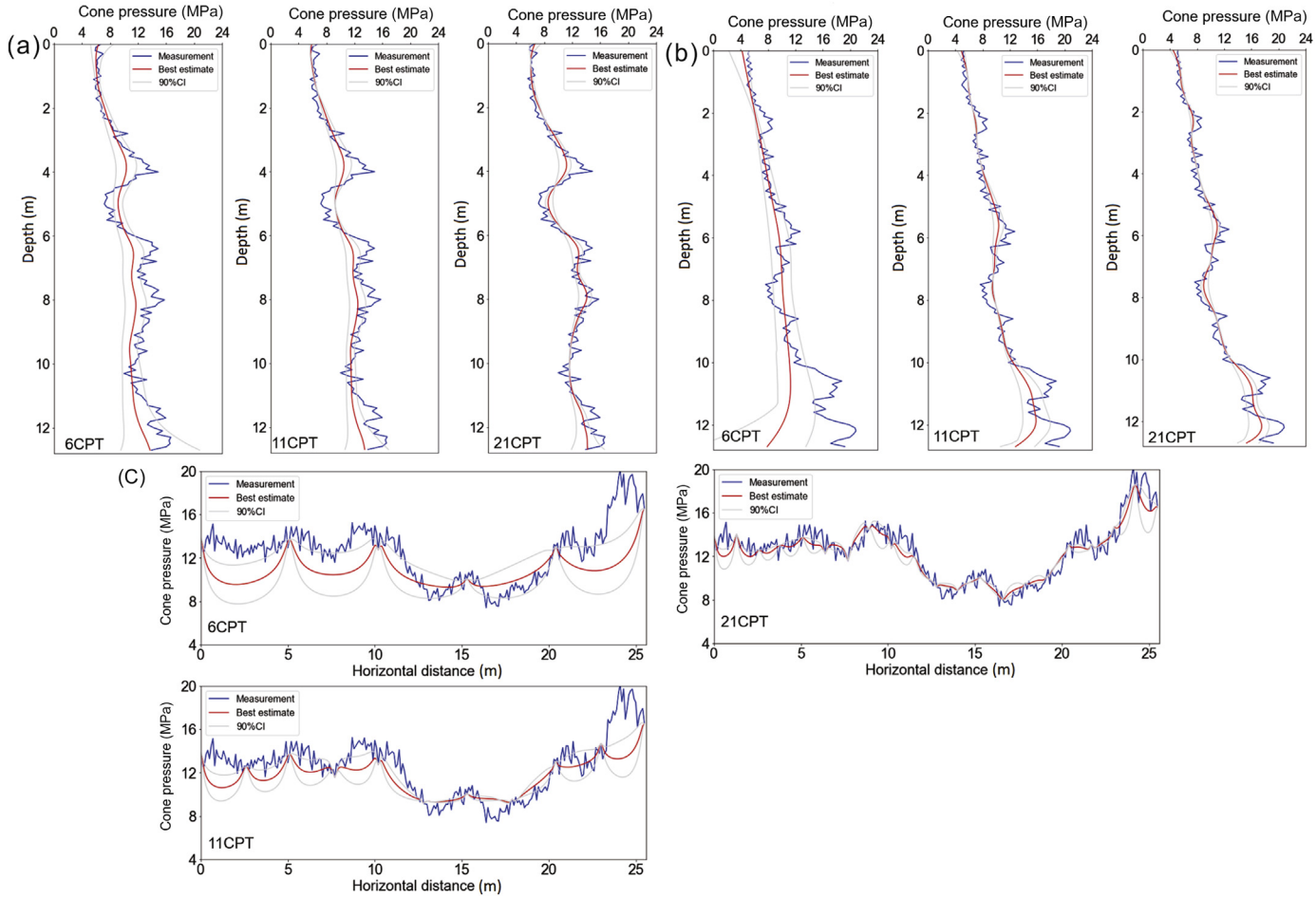

Fig. 13. Effect of data number on ensemble RBFN prediction: (a) A–A′; (b) B–B′; (c) C–C′.

7.2. Numerical results

Fig. 11 shows the reconstructed colormap of q_c by BCS. It is evident that the estimated field well captures the spatial variation of the original profile in Fig. 4. One thing worth noting is that BCS has assumed each sample point embeds with noise. The reconstructed colormap in Fig. 11 reflects the best estimate of the original field. In comparison, preceding MPS and RBFN methods have fully honored information associated with each sample point by assuming that it represents the true soil property without measurement error. The overall MAE and MAPE for the reconstructed field by BCS are 1.5 MPa and 12.9%, respectively, when 11 CPT soundings are available. Three representative profiles are also selected for further evaluation of the prediction capacity of BCS. The best estimate profile (i.e., red line) along A–A′ section (refer to Fig. 12a) well captures the overall trend of the original profile (i.e., blue line). Aside from the best estimate, BCS also explicitly quantifies interpolation uncertainty of the 2D field recovered from the 11 CPT soundings. It is noted that the 90%CI can almost enclose all the local variation of the original A–A′ section. Similar trends are observed along sections B–B′ and C–C′ as shown in Fig. 12b and c, respectively.

The prediction error can be further interpreted from the perspective of machine learning. The approximation error mainly associates with the choice of a 2D basis function (e.g., Fourier transform, wavelet transform and discrete cosine transform). Different basis functions can have different capacities in delineating natural processes of different scales. For instance, Fourier based transform (e.g., discrete cosine transform) are more efficient in exploiting the low frequency nature (Parmar and Scholar, 2014). 2D wavelet basis is adopted in this study with the purpose of characterizing the variation of q_c profile. Regarding estimation error, the amount of measurements determines the number and the coefficient of the 2D basis. To explicitly quantify the uncertainty associated with limited measurements, BCS has associated uncertainty with the coefficient of the 2D basis in the framework of Bayesian analysis. As revealed by Zhao et al. (2018), the sampling error can be effectively reduced when more measurements are available. In addition, the irreducible Bayes error has also been explicitly incorporated in BCS. In other words, the uncertainty estimated by BCS includes both sampling error and Bayes error.

Table 3
Summary of comparison results for different measurement numbers.

Method	MAE (MPa)				MAPE (%)			
	4CPT	6CPT	11CPT	21CPT	4CPT	6CPT	11CPT	21CPT
RBFN	2.7	1.7	1.1	0.6	21.1	14.0	9.4	5.7
MPS	3.1	1.4	1.0	0.7	34.4	15.5	10.4	6.8
BCS	2.0	1.7	1.5	1.5	18.0	14.9	12.9	12.6

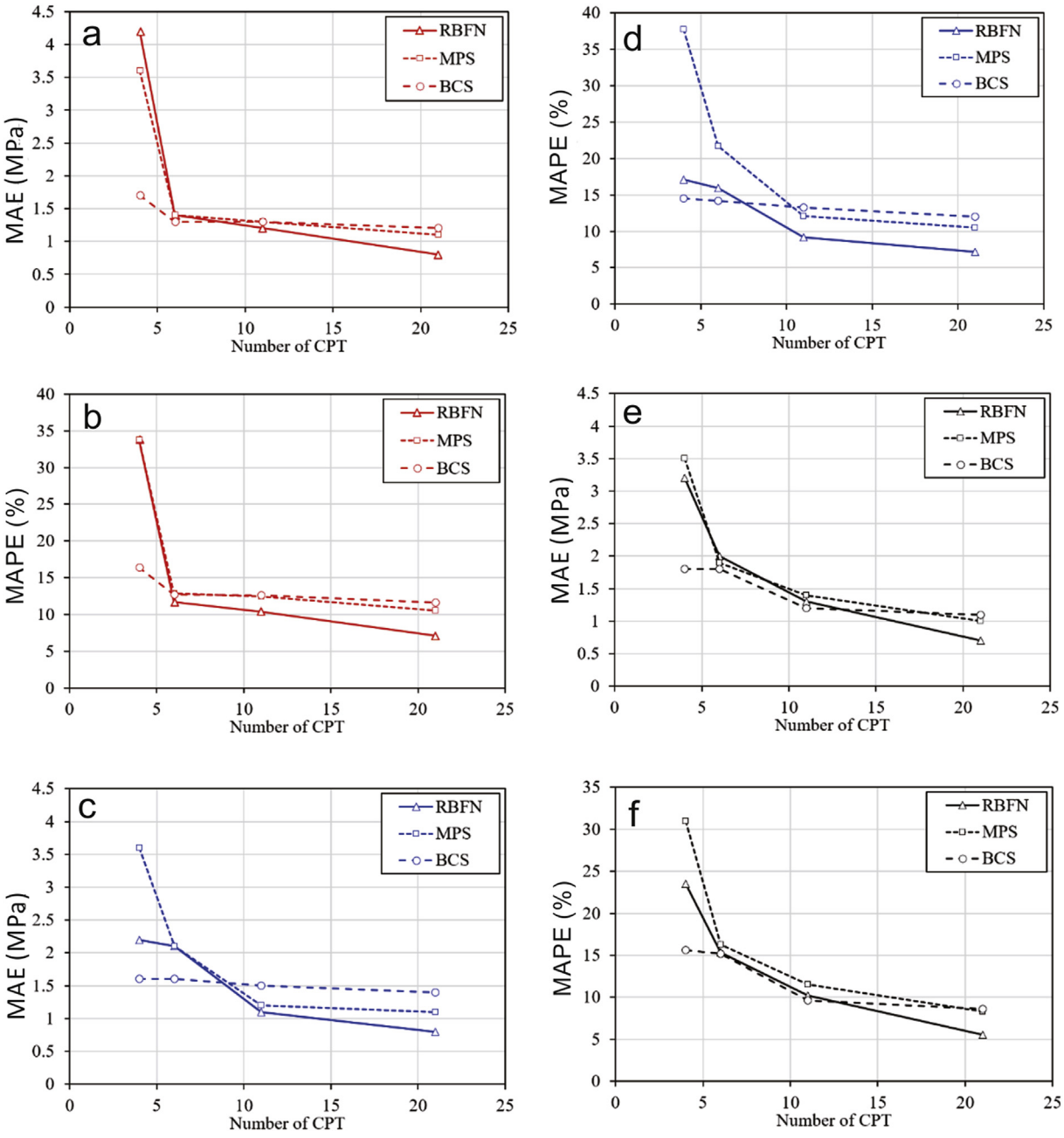

Fig. 14. Variation of MAE and MAPE with the number of CPT soundings: (a) MAE, A–A′ section; (b) MAPE, A–A′ section; (c) MAE, B–B′ section; (d) MAPE, B–B′ section; (e) MAE, C–C′ section; (f) MAPE, C–C′ section.

8. Effect of measurement data number

Intuitively, accuracy of spatial interpolation improves with an increase in measurement data number. In this section, the effect of sample size on quality of the reconstructed soil profiles is investigated. Specifically, three additional scenarios with the number of available CPT sounding to be 4, 6 and 21 are studied. The corresponding horizontal spacing between adjacent CPTs is 8.53 m, 5.12 m and 1.28 m.

Fig. 13 shows the evolution of best estimate and prediction uncertainty from ensemble RBFN with the number of CPT sounding. The comparison between prediction and original profiles along A–A′ section is shown in Fig. 13a. It is evident that as the number of CPT sounding increases from 11 to 21, the best estimate becomes more resemble to the original profile and associated 90%CI shrinks. More importantly, the 90%CI enlarges when the best estimate exhibits a large deviation from the original profile. Similarly, when the available CPT sounding reduces to 6, the best estimate becomes more dissimilar to the original profile, and the 90%CI along the whole profile broadens. Similar trends are also observed for sections B–B′ and C–C′ as shown in Fig. 13b and c. The evolutions of best estimate and associated uncertainty are quite intuitive. As more measurements are available, the governing parameters (i.e., anisotropic and shape factor) can be estimated with less uncertainty, leading to a reduction in estimation errors. Essentially, when the data are measured at all locations, the best estimate and 90%CI converge.

Similarly, effects of measurement data number on the variation of prediction accuracy for MPS and BCS are also investigated, and the results are summarized in Table 3. When the available CPT sounding is more than 11, both RBFN and MPS achieve better prediction accuracy than BCS in terms of MAE and MAPE. It should be noted that the prediction of MPS requires a high-quality training image, which is hardly

available in practice. Comparable performances among three methods are obtained with 6 CPT soundings. However, BCS outperforms ensemble RBFN and MPS when measurements become limited, which is an advantage of BCS. Comparison of prediction capacity from the three methods is also made with respect to the three most uninformed profiles (i.e., A–A′, B–B′ and C–C′ sections). As for section A–A′ (see Fig. 14a and b), RBFN shows a superior performance over MPS and BCS when more than 11CPT soundings are available. MAE and MAPE of RBFN reduce by 0.3 MPa and 3.3%, respectively, when CPT number increases from 11 to 21. Both error metrics increase when only 4 CPT soundings are available. Similar trends are observed for MPS prediction. In contrast, BCS predictions are more stable and MAE and MAPE increase from 1.2 MPa to 1.7 MPa and 11.6%–16.4%, respectively. Similar observations are obtained for sections B–B′ and C–C′. When number of CPT sounding increases, the prediction of all the three methods improves and the reconstructed profiles become more similar to the original profile. Conversely, when the available CPT sounding reduces, the performance of all the three algorithms deteriorates.

9. Summary and conclusion

Radial Basis Function Network (RBFN) is a popular machine learning method for spatial interpolation of non-stationary and non-Gaussian data. However, conventional radial basis functions are direction independent, and its prediction uncertainty cannot be quantified. In this study, the above two problems are explicitly tackled by employing Mahalanobis distance to account for spatial anisotropy and mobilizing ensemble learning to quantify prediction uncertainty. The improved RBFN is illustrated using a set of CPT data. The performance of ensemble RBFN is also compared with two other non-parametric data-driven methods, namely, Multiple Point Statistics (MPS) and Bayesian Compressive Sensing (BCS), in reconstructing cone pressure profiles from limited CPT soundings. Moreover, evolutions of best estimate and associated interpolation uncertainty of the reconstructed profiles with different measurement data numbers (i.e., 4, 6, 11 and 21 CPT soundings) are explicitly quantified and compared. Based on this study, the following conclusions may be drawn.

(1) The prediction accuracy of RBFN for reconstructing spatial distributions improves after incorporating Mahalanobis distance metric. In addition, the proposed ensemble RBFN method integrates results from both multiquadric and inverse multiquadric radial basis functions, and the prediction uncertainty can be quantified. The resulted 90% Confidence Interval (CI) reasonably encloses the original cone pressure profiles, and it effectively shrinks and enlarges with an increase and reduction in the measurement data number.
(2) For the 2D non-stationary and non-Gaussian CPT data, the ensemble RBFN and MPS outperform BCS in reconstructing CPT profiles when more than 11 CPT soundings are available. It is worth pointing out the application of MPS algorithm requires a high-quality training image for characterizing spatial variation, which is normally unavailable for practical site investigation. When the number of CPT sounding reduces to 4, performance of both RBFN and MPS deteriorates. In contrast, BCS prediction is less sensitive to the number of CPT soundings and performs best when CPT sounding becomes limited.
(3) The prediction error associated with machine learning practice can be decomposed into three components, namely, approximation error, estimation error and Bayes error. A prior understanding of the subsurface heterogeneity can help determine model complexity and minimize potential approximation error. The proposed ensemble RBFN enables the quantification of estimation error, which originates from model uncertainty (i.e., multiquadric and inverse multiquadric) and parameter (i.e., anisotropic ratio and shape factor) uncertainty when only spare measurements are available. While MPS relies heavily on training image to quantify approximation error and estimation error. In comparison, BCS can explicitly assess estimation error and Bayes error.

Declaration of competing 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.

Acknowledgements

The work described in this paper was supported by grants from the Research Grants Council of Hong Kong Special Administrative Region, China (Project No. CityU 11213119 and T22-603/15N). The financial support is gratefully acknowledged. The first author also acknowledges financial support from the Hong Kong Ph.D. Fellowship Scheme funded by the Research Grants Council of Hong Kong, China.

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Statistics 14, 91–113.
Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231.
Candes, E.J., Romberg, J.K., Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math.: A Journal Issued by the Courant Institute of Mathematical Sciences 59 (8), 1207–1223.
Carlson, C.A., 1991. Spatial distribution of ore deposits. Geology 19 (2), 111–114.
Cucker, F., Zhou, D.X., 2007. Learning Theory: an Approximation Theory Viewpoint. Cambridge University Press.
Devendra, K., 2008. Soft Computing: Techniques and its Applications in Electrical Engineering. Springer, Berlin, Germany.
Domingos, P.M., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.
Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inf. Theor. 52 (4), 1289–1306.
Du Toit, W., 2008. Radial Basis Function Interpolation. M.Sc. thesis. Stellenbosch University.
Ersoy, A., Yunsel, T., Cetin, M., 2004. Characterization of land contaminated by past heavy metal mining using geostatistical methods. Arch. Environ. Contam. Toxicol. 46 (2), 162–175.
Fenton, G.A., 1999. Random field modeling of CPT data. J. Geotech. Geoenviron. Eng. 125 (6), 486–498.
Guardiano, F.B., Srivastava, R.M., 1993. Multivariate geostatistics: beyond bivariate moments. In: Geostatistics Troia'92. Springer, pp. 133–144.
Hansen, T.M., Bach, T., 2016. MPSLIB: A C class for sequential simulation of multiple-point statistical models. Software 5, 127–133.
Hardy, R., 1971. Multiquadric equations of topography and other irregular surfaces. J. Geophys. Res. 76 (8), 1905–1915.
Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. Math. Intel. 27 (2), 83–85.
Hu, Y., Zhao, T., Wang, Y., Choi, C., Ng, C.W., 2019. Direct simulation of two-dimensional isotropic or anisotropic random field from sparse measurement using Bayesian compressive sampling. Stoch. Environ. Res. Risk Assess. 33 (8–9), 1477–1496.
Jones, E., Oliphant, T., Peterson, P., 2001. SciPy: Open Source Scientific Tools for Python. Retrieved from. https://www.scipy.org/about.html.
Li, J., Heap, A.D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record 2008/23, 137 pp.
Li, C., Wang, F.L., Chang, Y.Q., Liu, Y., 2010. A modified global optimization method based on surrogate model and its application in packing profile optimization of injection molding process. Int. J. Adv. Manuf. Technol. 48 (5–8), 505–511.
Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Software 26 (12), 1647–1659.
Li, J., 2019. A critical review of spatial predictive modeling process in environmental sciences with reproducible examples in R. Appl. Sci. 9 (10), 2048.
Lin, G., Chen, L., 2004. A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. J. Hydrol. 288 (3–4), 288–298.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. In: Proceedings of the National Institute of Science of India, 2, pp. 49–55.
Mariethoz, G., Renard, P., 2010. Reconstruction of incomplete data sets or images using direct sampling. Math. Geosci. 42 (3), 245–268.
Mariethoz, G., Caers, J., 2014. Multiple-point Geostatistics: Stochastic Modeling with Training Images. John Wiley & Sons.
Mariethoz, G., Renard, P., Froidevaux, R., 2009. Integrating collocated auxiliary parameters in geostatistical simulations using joint probability distributions and probability aggregation. Water Resour. Res. 45 (8), W08421.

Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M., Renard, P., 2013. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. Comput. Geosci. 52, 307–324.
Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113, 56–69.
Park, J., Sandberg, I.W., 1991. Universal approximation using radial-basis-function networks. Neural Comput. 3 (2), 246–257.
Parmar, H.M., Scholar, P., 2014. Comparison of DCT and wavelet based image compression techniques. Int. J. Exp. Diabetes Res. 2, 664–669.
Powell, M.J., 1987. Radial Basis Functions for Multivariable Interpolation: a Review. Algorithms for Approximation, 143–167.
Risser, M.D., 2016. Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate-Driven Approaches. arXiv: 1610.02447v1.
Rusu, C., Rusu, V., 2006. Radial basis functions versus geostatistics in spatial interpolations. In: IFIP International Conference on Artificial Intelligence in Theory and Practice. Springer, pp. 119–128.
Shcherbakov, M.V., Brebels, A., Shcherbakova, N.L., Tyukov, A.P., Janovsky, T.A., Kamaev, V.A., 2013. A survey of forecast error measures. World Appl. Sci. J. 24 (24), 171–176.
Shen, P., Zhang, L.M., Zhu, H., 2016. Rainfall infiltration in a landslide soil deposit: importance of inverse particle segregation. Eng. Geol. 205, 116–132.
Strebelle, S., 2002. Sequential Simulation Drawing Structures from Training Images. Ph. D. thesis. Stanford University, p. 374.
Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190 (3–4), 214–251.
Wang, Y., Zhao, T., 2016a. Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. Geotechnique 67 (6), 523–536.
Wang, Y., Zhao, T., 2016b. Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. Can. Geotech. J. 53 (9), 1547–1559.
Wang, Y., Zhao, T., Phoon, K., 2017. Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. Can. Geotech. J. 55 (6), 862–880.
Wang, Y., Hu, Y., Zhao, T., 2020. CPT-based subsurface soil classification and zonation in a 2D vertical cross-section using Bayesian compressive sampling. Can. Geotech. J. 57 (7), 947–958. https://doi.org/10.1139/cgj-2019-0131.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons.
Xu, Z., Weinberger, K.Q., Chapelle, O., 2012. Distance Metric Learning for Kernel Machines. arXiv: 1208.3422v2.
Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications. Springer Science & Business Media.
Zhang, W., Goh, A.T.C., 2014. Multivariate adaptive regression splines model for reliability assessment of serviceability limit state of twin caverns. Geomechanics and Engineering 7 (4), 431–458.
Zhang, W., Wu, C., Li, Y., Wang, L., Samui, P., 2019. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. Georisk 1–14.
Zhang, W., Goh, A.T., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci. Front. 7 (1), 45–52.
Zhang, X., Yue-Yu, S., Zhang, X., Kai, M., Herbert, S., 2007. Spatial variability of nutrient properties in black soil of northeast China. Pedosphere 17 (1), 19–29.
Zhao, T., Hu, Y., Wang, Y., 2018. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. Eng. Geol. 246, 162–175.
Zhou, Z.H., 2012. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC.