The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications (APSSRA2020)
October 4–7 2020, Tokyo, Japan
T. Takada, I. Yoshida & T. Itoi (editors)

Comparative study of data-driven machine learning methods for spatially interpolating non-stationary and non-Gaussian geotechnical properties

Chao Shi[1] and Yu Wang[2]

[1] *Department of Architecture and Civil Engineering, City University of Hong Kong (Email): chaoshi6-c@my.cityu.edu.hk*
[2] *Department of Architecture and Civil Engineering, City University of Hong Kong (Email): yuwang@cityu.edu.hk*

Abstract: A well understanding of subsurface heterogeneities is beneficial for risk assessment and decision making in geotechnical engineering practice. It is conventional to use geostatistics to estimate heterogeneous geotechnical properties at un-sampled locations. The successful application of traditional geostatistical models relies heavily on stationarity assumption to derive spatial auto-correlation functions, e.g. semivariogram. The conversion of geotechnical measurements, which are normally non-stationary and non-Gaussian, into stationary processes is a highly non-trivial task for engineering practitioners, particularly when only sparse measurements are available. Data-driven machine learning methods, e.g., radial basis function network (RBFN), are promising for spatial interpolation as they are non-parametric and can adaptively determine the optimal relationship between input and output. In this study, three data-driven machine learning algorithms, namely ensemble RBFN, Multiple Point Statistics (MPS) and Bayesian Compressive Sensing (BCS), are introduced and compared for spatial interpolation. The three approaches can provide best estimate as well as quantify prediction uncertainty of geotechnical properties at locations of interest. The performance of all the three methods is illustrated using a simulated example of cone penetration test (CPT) data. The results indicate that the ensemble RBFN can better predict the best estimate and associated uncertainty when a reasonable amount of measurements are available. Moreover, BCS algorithm is demonstrated to be robust and insensitive to measurement data number, exhibiting a superior performance over RBFN and MPS when only limited measurements are available.

Keywords: spatial interpolation, multiple point statistics, Bayesian compressive sampling, compressive sensing, sparse measurement

1. Introduction

It is an essential task to interpolate spatially varying field attributes from scatter measurements for geotechnical engineers. An accurate interpolation of geotechnical properties plays a key role in planning, risk assessment and decision making. Geostatistics has been a powerful tool for assessing heterogeneity and spatial variability. For instance, kriging is a popular geostatistics model and can provide both best estimate and interpolation uncertainty. However, the successful application of kriging is limited to stationary field (Webster and Oliver, 2007) and requires prior evaluation of a site- and data-specific autocorrelation structure (e.g., semivariogram) (Oliver and Webster, 2014) between spatial measurements. The accurate specification of parametric function forms and associated parameters (i.e., still, range and nugget) requires a large amount of measurements, which are normally unavailable in practical geotechnical site characterization.

It is well acknowledged that geotechnical measurements (e.g., cone pressure and undrained shear strength) are non-stationary and non-Gaussian in nature, which impedes the application of conventional parametric statistical models (e.g., kriging). Emerging machine learning methods provide alternative spatial interpolators for dealing with those non-stationary, non-Gaussian and non-heteroscedastic geotechnical processes (Li and Heap, 2008; Li et al., 2011). Machine learning is a branch of soft computing techniques, which solve practical problems by progressively and adaptively exploiting imprecision, uncertainty and partial truth (Devendra, 2008). The prominent advantages of machine learning approaches over conventional geostatistical models and other deterministic methods (e.g., inverse distance weighing) are data-driven and less assumption dependent (e.g., specification of certain parametric function forms are not needed for machine learning methods).

Of all the machine learning methods, network based models are appealing to engineers as they can adaptively exploit complex non-linear relationship between measurements. Shi and Wang (2020) developed an ensemble Radial Basis Function Network (RBFN) to account for spatial anisotropy and quantify prediction uncertainty for spatial interpolation in geotechnical site characterization. Other popular non-parametric data-driven approaches including Multiple Point Statistics (MPS) (e.g., Mariethoz and Caers, 2014) and Bayesian Compressive Sensing (BCS) (Wang and Zhao, 2016). In this study, a comparative study is performed to benchmark the above three data-driven approaches in spatial interpolation of non-stationary and non-Gaussian geotechnical processes. The accuracy of best estimate and uncertainty quantification are explicitly compared.

The reminder of this study is organized as follows. In the second section, a numerical example of a 2D non-stationary and non-Gaussian random field is simulated. Rationales behind the three non-parametric methods (i.e., RBFN, MPS and BCS) are briefed and implementation procedures are detailed in the third section. Comparison of the three data-driven methods in interpolating spatially varying geotechnical properties is discussed in the fourth and fifth sections. Subsequently, effects of measurement data number on the reconstructed fields by different methods are investigated. Finally, conclusions on capacity of different models in estimating spatially varying geotechnical properties are drawn.

2. Simulation of 2D random field

In this section, a vertical 2D random field representing cone pressure q_c from Cone Penetration Test (CPT) is simulated. According to Fenton (1999), cone pressure is considered the most representative of 'point' property of soil without local averaging. The q_c values after logarithmic transformation are assumed to increase linearly with depth (Fenton, 1999).

$$ln(q_c) = a + b \times z + \varepsilon \tag{1}$$

where a and b are intercept and coefficient for the linear equation and taken as 1.5 and 0.1, respectively; ε is the residual term and assumed to follow Gaussian distribution with mean and standard deviation of 0 and 0.15. The correlation ρ of ε between any two points in space is modeled as an exponential function (e.g., Shen et al., 2016).

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

where Δ_h and Δ_v are relative distance in the horizontal and vertical direction between any points in space; λ_h and λ_h are correlation lengths in the horizontal and vertical directions and taken to be 15m and 6m, respectively.

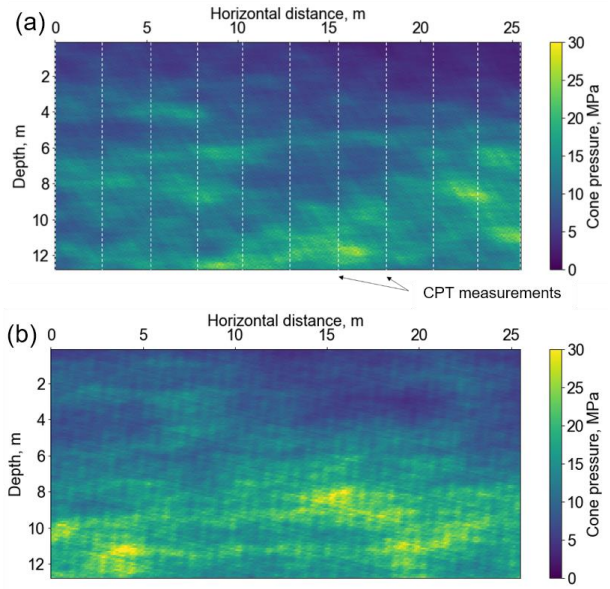

Fig.1 Simulated 2D random field: (a) Simulation image; (b) Training image for MPS algorithm

The total size of the simulated 2D random field is 12.8m in depth and 25.6m in horizontal length. The resolution in both directions is taken as 0.1m, resulting in a total of 32768 (i.e., 128×256) points. Fig.1a shows the colormap of the simulated image. For illustration, only 11 line measurements with equal horizontal separation of 2.56m are taken as measurements. A training image (refer to Fig.1b) follows the same set of random field parameters is also generated to facilitate the calculation of MPS algorithm. More discussion can refer to section 3.2.

3. Comparative study

3.1 Ensemble radial basis function network

Radial basis function network, i.e., RBFN, has been a popular method for solving multivariate interpolation problems. Mathematically speaking, the interpolant, *y(x)*, at an un-sample location is calculated as a weighted summation of basis functions at discrete points, i.e., x_i, i=1, 2…n.

$$y(x) = \sum_{i=1}^{n} \omega_i \varphi(\| x - x_i \|), x \in R^m \tag{3}$$

where ‖•‖ calculates the Euclidean distance between two points x and x_i. ψ is the radial basis function, whose value solely depends absolute radial distance to a central point, x_i. Any functions satisfy the above property can be called a radial basis function. Conventional radial basis functions include multiquadratic and inverse multiquadratic.

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

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

where σ is shape factor. Fig.2 illustrate the distribution of multiquadric and inverse multiquadric functions with raidal distance at different shape factors.

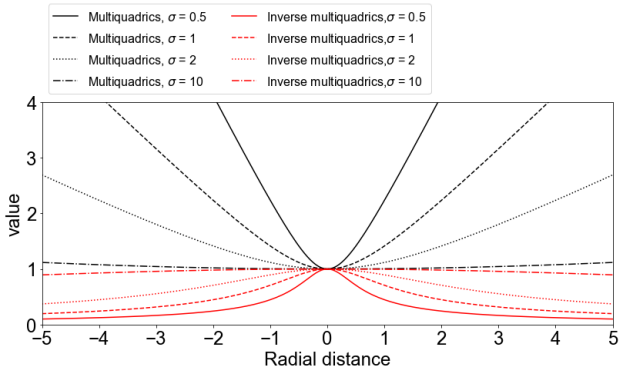

Fig.2 Illustration of commonly used radial basis function with different shape factors

It is worth pointing out that conventional geotechnical properties are normally depositional and exhibit strong horizontal patterns. Therefore, it is imperative to take spatial anisotropy into consideration. Another difficulty associated with conventional RBFN is the quantification of interpolation uncertainty.

In order to explicitly overcome the above limitations of conventional RBFN. An ensemble RBFN was proposed by Shi and Wang (2020). Spatial anisotropy was accounted for by implementing a Generalized Euclidean distance (Mahalanobis, 1936) for distance calculation in Eq.(6).

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

where $\boldsymbol{M}$ is 2 by 2 diagonal matrix for 2D problem in this study.

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

where a is the anisotropic ratio.

In addition, interpolation uncertainty is quantified by employing multiquadric and inverse multiquadric functions within an ensemble learning framework. Both multiquadric and inverse multiquadric functions are adopted based on the consideration that the two functions can fill up the whole space at any points by changing shape factor σ (see Fig.2). The interpolation results from both radial basis functions are then stacked with equal weights for deriving the final mean and prediction uncertainty. Eq.(6) is then integrated with Eqs.(4) and (5), yielding the following modified radial basis functions.

Multiquadric

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

Inverse multiquadric

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

It should be noted that there are two unknown parameters, i.e., anisotropic ratio a and shape factor σ, in above Eqs.(8) and (9), which are determined using training data. Fig. 3 shows implementation procedure for ensemble RBFN.

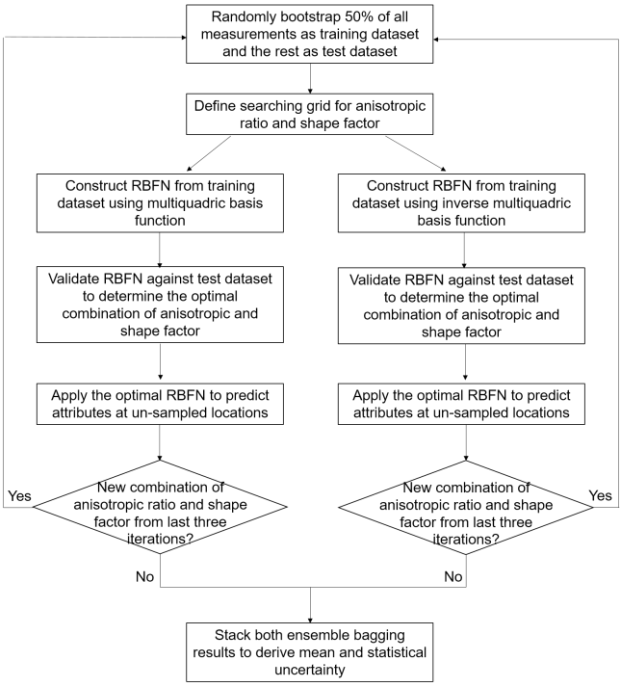

Fig.3 Flowchart of ensemble RBFN algorithm (Shi and Wang, 2020)

The flowchart in Fig.3 illustrates procedures of the ensemble RBFN algorithm to interpolate spatially varying geotechnical properties. For clear references, only key steps are summarized below.

1. Collect all measurement data (i.e., spatial coordinates and cone pressure value) and divide into training and test dataset with a split ratio of 50:50. Define all possible ranges of anisotropic ratio a and shape factor σ.
2. Calculate Mahalanobis distance between training dataset and spatial coordinates of test dataset using Eq.(6), and substitute into Eqs.(8) and (9) to derive corresponding interpolant.
3. Determine the optimal combination of anisotropic ratio and shape factor based on the minimization of error between predicted and actual values for the test dataset, and apply the optimal parameters to interpolate spatially varying cone pressure at un-sampled locations.
4. Repeat steps 1 to 3 separately for both radial basis functions until no new results are obtained from the last consecutive repetitions, and combine all the simulation results at un-sampled locations to derive mean and 90% Confidence Interval (CI).

The above procedures can be easily implemented using the standard Scipy in Python 3.7 (Jones et al., 2001).

3.2 Multiple point statistics

Multiple point statistics (MPS) is a well-developed non-parametric spatial interpolation method in geoscience community. MPS was first proposed by Guardiano and Srivastava (1993) to move beyond conventional two-point based indicator variogram to multiple point statistics. Rather than rely on assumed parametric function forms for spatial interpolation, MPS directly infers higher-order statistics from a training image (Mariethoz and Caers, 2014; Strebelle, 2002) based on the assumption that both the training image and the underlying simulation image share the same multiple-point covariance relationship. It should be noted that in practical geotechnical engineering, the complete training image is normally unavailable, particularly for a small or medium-size project.

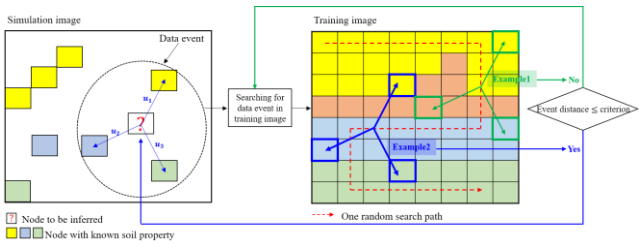

Fig.4 Illustration of direct sampling algorithm

MPS can be used for spatial interpolation of both categorical variable and continuous variable. Specifically for continuous variable, Direct Sampling (DS) was proposed by Mariethoz and Renard (2010) to perform conditional resampling of similar data events from training image. The conditional simulation process of direct sampling is illustrated in Fig.4. Each cell is a point

measurement. A target data event is defined as n closest known point values, e.g., the three colored cells (n=3) within the dashed circle as shown in Fig.4, around the unknown cell. Similar data events are searched within the training image and the similarity between target data event and potential replicates in the training image, e.g., examples 1 and 2, is regulated by a weighted Euclidean distance (d) (Mariethoz and Renard 2010). A smaller d value implies more resemblance of the searched example to the target data event. When the first replicate (e.g., example 2) with d less or equal to a specified value is located, the value of the unknown cell can be determined by directly assigning the central value of the replicate. The interpolated value is then treated as a known measurement. The above procedure is repeated until all the unknown cells within the simulation image are interpolated, finishing a realization. Multiple realizations are generated following random search path with training image and random simulation path in simulation image. The best estimate and associated interpolation uncertainty can be obtained by statistical analysis of those multiple realizations.

Hanson and Bach (2016) coded the direct sampling in C++ program, which can be accessed by Python and Matlab interfaces. Two text files are required for running direction sampling. One file should contain all the information for the training image, including spatial coordinates and q_c values. The other file documents all the available measurement (i.e., 11 line profiles) for the simulation image. In addition, hyper-parameters (e.g., distance criterion, number of conditioning points) for the algorithm should also be specified. Meerschman et al. (2013) provided a practical guidance on performing stochastic simulation with direct sampling algorithm. Table 1 lists the key hyper-parameters for geotechnical application of direct sampling. Multiple realizations are generated until the mean q_c value of the last consecutive simulations show an average difference of less than 0.1kPa.

Table 1. Input parameters for MPS algorithm

Input parameter	Value
Maximum number of counts for conditional probability density function	1
Maximum number of conditional point, n	10
Minimum Euclidean distance, d	0
Shuffle training path [0: sequential, 1: random]	1
Number of realization	181

3.3 Bayesian compressive sensing

Compressive Sensing (CS) was originally developed in electrical engineering in order to compress and recover signals. Wang et al. (2017) and Zhao et al. (2018) integrated CS algorithm within the Bayesian framework for spatially interpolating non-stationary and non-Gaussian geotechnical properties from sparse measurements. The fundamental assumption behind the Bayesian Compressive Sensing (BCS) is that most geotechnical processes are compressive (e.g., having trends or patterns). Therefore, the complete geotechnical process or field can be represented as a weighted summation of a limited number of basis functions (e.g., wavelet or discrete cosine functions) and recovered by remarkably few measurements. Mathematically speaking, the original geotechnical process $\hat{\boldsymbol{F}}$ recovered from limited measurements is formulated as follows:

$$\hat{\boldsymbol{F}} = \sum_{t=1}^{N_h \times N_v} \boldsymbol{B}_t^{2D} \hat{\omega}_t^{2D} \tag{9}$$

where $\boldsymbol{B}_t^{2D}$ is the t-th 2D basis function; ω_t^{2D} is coefficients associated with $\boldsymbol{B}_t^{2D}$; N_h and N_v are the total number of points in the horizontal and vertical directions.

The determination of coefficients for those basis functions are purely data-driven. The derived coefficients and associated basis functions can be combined to give best estimate and quantify interpolation uncertainties of interpolated profiles. The complete mathematical derivations of BCS are detailed by Zhao et al. (2018). Only key equations are reported here. The best estimate and variance of the reconstructed geotechnical profile are expressed as follows:

$$\mu_{\hat{F}} = E(\hat{\boldsymbol{F}}) = \sum_{t=1}^{N_h \times N_v} \boldsymbol{B}_t^{2D} \mu_{\hat{\omega}_t^{2D}} \tag{10}$$

$$Var(\hat{\boldsymbol{F}} = E\left[(\hat{\boldsymbol{F}} - \mu_{\hat{F}})(\hat{\boldsymbol{F}} - \mu_{\hat{F}})^T\right] \tag{11}$$

where $\mu_{\hat{F}}$ represents mean of the estimated geotechnical process $\hat{\boldsymbol{F}}$. BCS algorithm has been successfully applied to interpolation and simulation of 2D random fields (Hu et al., 2019; Wang et al., 2019; Zhao et al., 2018) and non-stationary and non-Gaussian Random fields (Wang et al. 2019; Montoya-Noguera et al. 2019).

Although the mathematical formation of 2D BCS looks complicated, the implementation and simulation are quite straightforward. A packaged Matlab function is available in Zhao et al. (2018). The key simulation steps are summarized in Table 2.

Table 2. Implementation procedure for BCS algorithm

Step	
1	Collect all measurements and discretize the whole domain
2	Specify the resolution (e.g., 0.1m) and calculate total size of 2D field (e.g., 25.6m ×12.8m)
3	Construct 2D orthogonal wavelet basis
4	Calculate non-zero coefficients for the wavelet basis
5	Reconstruct 2D cone pressure field and derive associated prediction uncertainty

4. Performance measure

In this study, two major measures, namely Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE), are used to compare interpolation results of different methods. The formulations for MAPE and MAE calculation are as follows.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_t - y_t| \tag{12}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \quad (13)$$

where $\hat{y}_t$ and y_t are predicted and actual value at t-th point.

5. Numerical results of different methods

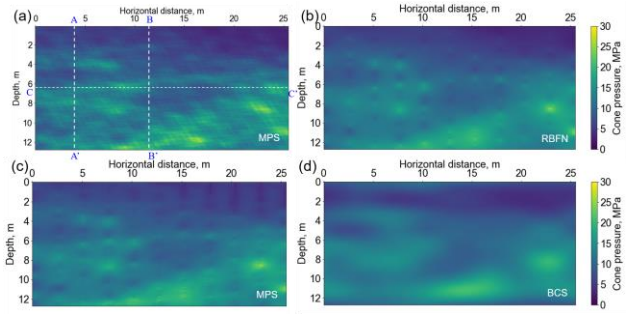

Fig.5 (a) True test image; (b) Interpolation results of RBFN; (c) Interpolation results of MPS; (d) Interpolation results of BCS

Fig.5 shows colormaps of underlying true test image and interpolation results from the above three methods. It is clear that all the three reconstructed colormaps conditional on 11 CPT soundings can essentially recover the spatial patterns of the test cone pressure field. The calculated MAPE for the three methods (i.e., RBFN, MPS and BCS) ranges between 9.4% and 12.9%. The corresponding MAE values vary between 1.0MPa and 1.5MPa. For a cone pressure field with a maximum value of about 30MPa, the differences in MAE are considered negligible. To further compare the interpolation performance, vertical profiles along A-A' section are also extracted and compared.

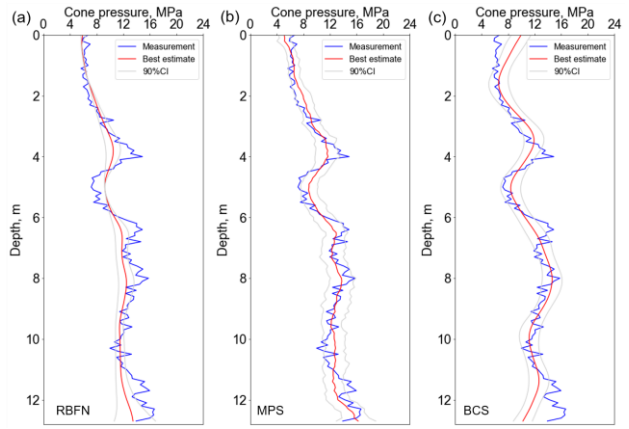

Fig.6 Comparison of interpolation results along A-A' section: (a) RBFN; (b) MPS; (c) BCS

Fig.6a shows the comparison between the true measurement and the best estimate from ensemble RBFN. The 90% confidence interval (CI) from RBFN simulations are also superimposed for better comparison. Essentially, the variation of best estimate along depth follows that of true measurements. In addition, CI enlarges whenever there is a large variation of the true measurements. Similar interpolation results obtained from MPS and BCS are shown in Figs.6b and 6c. Clearly, the above three methods can not only provide the best estimate, but also explicitly quantify interpolation uncertainty.

6. Effect of measurement data number

Intuitively, the interpolation performance of different methods improves as more measurements are used. Fig.7 compares the prediction performance of three different methods at different measurement numbers. The calculated MAPE along A-A' section is shown in Fig.7a. It is evident that as the number of CPT increases from 11 to 21, all MAPE values reduce, implying an improved prediction. The best accuracy is obtained by RBFN. Conversely, when the available CPT number reduces from 11 to 4. Performances of RBFN and MPF deteriorate significantly with MAPE increasing from about 10% to over 35%. In comparison, the interpolation performance of BCS is relatively insensitive to the number of CPT and exhibits a mild increase in MAE. Moreover, when MAE is used as the comparison measure (see Fig.7b), RBFN performs best when more measurements are added. BCS achieves the best performance when only limited CPT soundings are available. Similar trends of MAPE and MAE along B-B' and C-C' sections are shown in Figs. 7c-7f.

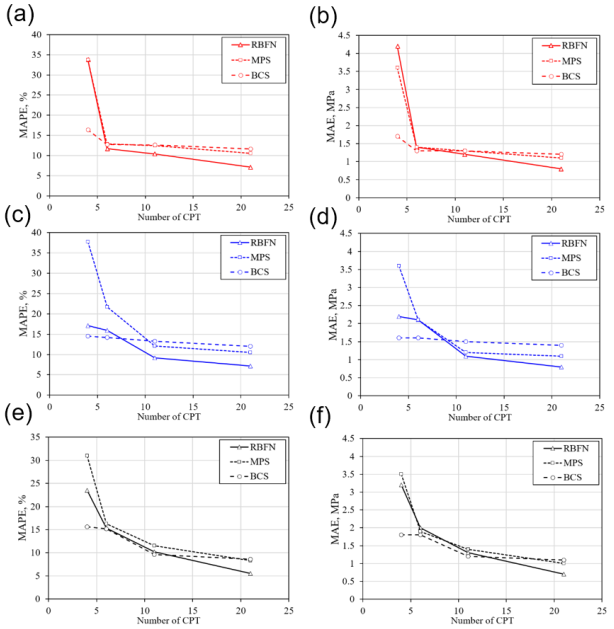

Fig.7 Comparison of interpolation performance with different measurement numbers along selected profiles: (a) MAPE along A-A' section; (b) MAE along A-A' section; (c) MAPE along B-B' section; (d) MAE along B-B' section; (e) MAPE along C-C' section; (f) MAE along C-C' section;

7. Summary and conclusion

The performance of three non-parametric data-driven approaches, namely ensemble Radial Basis Function Network (RBFN), Multiple Point Statistics (MPS) and Bayesian Compressive Sensing (BCS), in interpolating spatially varying non-Stationary and non-Gaussian cone

pressure profiles from sparse measurements was compared in this study. Moreover, the evolutions of best estimate and associated interpolation uncertainty with different line measurements (i.e., 4, 6, 11 and 21 CPT soundings) are explicitly investigated.

It is found that the ensemble RBFN and MPS outperforms BCS in reconstructing CPT profiles when more than 11 CPT soundings are taken as measurements. Conversely, when the number of available CPT profiles reduces to 4, the performance of ensemble RBFN and MPS deteriorates. In comparison, BCS interpolation is less sensitive to the number of measurement data and achieves superior interpolation performance when only sparse and limited measurements are available.

Acknowledgement
The work described in this paper was supported by grants from the Research Grant Council of Hong Kong Special Administrative Region, China (Project no. CityU 11213119 and T22-603/15N). The financial support is gratefully acknowledged.

References

Devendra, K. 2008. Soft Computing: Techniques and Its Applications in Electrical Engineering. Springer, Berlin, Germany.

Fenton, G. A. 1999. Random field modeling of CPT data. Journal of Geotechnical and Geoenvironmental Engineering, 125(6): 486-498.

Guardiano, F.B. and Srivastava, R.M. 1993. Multivariate geostatistics: beyond bivariate moments. In Geostatistics Troia'92.Springer. pp. 133-144.

Hu, Y., Zhao, T., Wang, Y., Choi, C., and Ng, C.W. 2019. Direct simulation of two-dimensional isotropic or anisotropic random field from sparse measurement using Bayesian compressive sampling. Stochastic Environmental Research and Risk Assessment. 33 (8-9), 1477-1496.

Jones, E., Oliphant, T., Peterson, P., and others. (2001). SciPy: Open source scientific tools for Python. Retrieved from "http://www.scipy.org/".

Li, J., and Heap, A. D. 2008. A review of spatial interpolation methods for environmental scientists. Geoscience Australia, Record 2008/23, 137pp.

Li, J., Heap, A. D., Potter, A., and Daniell, J.J. 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, 26(12): 1647-1659.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. In. National Institute of Science of India.

Mariethoz, G., and Renard, P. 2010. Reconstruction of incomplete data sets or images using direct sampling. Mathematical Geosciences, 42(3): 245-268.

Mariethoz, G. and Caers, J. 2014. Multiple-point geostatistics: stochastic modeling with training images. John Wiley & Sons.

Montoya-Noguera, S., Zhao, T., Hu, Y., Wang, Y., and Phoon, K. K. (2019). Simulation of non-stationary non-Gaussian random fields from sparse measurements using Bayesian compressive sampling and Karhunen-Loève expansion. Structural Safety, 79, 66-79.

Oliver, M., and Webster, R. 2014. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. Catena, 113: 56-69.

Shen, P., Zhang, L. M., and Zhu, H. (2016). Rainfall infiltration in a landslide soil deposit: Importance of inverse particle segregation. Engineering geology, 205, 116-132.

Shi, C. and Wang, Y. 2020. Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. Geoscience Frontiers, https://doi.org/10/1016/j.gsf.2020.01.011.

Strebelle, S. 2002. Sequential simulation drawing structures from training images. Ph. D. thesis, Stanford University, 374 pp.

Wang, Y., and Zhao, T. 2016. Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. Géotechnique, 67(6): 523-536.

Wang, Y., Zhao, T., and Phoon, K. 2017. Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. Canadian Geotechnical Journal, 55(6): 862-880.

Wang, Y., Zhao, T., Hu, Y., and Phoon, K. K. (2019). Simulation of random fields with trend from sparse measurements without detrending. Journal of Engineering Mechanics, ASCE, 145(2), 04018130.

Wang, Y., Hu, Y., and Zhao, T. 2020. CPT-based subsurface soil classification and zonation in a 2D vertical cross-section using Bayesian compressive sampling. Canadian Geotechnical Journal, https://doi.org/10.1139/cgj-2019-0131.

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

Zhao, T., Hu, Y., and Wang, Y. 2018. Statistical interpretation of spatially varying 2D geo-data from sparse measurements using Bayesian compressive sampling. Engineering Geology, 246: 162-175.