Proc. of the 6th Intl. Symposium on Reliability Engineering and Risk Management (6ISRERM)
31 May – 1 June 2018, Singapore
Editor(s) Xudong Qian, Sze Dai Pang, Ghim Ping Raymond Ong, Kok-Kwang Phoon

STATISTICAL INTERPOLATION OF SPATIALLY VARYING BUT SPARSELY MEASURED GEOTECHNICAL DATA USING BAYESIAN COMPRESSIVE SAMPLING

YUE HU[1], TENGYUAN ZHAO[1]and YU WANG[1]

[1]*Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Hong Kong SAR.*
E-mail: yuehu47-c@my.cityu.edu.hk

Geotechnical property data obtained from in situ or laboratory tests are usually sparse. In engineering practice, however, high-resolution geotechnical property profiles (e.g., variation of soil property along depth) are often preferred for engineering analysis and design. How to interpret a high-resolution profile from sparse measurement data in an objective manner is a challenge in geotechnical practice. Moreover, because of sparsity of the measured data, the interpolated results contain significant statistical uncertainty, which unavoidably affects subsequent geotechnical design and analysis, especially reliability-based design or analysis. Therefore, quantification of the statistical uncertainty is of great importance. However, such quantification is also challenging due to the spatially varying nature and auto-correlated pattern of soil properties. To address these two challenges, a Bayesian Compressive Sampling (BCS) approach is presented in this paper. It not only provides a high-resolution geotechnical property profile from sparse measurement data, but also quantifies the uncertainty in the interpolated results. An example of cone penetration test (CPT) tip resistance q_c profile is used for illustration. It shows that the BCS approach performs well and, as the number of measurement data points increases, the statistical uncertainty reduces quickly.

Keywords: site investigation, Bayesian approach, data interpolation, statistical uncertainty.

1 Introduction

Geotechnical data obtained from in situ or laboratory tests are usually sparse and limited in engineering practice due to budget, time limit etc. However, interpolated geotechnical data profiles (e.g., soil property variation along depth) with high resolution are often preferred when implementing engineering analysis and design (e.g., Mayne et al. 2002). Rational interpolation of geotechnical data from available measurement is a critical issue especially when the number of measurement is too small. There are many interpolation techniques including polynomial interpolation, kriging interpolation etc (e.g., Barthelmann et al. 2000). However, determination of the most appropriate order for polynomial is non-trivial, and estimation of spatial auto-correlation structure for kriging is difficult, when only sparse data are available (e.g., Wang et al. 2017a). Moreover, the sparsity of available measurement data results in significant statistical uncertainty within interpolated geotechnical data profile. The statistical uncertainty unavoidably propagates to and affects the subsequent engineering analysis and design, particularly for probabilistic analysis and reliability-based design (e.g., Bathurst et al. 2008). Traditional geostatistics is capable of deriving the uncertainty of interpolated geotechnical data profile based on conventional statistics. While the case is challenging when the number of measurement data

is too small (e.g., Webster and Oliver 2007). Quantification of statistical uncertainty becomes even more complex when the spatially varying and auto-correlated pattern are considered.

To address the above challenges, this paper develops a Bayesian approach to statistically interpolate geotechnical data profile from sparsely measured data, namely Bayesian compressive sampling (BCS). BCS method is able to not only provide interpolated geotechnical data with high resolution from limited measurement data but also quantify the associated statistical uncertainty. The quantified statistical uncertainty indicates the accuracy and reliability of interpolated geotechnical data profile and can be used to implement probability-based analysis (e.g., Wang et al. 2017b). It has also been suggested that BCS is more suitable for sparse measurements than the kriging method (e.g., Wang et al. 2017a). Moreover, the BCS method allows engineers to observe how the interpolated geotechnical data profile evolves when more and more measurement data are available as input. One set of cone penetration test (CPT) tip resistance profile collected from United States Geological Survey (USGS) database (Holzer et al. 2010) is adopted for illustration. This paper firstly covers the background of compressive sampling (CS), followed by the mathematical formulation of BCS. Then the real data example is presented.

2 Brief Review of Compressive Sampling

Compressive sampling is a novel sampling paradigm in signal processing (e.g., Candes and Wakin 2008; Donoho 2006). CS asserts that a signal can be completely recovered from partial samples based on the fact that many natural signals (e.g., the spatial variation of soil property along depth) are "compressible". The term "compressible" means a signal can be concisely expressed by weighted summation of a few pre-specified convenient basis functions (e.g., discrete cosine function, discrete wavelet function). This transformation process can be expressed as:

$$\boldsymbol{f} = \mathbf{B}\boldsymbol{\omega} \tag{1}$$

where $\boldsymbol{f}$ is a discrete real-valued signal, expressed as a column vector with a length of N; $\mathbf{B}$ is a pre-specified basis function matrix with dimension $N \times N$ (i.e., each column of $\mathbf{B}$ is a basis function); $\boldsymbol{\omega}$ is weight coefficients vector with a length of N while each component corresponding to the a basis function in $\mathbf{B}$. For a compressible signal, note that most elements in weight vector $\boldsymbol{\omega}$ are almost zeros, and just a few non-trivial elements possess significant or relatively large magnitude. Therefore, the underlying signal $\boldsymbol{f}$ can be approximated if those non-trivial coefficients in $\boldsymbol{\omega}$ can be identified by limited measurement from $\boldsymbol{f}$, namely, measurement data $\boldsymbol{y}$, which is expressed as:

$$\boldsymbol{y} = \mathbf{\Psi}\boldsymbol{f} = \mathbf{\Psi}\mathbf{B}\boldsymbol{\omega} = \mathbf{A}\boldsymbol{\omega} \tag{2}$$

where $\mathbf{\Psi}$ is measurement matrix with dimension $M \times N$ ($M<N$), reflecting the positions of measurement data $\boldsymbol{y}$; $\mathbf{\Psi}$ can be easily constructed from an identity matrix with dimension $N \times N$ according to the positions of measurement data $\boldsymbol{y}$ in $\boldsymbol{f}$. Since $M<N$, the above Eq. (2) is underdetermined, it cannot be solved directly. It can be solved by some other algorithms such as orthogonal matching pursuit (e.g., Cai and Wang 2011; Wang and Zhao 2016). If the non-trivial coefficients in $\boldsymbol{\omega}$ can be properly estimated as $\boldsymbol{\omega}_s$ by measurement data $\boldsymbol{y}$, the underlying signal $\boldsymbol{f}$ can be approximated as $\boldsymbol{f}'$, which is expressed as:

$$\boldsymbol{f} \approx \boldsymbol{f}' = \mathbf{B}\boldsymbol{\omega}_s \tag{3}$$

where $\boldsymbol{\omega}_s$ is the estimated non-trivial approximation coefficients with the same length as $\boldsymbol{\omega}$, while the trivial part are set to zeros. When measurement data $\boldsymbol{y}$ are sparse and limited from the underlying signal, the estimated approximation coefficients involves statistical uncertainty which can propagate to the recovered signal. To quantify the associated statistical uncertainty, the approximation coefficients are formulated as random variables under Bayesian framework (e.g., Wang and Zhao 2017), which is discussed in the following section.

3 Bayesian compressive sampling

3.1 *Bayesian framework*

Based on Bayes' theorem, the posterior PDF of $\boldsymbol{\omega}_s$ is expressed as:

$$p(\boldsymbol{\omega}_s \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\omega}_s) p(\boldsymbol{\omega}_s)}{p(\boldsymbol{y})} \tag{4}$$

where $p(\boldsymbol{\omega}_s \mid \boldsymbol{y})$ is the posterior PDF of $\boldsymbol{\omega}_s$ given measurement data $\boldsymbol{y}$; $p(\boldsymbol{y} \mid \boldsymbol{\omega}_s)$ is the likelihood of observing measurement data $\boldsymbol{y}$ given non-trivial approximation coefficients $\boldsymbol{\omega}_s$; $p(\boldsymbol{\omega}_s)$ is the prior PDF of non-trivial approximation coefficients $\boldsymbol{\omega}_s$; $p(\boldsymbol{y})$ is a constant to guarantee the integration of posterior PDF is unity.

The likelihood function $p(\boldsymbol{y} \mid \boldsymbol{\omega}_s)$ is formulated based on the residuals between measurement data $\boldsymbol{y}$ and recovered complete signal at corresponding locations. Suppose the residual follows a normal distribution with zero mean and unknown standard deviation σ. To facilitate the derivation of the Bayesian framework, a random variable α_0 is defined as the reciprocal of variance of the residual, i.e., $\alpha_0=\sigma^{-2}$. The likelihood function $p(\boldsymbol{y} \mid \boldsymbol{\omega}_s, \alpha_0)$ considering effect of variance of residual is formulated as a multivariate normal distribution assuming that the residuals at all measurement locations are independent of each other.

To be consistent with the likelihood formulation, the prior distribution $p(\boldsymbol{\omega}_s, \alpha_0)$ considering the effect of α_0 is formulated rather than $p(\boldsymbol{\omega}_s)$. To facilitate the derivation of prior distribution, $p(\boldsymbol{\omega}_s, \alpha_0)$ is taken to follow a multivariate normal-gamma distribution since normal likelihood and normal-gamma prior is a frequently used conjugate pair in Bayesian formulation (e.g., Murphy 2007) which provides analytical derivation for posterior distribution.

Given the conjugate pair of likelihood function $p(\boldsymbol{y} \mid \boldsymbol{\omega}_s, \alpha_0)$ and joint prior distribution $p(\boldsymbol{\omega}_s, \alpha_0)$, the posterior distribution $p(\boldsymbol{\omega}_s, \alpha_0 \mid \boldsymbol{y})$ considering α_0 can be derived based on Bayes' theorem in Eq. (4). According to the features of conjugate pair, the posterior distribution $p(\boldsymbol{\omega}_s, \alpha_0 \mid \boldsymbol{y})$ also follows a multivariate normal-gamma distribution. Since the α_0 is of little interest in this study, it will be removed after marginalization which is discussed in the following subsection.

3.2 *Marginalization*

Given the posterior distribution $p(\boldsymbol{\omega}_s, \alpha_0 \mid \boldsymbol{y})$, the marginal posterior distribution $p(\boldsymbol{\omega}_s \mid \boldsymbol{y})$ can be derived by marginalization which is expressed as the following integration (e.g., Sivia and Skilling 2006):

$$p(\boldsymbol{\omega}_s \mid \boldsymbol{y}) = \int p(\boldsymbol{\omega}_s, \alpha_0 \mid \boldsymbol{y}) d\alpha_0$$
$$= \frac{\Gamma\left(\frac{2c_n + N}{2}\right)}{\Gamma(c_n)\pi^{N/2}(2c_n)^{N/2}\left(\det\left[\frac{d_n}{c_n}\mathbf{H}\right]\right)^{1/2}\left(1 + \frac{1}{2c_n}(\boldsymbol{\omega}_s - \boldsymbol{\mu}_{\omega_s})^{\mathrm{T}}\left(\frac{d_n}{c_n}\mathbf{H}\right)^{-1}(\boldsymbol{\omega}_s - \boldsymbol{\mu}_{\omega_s})\right)^{\frac{2c_n+N}{2}}} \quad (5)$$

The posterior PDF of $\boldsymbol{\omega}_s$ given measurement data $\boldsymbol{y}$ is found to follow a multivariate Student's t distribution, with mean vector $\boldsymbol{\mu}_{\omega_s}$ and covariance matrix $\mathbf{COV}_{\omega_s}$ expressed as follow:

$$\boldsymbol{\mu}_{\omega_s} = \mathbf{H}\mathbf{A}^{\mathrm{T}}\boldsymbol{y} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{D}\right)^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{y}$$
$$\mathbf{COV}_{\omega_s} = \frac{d_n\mathbf{H}}{c_n - 1} = \frac{d_n\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{D}\right)^{-1}}{c_n - 1} \quad (6)$$

where $\mathbf{D}$ is a $N \times N$ diagonal matrix with diagonal elements $\mathbf{D}_{i,i}=\alpha_i$ (i=1, 2, 3,…, N). α_i is parameter related to variance of the i-th approximation coefficient. $c_n = M/2+c$; $d_n = d+(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{\mu}_{\omega_s}^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\mu}_{\omega_s})/2$. c and d shall be taken as small values (e.g., $c=d=10^{-4}$) to achieve an uninformative prior distribution of α_0, which is of little interest in this study. Note that, the statistics of $\boldsymbol{\omega}_s$ in Eq. (6) determine the distribution of approximation coefficients once the variance parameters α_i are determined, which are discussed in the next subsection.

3.3 *Determination of hyper-parameters*

Note that the statistics of $\boldsymbol{\omega}_s$ in Eq. (6) are conditional to α_i, which are hyper-parameters since α_i also depend on other parameters. The hyper-parameters can be estimated by maximum likelihood method (e.g., Bishop 2006). The idea of maximum likelihood is that the most probable α_i should be the ones that maximize the likelihood of measurement data $\boldsymbol{y}$, or equivalently its logarithm (e.g., Ji et al. 2009):

$$L = \ln(p(\boldsymbol{y})) = -\frac{1}{2}\left[(M + 2c)\ln\left(\boldsymbol{y}^T\mathbf{C}^{-1}\boldsymbol{y} + 2d\right) + \ln\left(\det(\mathbf{C})\right)\right] + const \quad (7)$$

where $\mathbf{C}=\mathbf{I}_{M\times M}+\mathbf{A}\mathbf{D}^{-1}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{I}_{M\times M}$ is an identity matrix with dimension $M \times M$. The most probable α_i can be derived by differentiating the above Eq. (7) with respect to α_i. Setting the derivative to zero gives:

$$\alpha_i = \frac{1}{\mu_i^2(M + 2c)/(\boldsymbol{y}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{y} + 2d) + H_{i,i}} \quad (8)$$

where μ_i is the i-th element in $\boldsymbol{\mu}_{\omega_s}$; $H_{i,i}$ is the i-th diagonal in matrix $\mathbf{H}$. Note that the statistics of $\boldsymbol{\omega}_s$ in Eq. (6) depend on α_i while α_i also depend on statistics of $\boldsymbol{\omega}_s$ in Eq. (6). This suggests that iterations should be conducted between Eq. (6) and (8). By setting an initial set of α_i (i=1, 2, 3,…, N), Eq. (6) can be determined. Then updated α_i can be obtained according to Eq. (8). This process can be repeated until the likelihood function (7) reaches its maximum. After that, the most probable statistics of $\boldsymbol{\omega}_s$ can be derived. Therefore the best estimate of signal of interest

$\boldsymbol{\mu}_{f'}$ can be reconstructed using $\boldsymbol{\mu}_{\omega_s}$, and the associated statistical uncertainty can be quantified easily by covariance matrix $\mathbf{COV}_{f'}$, expressed as follow:

$$\begin{aligned} \boldsymbol{\mu}_{f'} &= E(\boldsymbol{f}') = E(\mathbf{B}\boldsymbol{\omega}_s) = \mathbf{B}\boldsymbol{\mu}_{\omega_s} \\ \mathbf{COV}_{f'} &= \mathbf{B}\mathbf{COV}_{\omega_s}\mathbf{B}^{\mathrm{T}} \end{aligned} \tag{9}$$

4 Real data example

To illustrate the performance of BCS for interpolating spatially varying but sparsely measured geotechnical data, BCS is applied to a cone penetration test (CPT) tip resistance profile collected from United States Geological Survey (USGS) database (Holzer et al. 2010). The test was conducted within a sand layer in Mississippi river valley, Poinsett County, Arkansas. For illustration, a profile with length 256 ranging from 7.75m to 20.5m with sampling interval 0.05m is considered as original profile represented by solid line in Figure 1. Limited data points from that profile are selected as measurement data input to implement BCS shown by open circle. The whole CPT tip resistance profile can be interpolated by BCS using only M=12 measurement data points shown by red dash line in Figure 1 (a) with quantified statistical uncertainty expressed as confidence interval (95%). The confidence interval of 95% which indicates the reliability of interpolated CPT profile covers most of variation of the original CPT profile.

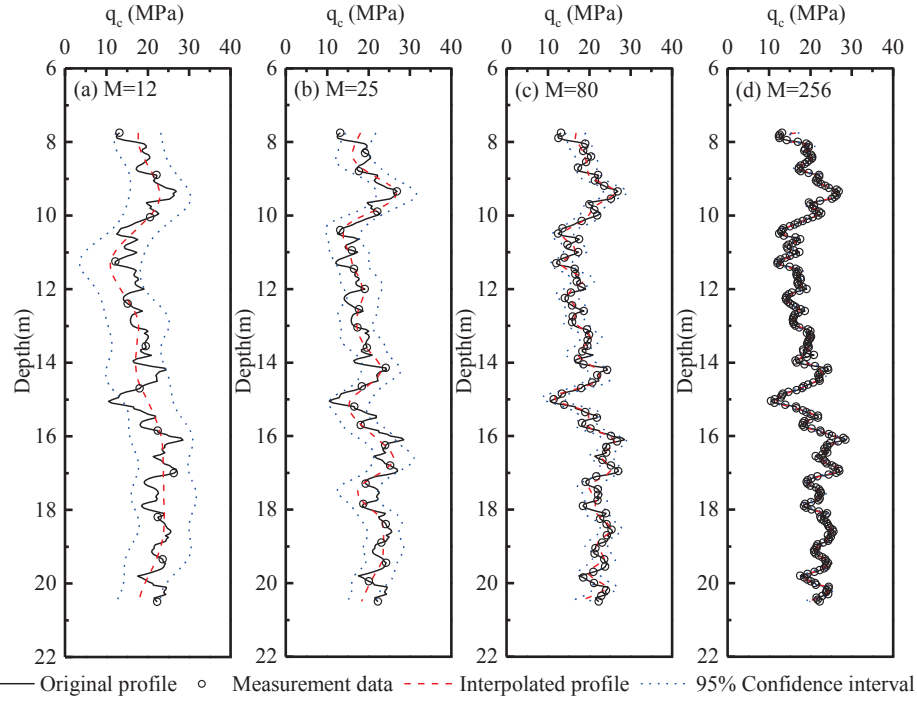

Figure 1. Comparison between original profile and statistically interpolated profile under different measurement scenarios: (a) M=12; (b) M=25; (c) M=80; (d) M=256

Moreover, complete profile interpolated by BCS converges to the original profile as the number of measurement data increases. Three more measurement scenarios, i.e., M=25, M=80 and M=256 are provided as shown in subplot (b)-(d) in Figure 1. It shows that the interpolated profile becomes more and more similar to the original profile as more and more measurement

data involved. Note that the quantified uncertainty also becomes narrower as the number of measurement data increases. When all data points are selected as measurement input as shown in Figure 1 (d), the interpolated profile are almost identical to the original one and the quantified statistical uncertainty reduces to almost zero.

5 Conclusion

Bayesian compressive sampling (BCS) method is presented to statistically interpolate spatially varying but sparsely measured data. This method is applied to a set of real engineering data. It shows that BCS can not only rationally interpolate a complete geotechnical profile from limited measurement but also quantify the statistical uncertainty. The quantified uncertainty indicates the accuracy and reliability of the interpolated profile. Moreover, the interpolated profile converges to the original profile and the uncertainty region reduces to almost zero when the number of measurement data increases. The BCS method is robust and can be applied to different geotechnical data.

Acknowledgement

The work described in this paper is supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 9042172 (CityU 11200115) and Project No. 8779012 (T22-603/15N)). The financial support is gratefully acknowledged.

References

Barthelmann, V., Novak, E. & Ritter, K. 2000. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, 12, 273-288.

Bathurst, R.J., Allen, T.M. & Nowak, A.S. 2008. Calibration concepts for load and resistance factor design (LRFD) of reinforced soil walls. *Canadian Geotechnical Journal*, 45, 1377-1392, doi: 10.1139/T08-063.

Bishop, C.M. 2006. Pattern recognition. *Machine Learning*, 128, 1-58.

Cai, T.T. & Wang, L. 2011. Orthogonal matching pursuit for sparse signal recovery with noise. IEEE Transactions on Information Theory, 57, 4680-4688.

Candes, E.J. & Wakin, M.B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25, 21-30, doi: 10.1109/MSP.2007.914731.

Donoho, D.L. 2006. Compressed sensing. *IEEE Transactions on Information Theory*, 52, 1289-1306, doi: 10.1109/TIT.2006.871582.

Holzer, T.L., Noce, T.E. & Bennett, M.J. 2010. *Maps and documentation of seismic CPT soundings in the Central, Eastern, and Western United States.* US Geological Survey Report 2331-1258.

Ji, S., Dunson, D. & Carin, L. 2009. Multitask compressive sensing. *IEEE Transactions on Signal Processing*, 57, 92-106.

Mayne, P.W., Christopher, B.R. & DeJong, J. 2002. *Subsurface investigations—geotechnical site characterization* Report FHWA NHI-01-031.

Murphy, K.P. 2007. *Conjugate Bayesian analysis of the Gaussian distribution*. 1, 16.

Sivia, D. & Skilling, J. 2006. *Data analysis: a Bayesian tutorial*. OUP Oxford.

Wang, Y. & Zhao, T. 2016. Interpretation of soil property profile from limited measurement data: a compressive sampling perspective. *Canadian Geotechnical Journal*, 53, 1547-1559.

Wang, Y. & Zhao, T. 2017. Statistical interpretation of soil property profiles from sparse data using Bayesian compressive sampling. *Géotechnique*, 67, 523-536, doi: 10.1680/jgeot.16.P.143.

Wang, Y., Akeju, O.V. & Zhao, T. 2017a. Interpolation of spatially varying but sparsely measured geo-data: a comparative study. Engineering Geology, 231, 200-217.

Wang, Y., Zhao, T. & Phoon, K.-K. 2017b. Direct simulation of random field samples from sparsely measured geotechnical data with consideration of uncertainty in interpretation. Canadian Geotechnical Journal.

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