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)

Bayesian inference on the estimation of cross-variogram

J. B. Xu[1], L. L. Zhang[2*] and Y. Wang [3]

[1]*State Key Laboratory of Ocean Engineering, Department of Civil Engineering, Shanghai Jiao Tong University, Shanghai, China; Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Hong Kong SAR. Email: xujiabao@sjtu.edu.cn*
[2*]*State Key Laboratory of Ocean Engineering, Department of Civil Engineering, Shanghai Jiao Tong University, Shanghai, China. Email: lulu_zhang@sjtu.edu.cn*
[3]*Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Hong Kong SAR. Email: yuwang@cityu.edu.hk*

Abstract: Site characterization is an essential step in geotechnical design and analysis. Because the measurements of interested soil properties are limited and sparse, data fusion methods are commonly utilized to characterize a site in geotechnical engineering. Among kinds of data fusion methods, the cokriging method is a widely used geostatistical interpolation method. Cokriging can improve the performance of site characterization of interested soil properties by merging the measurements of correlated soil properties. The correlation between a primary variable and a secondary variable is represented by a cross-variogram model in cokriging. Traditionally, the cross-variogram can only be calculated when the measurements of the primary and secondary variables are co-located. However, the measurements of different kinds of geotechnical properties are commonly not co-located, which restricts the extensive utilization of cokriging. In this study, a Bayesian inference method is proposed to estimate the cross-variogram model when the measurements of the primary and secondary variables are not co-located. The proposed method is illustrated and validated by the dataset of the Ely site in Ely, Nevada. The results show that the Bayesian inference method can estimate the cross-variogram model when the measurements of the primary and secondary variables are not co-located. Moreover, the uncertainty of variogram models and cokriging estimation can also be measured by the proposed Bayesian inference method.

Keywords: correlation; cross-variogram; cokriging; Bayesian inference; uncertainty.

1. Introduction

Site characterization is an indispensable step in geotechnical design and analysis, which can show the spatial distribution of geotechnical properties. Traditionally, a proper site characterization requires plenty of in situ or tested measurements at different locations. However, the number of geotechnical measurements is usually limited and sparse (Wang and Zhao 2017; Zhao and Wang 2018). Therefore, how to characterize a site adequately based on limited and sparse measurements in geotechnical engineering is an intractable problem.

In a geotechnical engineering site, there are always some measurements of other soil properties, which are correlated with the interested soil property. The measurements of correlated soil properties can help improve the performance of site characterization of interested soil property by data fusion methods.

Among a variety of data fusion methods, cokriging is a widely used geostatistical interpolation method in geotechnical engineering. Cokriging can improve the prediction of an interested soil property (primary variable) by merging measurements of correlated soil properties (secondary variable). The correlation between the primary variable and the secondary variable is expressed by the cross-variogram model. The variability of each variable is described by an auto-variogram model. The auto-variogram model and cross-variogram model are calculated based on the measurements of the primary variable and secondary variable in cokriging. Moreover, the cross-variogram can only be calculated from the co-located measurements of the primary variable and secondary variable (Papritz et al. 1993). In a geotechnical site investigation, however, the measurements of a soil property are usually obtained by destructive sampling. Therefore, there are seldom co-located measurements of the primary variable and secondary variable for most geotechnical conditions.

In this study, a Bayesian inference method is firstly developed to estimate the cross-variogram when the measurements of the primary variable and secondary variable are not co-located. This paper firstly reviews the theory of cross-variogram. Then, the Bayesian inference method is introduced to estimate the cross-variogram model when the measurements of the primary variable and secondary variable are not co-located. Subsequently, a set of real elevation data is used to demonstrate this proposed Bayesian inference method.

2. Proposed method

2.1 Review of the cross-variogram

In the cokriging method, the cross-variogram model is used to express the correlation between a primary variable and a secondary variable. If the primary variable and secondary variable are second-order stationary, the cross-variogram is defined as (Journel 1986; Pyrcz and Deutsch 2014):

$$\gamma_{12}(\mathbf{h})=\frac{1}{2}E\left\{\left[Z_1(\mathbf{x}+\mathbf{h})-Z_1(\mathbf{x})\right]\left[Z_2(\mathbf{x}+\mathbf{h})-Z_2(\mathbf{x})\right]\right\} \quad (1)$$

where $\gamma_{12}(\mathbf{h})$ is the cross-variogram of the primary variable $Z_1(\mathbf{x})$ and the secondary variable $Z_2(\mathbf{x})$; $\mathbf{x}$ is the spatial coordinates; $\mathbf{h}$ is the lag distance; E {} is the expectation symbol. The definition implies that the cross-variogram can only be calculated from the co-located measurements of the primary variable and

secondary variable. However, the measurements of correlated geotechnical properties are usually non-co-located, which restricts the estimation of the cross-variogram in geotechnical engineering.

When the primary variable and secondary variable are assumed to be second-order stationary, the cross-variogram can be expressed by the cross-covariance:

$$\gamma_{12}(\mathbf{h}) = C_{12}(0) - C_{12}(\mathbf{h}) \quad (2)$$

where C_{12} (**h**) is the cross-covariance between the primary variable and secondary variable.

For second order stationary correlated geotechnical variables, the joint probability density function (pdf) can be modeled as a joint Gaussian distribution. The covariance matrix in the joint pdf shows the correlation between the primary variable and the secondary variable. The parameters of the cross-covariance model used in the covariance matrix can be used to obtain the parameters of the cross-variogram model by Bayesian inference. In the next section, a Bayesian inference will be developed to estimate the cross-variogram directly from measurements when the measurements of the primary variable and secondary variable are not co-located.

2.2 Bayesian inference for the cross-variogram

Bayesian inference is a probabilistic approach that can update prior knowledge with measurements to estimate the posterior distribution of model parameters (Zhang et al. 2014; Zhang et al. 2018; Xu et al. 2020). One of the most critical parts of Bayesian inference is the likelihood function, which measures the goodness of fit of a model to measurement data. When random variables are Gaussian distribution and second order stationary, the likelihood function can be constructed by the joint probability distribution function of the measurements of correlated random variables.

In this research, the corresponding logarithmic likelihood function of the parameters can be determined based on the measurements of primary and secondary variables:

$$\log[P(\mathbf{z}\,|\,\boldsymbol{\theta})] = \text{constant-}\frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \quad (3)$$

where log [P (**z** | **θ**)] is the logarithmic likelihood function, $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]^{\mathrm{T}}$, $\mathbf{z}_1$ and $\mathbf{z}_2$ are the vectors of measurements of primary and secondary variables, respectively. $\boldsymbol{\mu} = [\mu_1, \mu_2]$, μ_1 and μ_2 are the mean values of primary and secondary variables, respectively. The covariance matrix C is defined as

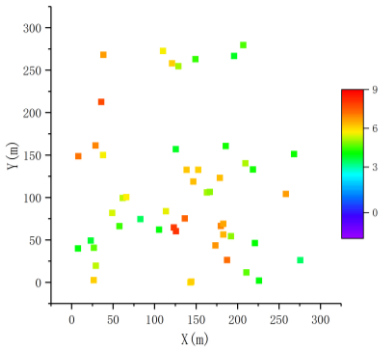
$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \quad (4)$$

where $\mathbf{C}_{11}$ is the covariance matrix between the measurements of the primary variable, $\mathbf{C}_{22}$ is the covariance matrix between the measurements of the secondary variable, $\mathbf{C}_{12}$ and $\mathbf{C}_{21}$ are the cross-covariance matrices between the measurements of the primary variable and secondary variable, $\mathbf{C}_{21} = \mathbf{C}_{12}$. **θ** is a vector containing the parameters of covariance and cross-covariance and the mean values of the primary and secondary variables. In this study, the Matérn model is adopted as the cross-variogram and cross-covariance model (Minasny and McBratney 2005).

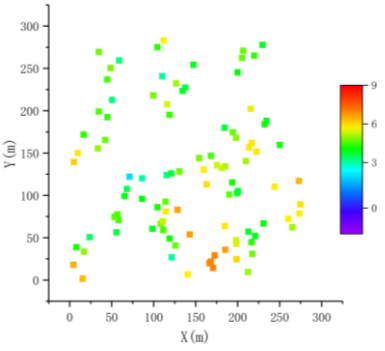
When the likelihood function is constructed, a set of prior information is assigned. Subsequently, the posterior information of covariance and cross-covariance parameters can be obtained by Markov chain Monte Carlo (MCMC) simulation. In this study, the differential evolution adaptive Metropolis (DREAM) algorithm proposed by Vrugt et al. (2008) is adopted. The DREAM algorithm is a multi-chain MCMC simulation algorithm. The Bayesian inference using MCMC simulation can generate plenty of stably convergent posterior parameter samples of the covariance and cross-covariance, which can be utilized to model the auto-variogram and cross-variogram models. Subsequently, the auto-variogram and cross-variogram models can be utilized to interpret at unsampled locations using the cokriging method.

3. Real data example

In this study, a 2D dataset derived from the published Ely dataset (Journel and Kyriakidis 2004; Remy et al. 2009) is used as an example. The Ely dataset contains two sets of data with 10, 000 co-located measurements, which can be utilized as the measurements of the primary variable and secondary variable, respectively. The two datasets were transformed from a digital elevation model of a study site near the town of Ely in eastern Nevada. This site is a squared area of approximately 300 m by 300 m.



(a) 50 measurements of the primary variable

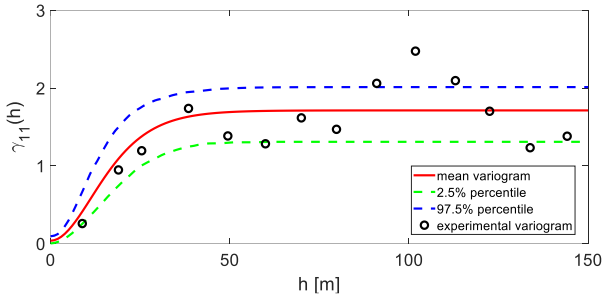


(b) 100 measurements of the secondary variable

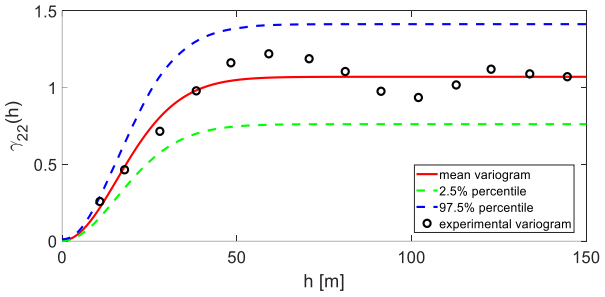
Figure 1. Locations of (a) 50 measurements of the primary variable and (b) 100 measurements of the secondary variable in the Ely site.

50 data points of the primary variable and 200 non-co-located data points of the secondary variable, shown in Fig. 1, are chosen to investigate the ability of the Bayesian inference method in estimating the cross-variogram when measurements are not co-located.

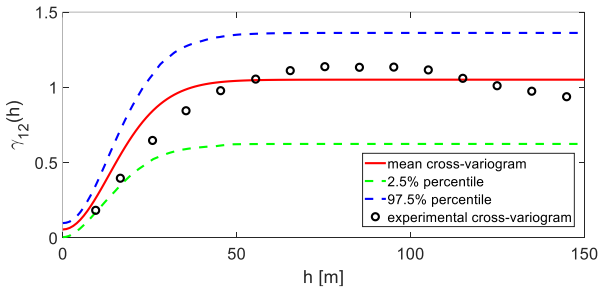
In this research, a set of uniform distribution is assigned as the prior information. The Bayesian inference method using MCMC simulation runs 10 different Markov Chains simultaneously in parallel. After 5000 iterations, 50, 000 posterior samples of parameters are obtained. The last converged 10, 000 posterior samples are used to explore the variogram model.



(a) auto-variogram of the primary variable



(b) auto-variogram of the secondary variable



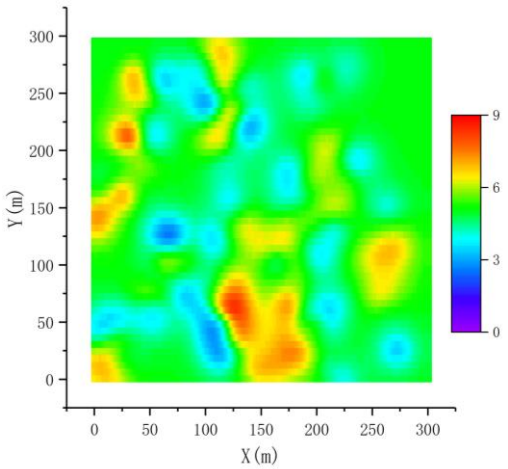
(c) cross-variogram

Figure 2. Estimated variogram of the Ely dataset: (a) auto-variogram of the primary variable; (b) auto-variogram of the secondary variable; (c) cross-variogram.

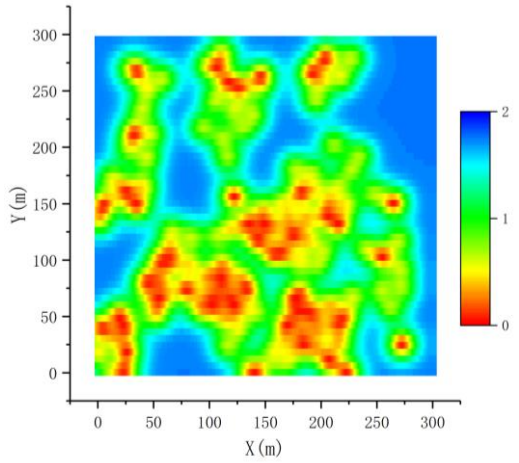
Fig. 2 displays the posterior mean and 95% confidence intervals of the estimated variogram models from the Bayesian inference method. The Bayesian inference method can estimate not only the cross-variogram model but also the auto-variogram models of the primary and secondary variables simultaneously. The experimental auto-variograms are calculated from the method of moments and plotted in this graph to compare with the results of the Bayesian inference method. In this research, the experimental cross-variogram cannot be calculated because the measurements of the primary variable and secondary variable are not co-located. Therefore, the original 10,000 co-located measurements of the primary variable and the secondary variable are used to calculate the experimental cross-variogram and evaluate the estimated cross-variogram from the Bayesian inference.

Fig. 2 shows that the 95% confidence intervals of estimated auto-variogram and cross-variogram models encompass the experimental auto-variogram and cross-variogram models. The mean auto-variogram and cross-variogram models coincide with most of the experimental auto-variogram and cross-variogram values, especially when the lag distance is small. Therefore, the Bayesian inference method is appropriate to estimate the cross-variogram by measurements directly when the measurements of primary and secondary variables are not co-located.

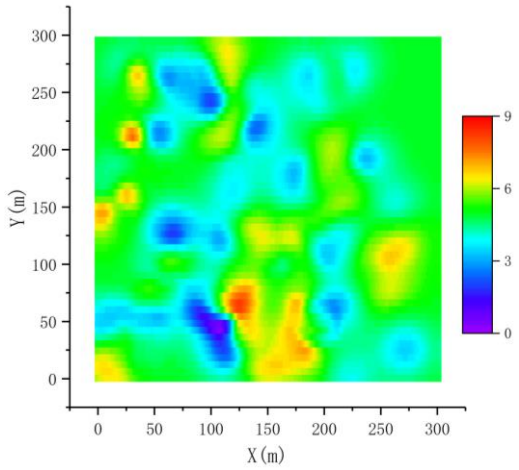
The Bayesian inference method proposed in this study can ensure that the covariance matrix in Eq. 3 is positive definite. Therefore, all posterior auto-variogram and cross-variogram models can be used in cokriging to produce amounts of cokriging predictions and associated cokriging variances accordingly. When the posterior variogram models are utilized to produce cokriging predictions at unsampled locations, the uncertainty of variogram parameters will propagate to the cokriging prediction. In this study, the last 100 converged posterior variogram models are applied to characterize this site by cokriging. Fig. 3 shows the mean cokriging interpretation and the mean cokriging variance. The graph of mean cokriging variance (Fig. 3b) shows that cokriging variance is large at those positions away from measurements. The 2.5th percentile of cokriging prediction (Fig. 3c) and 97.5th percentile of cokriging prediction (Fig. 3d) are also plotted to show the uncertainty of site characterization using the uncertain variogram models by cokriging method.



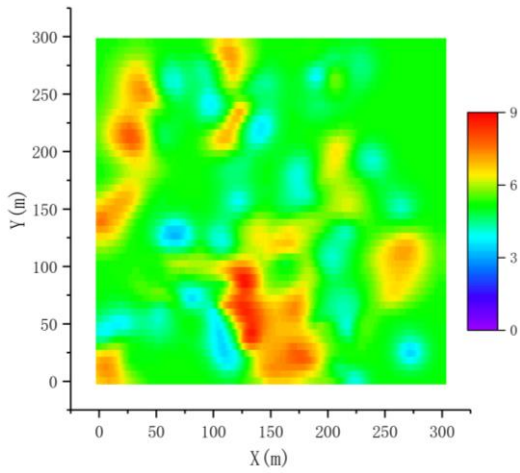
(a) mean cokriging interpretation



(b) mean cokriging variance



(c) 2.5th percentile of cokriging interpretation



(d) 97.5th percentile of cokriging interpretation

Figure 3. (a) Mean cokriging interpretation; (b) mean cokriging variance; (c) 2.5th percentile, and (d) 97.5th percentile
of the cokriging interpretation.

4. Conclusions

In this study, a Bayesian inference method is developed to estimate the cross-variogram when the measurements of primary and secondary variables are not co-located. A set of real elevation data is used to demonstrate this proposed method. The results show that the proposed method can estimate an accurate cross-variogram model directly from non-co-located measurements. Furthermore, the uncertainty of the estimated variogram models and cokriging interpretation can also be measured by the Bayesian inference method.

Acknowledgements
The work in this paper was supported by the Natural Science Foundation of China (Project No. 51679135, 51979158, 51639008 and 51422905) and the Program of Shanghai Academic Research Leader by Science and Technology Commission of Shanghai Municipality (Project No. 19XD1421900).

References

Journel, A. G. 1986. Geostatistics: Models and Tools for the Earth Sciences. *Mathematical Geology*, 18 (1): 119-140.

Journel, A.G., and Kyriakidis, P.C. 2004. Evaluation of Mineral Reserves: A Simulation Approach: Oxford University Press, New York, pp. 49–51.

Minasny, B., and McBratney, A.B. 2005. The Matern Function as a General Model for Soil Variograms. *Geoderma*, 128 (3-4):192–207.

Papritz, A., Künsch, H.R., and Webster, R. 1993. On the Pseudo Cross-Variogram. *Mathematical Geology*, 25 (8): 1015–1026.

Pyrcz, M. J., and Deutsch, C. V. 2014. Geostatistical Reservoir Modeling, second edition: New York: Oxford University Press.

Remy, N., Boucher, A., and Wu, J. 2009. Applied Geostatistics with SGeMS: A User's Guide. Cambridge University Press, New York, pp. 80–81.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J. M., and Robinson, B.A. 2008. Treatment of Input Uncertainty in Hydrologic Modeling: Doing Hydrology Backward with Markov Chain Monte Carlo Simulation. *Water Resources Research*, 44: W00B09.

Wang, Y., and Zhao, T.Y. 2017. Statistical Interpretation of Soil Property Profiles from Sparse Data Using Bayesian Compressive Sampling. *Geotechnique*, 67 (6): 523–536.

Xu, J.B., Zhang, L.L., Li, J.H., Cao, Z.J., Yang, H.Q., and Chen, X. 2020. Probabilistic estimation of variogram parameters of geotechnical properties with a trend based on Bayesian inference using Markov chain Monte Carlo simulation: *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards.* doi:10.1080/17499518.2020.1757720

Zhang, L.L., Zheng, Y.F., Zhang, L.M., Li, X., and Wang, J.H. 2014. Probabilistic Model Calibration for Soil Slope under Rainfall: Effects of Measurement Duration and Frequency in Field Monitoring. *Geotechnique*, 64 (5): 365–378.

Zhang, L.L., Wu, F., Zheng, Y.F., Chen, L.H., Zhang, J., and Li, X. 2018. Probabilistic Calibration of a Coupled

Hydro-Mechanical Slope Stability Model with Integration of Multiple Observations. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 12 (3): 169–182.

Zhao, T.Y., 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.