Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5182-7_01-006-cd

Bayesian Subsurface Mapping Using CPT Data

Antonis Mavritsakis[1], Timo Schweckendiek[1], Ana Teixeira[1] and Eleni Smyrniou[1]

[1]Safe and Resilient Infrastructure, Deltares, Boussinesqweg 1, the Netherlands.
[2]Department of Hydraulic Engineering Delft University of Technology, The Netherlands
E-mail:Antonis.Mavritsakis@deltares.nl

Abstract: Data-Driven Site Characterization aids geotechnical engineering and decision-making by producing a 3D soil parameter map of the subsurface. Unfortunately, limited site investigation data availability renders most traditional Machine Learning methods inadequate. In this paper, a framework for Bayesian Site Characterization (BaySiC) is applied on an artificial site investigation dataset. The framework aims to evaluate the statistics of the CPT measurements (cone resistance and sleeve friction) and establish their relationship, but also identify the soil heterogeneity patterns and classify the soil type at each point of the subsurface. The Bayesian framework can deal with small site investigation datasets and quantifies the prediction uncertainty with runtimes that are feasible for practical application. Essentially, the Bayesian framework maps the parameters over the subsurface on a probabilistic level. Therefore, the potency of the framework is not only judged based on traditional metrics, such as the adjacency of the mean prediction to the data, but also the likelihood attributed to the data, which is intuitively expressed by the visualization of the prediction credible intervals over the subsurface. This is shown by means of a case from a benchmark exercise.

Keywords: Site Characterization; Bayesian Inference; Subsurface Mapping, CPT measurements; Machine Learning

1 Introduction

Ideally, site characterization would give us full knowledge of the subsurface, i.e.soil layering and all relevant soil parameters. But the reality is that geotechnical engineering must find ways to merge the limited amount of data with the existent/empirical knowledge.Data-driven site characterization aids geotechnical engineering to predict soil parameter values and mapping them over the subsurface. Limited site investigation data availability renders most traditional Machine Learning methods inadequate for site characterization. However, Machine Learning approaches enabled by Bayesian techniques prove effective in such settings. Bayesian inference updates parameter distributions by combining pre-existing knowledge to data, accurate predictions, as well as quantification of the associated uncertainty.Especially the quantification of the uncertainty is of extreme importance during the decision-making process. In addition to that, identifying spatial correlation and autocorrelation patterns and producing a3D mapping of the subsurface, can highly help understanding the problem at hands, and allow the prediction of soil parameter values (and their probability distribution) per location of the subsoil. The Bayesian Site Characterization (BaySiC) framework presented in this paper aims thus to interpret site investigation data on a probabilistic level using Bayesian inference. It can establish relationships between parameters, identify spatial correlation patterns and assess the uncertainty of parameter predictions.

Section 1 and 2 are introductory sections to the problem and the theory behind the Bayesian site characterization framework. Section 2 also presents the methodology for the subsurface mapping using Bayesian techniques, and Section 3 demonstrates the application and performance by means of the benchmark exercise presented in (Phoon et al., 2022)(Phoon et al., 2022)(Phoon et al., 2022)(Phoon et al., 2022)using CPT measurements. Finally, Section 4 presents the conclusions and recommendations.

2 Theoretical background

2.1 Bayesian updating

Bayesian inference employs Bayes' theorem to draw conclusions on the variables (X) by conditioning on observations (ε)(Gelman et al., 2013). Bayesian inference entails the setup of a statistical model, that determines the sources of epistemic (reducible) and aleatory (irreducible) uncertainty present in the examined problem. Besides, Bayesian inference aims to reduce epistemic and provide a better description of aleatory uncertainty. In this study, Bayesian inference is performed by sampling using the Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011).

2.2 Bayesian model formulation

The benchmark exercise entails training the inference modelusing the CPT measurement data (observations ε) at the training locations and predicting the values of the measurements at the validation locations in a multi-layer

subsoil domain. Specifically, the CPT measurements are the cone resistance (q_t) and the sleeve friction (f_s).Moreover, the exercise requires the prediction of the layer type at the validation locations.The BaySiC framework employed in this study examines the benchmarking exercise through a Bayesian perspective. The goal is to use a statistical model that describes the observations and connects them to predictions at the test locations.

The basis of the statistical model is the assumption that the observations (ε) at each training point can be approximated by a constant mean (μ), which isa function of variablevector(X) and the coordinates per point (Eq. 1)(Geyer, Papaioannou and Straub, 2021), as well as a measurement error term (e), which in this study is assumed to be null.BaySiC models the subsurface using random fields (RF) (Vanmarcke, 2010).For this example, where RFs for 2 parameters (q_t and f_s) have to be jointly modelled, RF modellingadopts a multivariate normal distribution, whose covariance matrix (Σ) incorporates the variance of the observed variables q_t and f_s, the cross-correlation between them, as well as the autocorrelation due to spatial variability.Consequently, the observations can be described by the likelihood function of Eq. 2, sinceof q_t and $f_s$at all measurement points is a multivariate normal distribution. Even though this assumption might lead to issues such as negative values, it is adopted for compliance with the benchmark's settings.

$$\varepsilon(x,y,z) = [\varepsilon_{q_t}(x,y,z) \quad \varepsilon_{f_s}(x,y,z)]^T = \mu(X,(x,y,z)) + e \tag{1}$$

$$L_\varepsilon(X) = \mathrm{N}(\varepsilon|\mu,\Sigma) \tag{2}$$

The statistical model uses the variable vector Xas the basis of the assumed behavior. It should be noted that the model follows a "lumped" layer approach.Essentially, it assumes the existence of a single layer in the subsoil domainandcannotdistinguish between the different layers of the subsoil.The main random variables of X are given in Eq. 3. In detail, they are:the vector of global mean per observed variable (μ),the vector of standard deviation per observed variable (σ),the cross-correlation matrix (C_{cross}), and the vector of autocorrelation lengths per direction of the subsoil (θ).

$$X = \left[\mu_{q_t}, \mu_{f_s}, \sigma_{q_t}, \sigma_{f_s}, C_{cross}, \theta_v, \theta_h\right]^{\mathrm{T}} \tag{3}$$

For the sake of showcasing the effectiveness of the method, as well as due to the lack of context about the subsoil of the exercise, weakly informative priors are assumed for all variables of vector X. It should be noted that the lack of informative priors means that the influence of the likelihood function and the observations on the inference of the posteriors is expected to be greater.

Following this, some auxiliary variables are defined by the statistical model to aid the evaluation of the likelihood function. In detail, $\mathrm{C_{cross}}$is the cross-covariance matrix, estimated by the standard deviation vector and the cross-correlation matrix. Also, the autocorrelation matrix (C_{auto}) is derived by applying the distances between the training points ($d_{mat,h}$ and $d_{mat,v}$ for the horizontal and vertical directions respectively) and the vector of autocorrelation lengths in the Markov autocorrelation function (Eq. 4).Lastly, Σ is defined as the Kronecker product of the cross-covariance and autocorrelation matrices (Eq 5) ,and integrates their influence in the likelihood function.

$$C_{auto}(\theta) = \exp\left(\sqrt{\left(\frac{d_{mat,h}}{\theta_h}\right)^2 + \left(\frac{d_{mat,v}}{\theta_v}\right)^2}\right) \tag{4}$$

$$\Sigma = C_{auto} \otimes \mathrm{C_{cross}} \tag{5}$$

2.3 Posterior predictive distribution at the validation points

The points in the 3D subsurface can be distinguished between trainingand validation points, which belong to the training and validation locations respectively. The training points, where the observations are available, have already been used in conditioning the Bayesian model and deriving the posterior distributions of the variables, as described in paragraph 2.2.The assumption made for the likelihood function of BaySiC is expanded to the validation points. Now, the q_t and f_s values at the training and predictions points are assumed to follow a multivariate normal distribution with mean μ_S and covariance Σ_s, both of them being functions of the inferred variable vector X. The values at the training points are given, so making predictions requires conditioning the multivariate distribution to the observations at the training points. According to (Marriott and Eaton, 1984), the conditional distribution of the CPT measurements at the prediction points (ε_p) given the observations at the training points ($\varepsilon_t = \varepsilon$) is a multivariate normal distribution parametrized by the mean $\mu_{v|\varepsilon_t=\varepsilon}$ and the covariance $\Sigma_{v|\varepsilon_t=\varepsilon}$ (Eq. 6 and Eq. 7). Eq. 8to Eq. 10 formulate the derivation of these parameters as a function of the components of the global mean vector and covariance matrix. As a result, the conditional distribution is a function of the variables inferred by the Bayesian model and the measurements at the training points.

$$\mu_s = [\mu_t \quad \mu_v]^T \tag{6}$$

$$\Sigma_s = \begin{bmatrix} \Sigma_{tt} & \Sigma_{tv} \\ \Sigma_{vt} & \Sigma_{vv} \end{bmatrix} \tag{7}$$

$$(\varepsilon_v|\varepsilon_t = \varepsilon) \sim \mathrm{N}(\mu_{v|\varepsilon_t=\varepsilon}, \Sigma_{v|\varepsilon_t=\varepsilon}) \tag{8}$$

$$\mu_{v|\varepsilon_t=\varepsilon} = \mu_v + \Sigma_{vt}\Sigma_{tt}^{-1}(\varepsilon - \mu_t) \tag{9}$$

$$\Sigma_{v|\varepsilon_t=\varepsilon} = \Sigma_{vv} - \Sigma_{vt}\Sigma_{tt}^{-1}\Sigma_{tv} \tag{10}$$

This conditional distribution is used to determine the posterior predictive distribution at the validation points, by marginalizing over the posterior distribution of the Bayesian model's variables (X)(Gelman et al., 2013).Having established that the parameters of the conditional distribution are function of the inferred variables, the posterior predictive distribution at the validation points can be expressed through the posterior sample of the Bayesian model as in Eq. 11. Samples from the posterior predictive take the form of random fields of CPT measurements, conditioned to the measurements at the training locations.

$$P\left(\varepsilon_p|\varepsilon_t = \varepsilon\right) = \frac{1}{N}\sum_{i=1}^{N} N(\mu_{v|\varepsilon_t=\varepsilon}(X_i), \Sigma_{v|\varepsilon_t=\varepsilon}(X_i)) \tag{11}$$

2.4 Prediction evaluation metrics

This study employs three metrics for the evaluation of the competency of the presented approach per case. Two metrics were provided by the benchmark exercise creators as common comparison ground for all methods(Phoon et al., 2022). These are the Root Mean Squared Error (RMSE) and the Identification Rate (IR).Firstly, the RMSE (Eq. 12)is a means of assessing the average distance between the cone resistance of the validation dataset and the prediction per location (n points are present over the vertical profile of each validation location).Since BaySiCderives conclusions about X on a probabilistic level, the expectation of RMSE is used as the comparison metric. The RMSE meanis formulated in Eq. 13, where N is the number of posterior samples of X. Secondly, the IR (Eq. 14)counts the points per validation location where the layer prediction, quantified by the Robertson's Soil Behavior Type (SBT) numbermatches the validation dataset (Robertson, 2016). The estimation of the SBT, and thus the IR, are functions of the variable vector X.Themean of $IR$is evaluated according to Eq. 15.

$$RMSE(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i - \mu_{p|\varepsilon_t=\varepsilon}(\mathrm{X})\right]\Big)^2} \tag{12}$$

$$\overline{RMSE} = \int RMSE(X)p(X|\varepsilon)dX = \frac{1}{N}\sum_{i=1}^{N} RMSE(X_i) \tag{13}$$

$$IR(X) = \frac{1}{n}\sum_{i=1}^{n_{points}} \mathrm{I(X_i)}\,, \mathrm{I_i(X)} = \begin{cases} 1, \; SBT(\varepsilon_i) = SBT\left([\mu_{p|\varepsilon_t=\varepsilon}(X)]_{z=z_i}\right) \\ 0, \; SBT(\varepsilon_i) \neq SBT\left([\mu_{p|\varepsilon_t=\varepsilon}(X)]_{z=z_i}\right) \end{cases} \tag{14}$$

$$\overline{IR} = \int IR(X)p(X|\varepsilon)dX = \frac{1}{N}\sum_{i=1}^{N} IR(X_i) \tag{15}$$

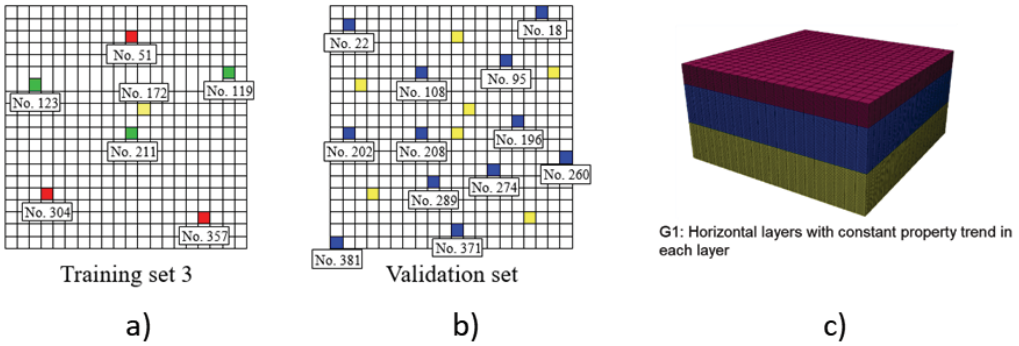

Figure 1. Panel a) shows the locations of the training CPTs and panel b) the locations of the validation CPTs in the top view of the soil domain. Panel c) shows the stratigraphy of the subsoil. The figures are taken from the benchmark example.

3 Subsurface mapping results

3.1 Benchmark exercise description

This sectiondemonstrates the application and performance of the BaySiC framework described in section2 in the benchmarking exercise of (Phoon et al., 2022).The exercise focuses only on one of the available benchmark settings, which employs training set T3 and subsoil case G1. The top view of the training and validation set

formations isshown in Figure 1. As observed, validation CPT locations can exist outside the formation of training locations. The CPT profiles of all locations reach the full 10meter depth of the subsoil domain. Moreover, the same Figure 1presents the subsoil formation of the exercise, which consists of horizontal layers of sand, clay and silt.

The training data of the benchmark exercise is the input of the BaySiC framework.BaySiC conditions the random variables (Eq. 3) to this data and infers their posterior distributions and the posterior predictive distribution at the validation points. The latter is used to determine the expectations of the validation metrics, which indicate the performance of BaySiC.

3.2 Inference results

Inference is performed on the training set T3 and 10,000 samples are collected from the posterior. The inference time for collecting the samples is approximately 60minutes, as performed by a laptop with an i5 processor and 8GB of RAM.

The posterior of the global means and standard deviations for the CPT measurements are presented in Figure 2. Considering that this is a lumped layer approach, the posteriors are not expected to approximate the underlying target values that generated the training data. The same figureshows the posterior distribution of the coefficient of correlation between q_t and f_s (ρ), which is the off-diagonal term of the covariance matrix. The posterior shows undeniable correlation between the observed variables. Specifically, the 5th quantile of the posterior sample is $\rho_{0.05} = 0.37$, while the lowest sample found is $\rho_{min} = 0.33$.

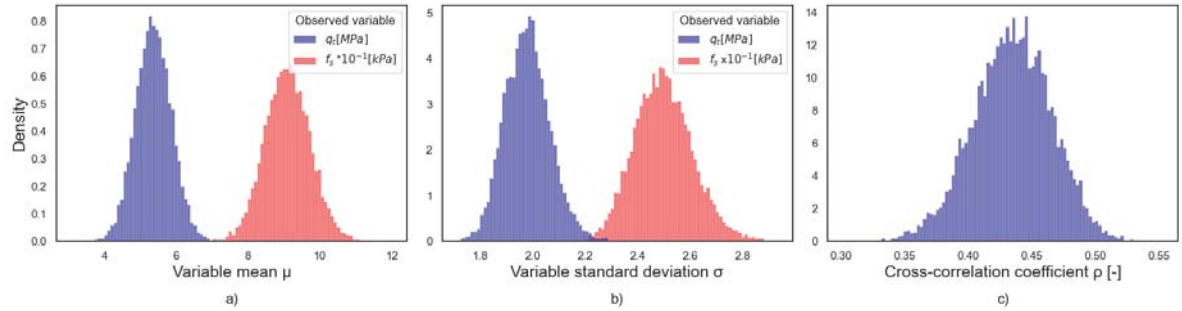

Figure 2. Posteriors of the a) mean and b) standard deviation for the q_t and f_s. The mean and standard deviation of f_s have been scaled by 10^{-1} for better visualization. Panel c) shows the posterior of the cross-correlation factor between q_t and f_s.

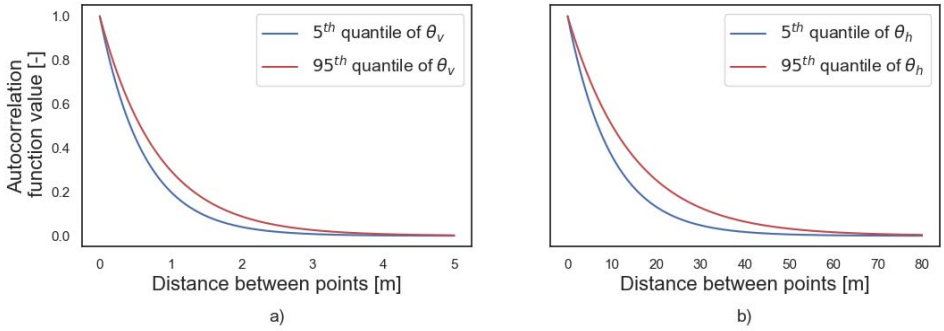

Figure 3. Comparison of the autocorrelation values for the 5th and 95th quantiles of the posteriors of: a) θ_v and b) θ_h.

The spatial variability of the "lumped" layer is controlled by the autocorrelation lengths. The impact of the variance of these variables is not directly clear, so Figure 3 compares the autocorrelation function plot over distance for the 5th and 95th quantiles of each θ posterior. The difference between the quantile lines is small for the vertical direction but can be considerable for the horizontal direction. The observations are densely packed over the vertical (along the CPT profile), so characterization of θ_v is quite accurate. On the other hand, the CPT formation appears to have considerably greater distances between the locations than θ_h, which leads to poorer inference

3.3 Prediction results

Following, the inferred posterior is used to predict the observed variables (q_t and f_s) at the validation points and evaluate the adopted metrics, as described in paragraph 2.4. The scores of the method are exhibited in Table **1**.

Table 1. Resulting meanper metric

Metric	RMSE[-]	IR [-]
Mean	1.23	0.81

Figure 4 presents the meanprediction of q_t at the validation locations. The q_t meanlies closely to the validation data and the conditional fields from the posterior predictive achievean RMSE of 1.23 MPa.The expectation profile in all locations follows the$q_t$of the clay layer closely, due to its low deviation. At some depths, the$q_t$prediction profile deviates from the validation data in the sand and silt layers, which fluctuate more intensely. Although such deviationsexist, the method is still able to grasp the general trend and follow the shape of the validation data per layer. Moreover, the effect of adopting a "lumped" layer approach is apparent; the predictioncurvetransfers smoothly between layers, because it has not been set up to comprehend the existence of multiple layers. Additionally, the credible intervals envelope the validation data over a considerable range of the profiles, while also closely capturing the validation curve shape. Lastly, a shortcoming of adopting the 'lumped' layer approach is that the prediction uncertainty per layer is influenced by the uncertainty of all layers. As a result, the prediction uncertainty of the clay layer is relatively large, as it is affected by the uncertainty of the sand and silt layers, which are considerably greater.

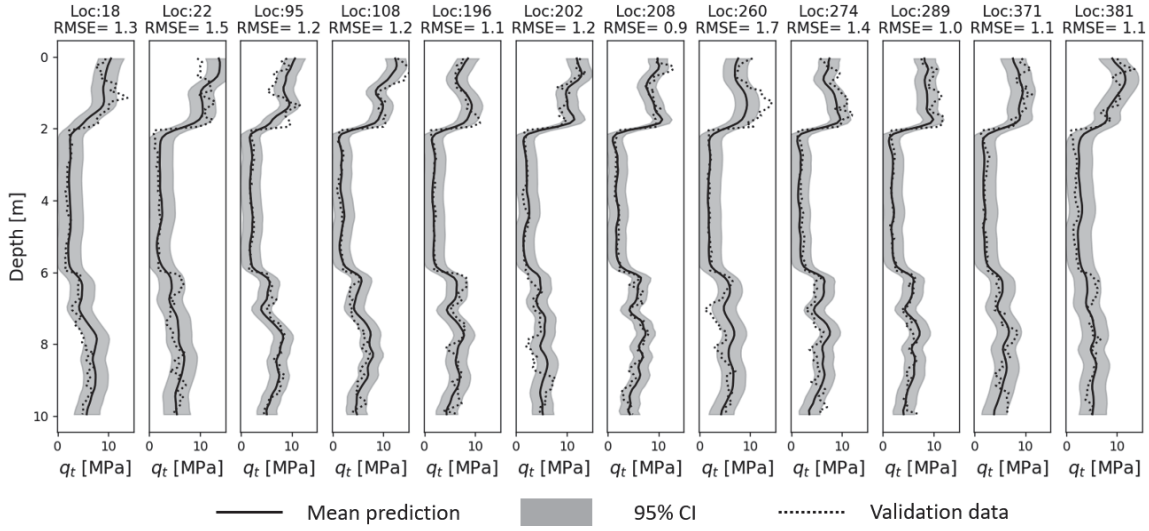

Figure 4. Mean predictionof q_t, 95% credible interval and validation data at the validation locations.

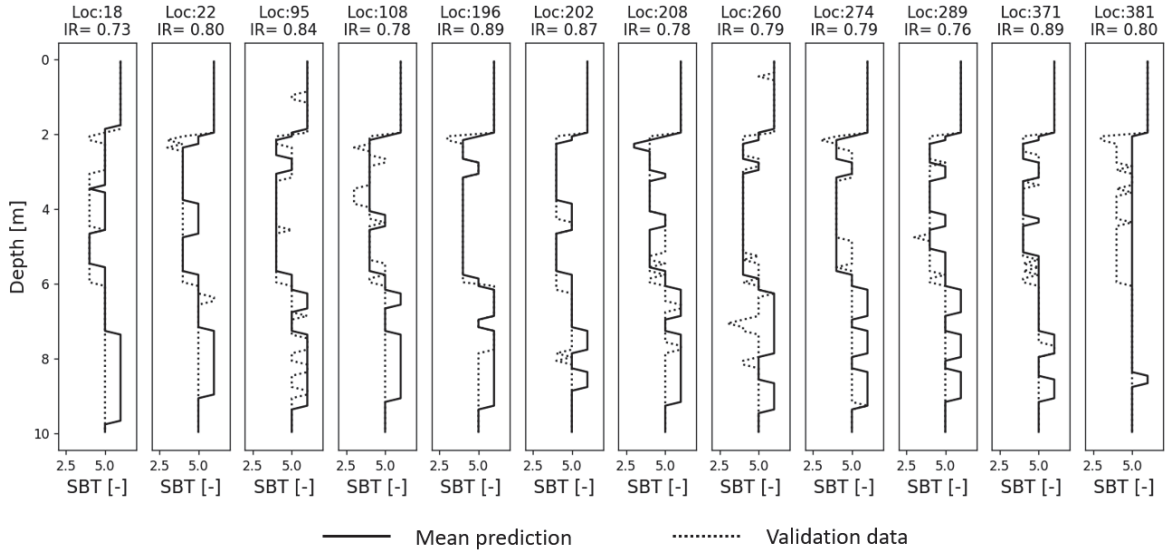

Figure 5. Layer type prediction based on the meanSBT prediction along with the validation data at the validation locations.

Figure 5 demonstrates the accuracy of the layer type prediction. BaySiC achieves an average IR of 0.81. Specifically, BaySiC is competent in following recognizing the body of the layer, meaning that it can follow the main type of each formation. According to the SBT estimation for the validation data, each layer contains lenses of other material types. Although the prediction has been accurate for some lenses, it is overall not able to detect them consistently This behavior is expected; because lenses occur as localized anomalies for the training and validation sets and hardly follow a spatial pattern.

Lastly, it is worthwhile to mention that expectedly, the greatest RMSE and lowest IR scores are met at validationCPTsthat are located further from the training CPT formation.Greater distances from the training set reduce the horizontal autocorrelation andmakes the predictions more uncertain and less accurate.

4 Conclusions

This paper presentsBaySiC, a framework for Bayesian Site Characterization,by approaching a benchmarking exercise for subsurface mapping of CPT parameters provided by(Phoon et al., 2022). The goal of the exercise is to illustrate the effectiveness and efficiency of Bayesian methods in a data drivensite characterization problem.The study adopts a Bayesian perspective and brings the problem to a probabilistic level,where all predictions are associated with the corresponding probability of occurrence. Naturally, the method focuses on accurate predictions through uncertainty reduction.

Firstly, a statistical model that describes the global uncertainty and spatial variability of CPT parameters has been set up. A centerpiece of BaySiC is the definition of the likelihood function, which establishesrelationships between observations over a 3Dsubsoil domain. Bayesian inference is used to condition the variables of the statistical model on the CPT observations at the training locations of the exercise. Subsequently, a prediction step determines the parameter distributions at the validation points.

One case of the benchmark has been analyzed in the paper, to the end of demonstrating the functionof BaySiC.The predictions of the method are competent in approximating the validation datain terms of parameter values and profile shape.Additionally, the method is in general effective in recognizing the layer type per validation point but suffers from the presence of localized anomalies.Ultimately, BaySiC has proven to be competent in deriving posterior distributions that lead to accurate and precise predictions both in terms of CPT parameter values and layer type identification.

In conclusion, the paper suggests a Bayesian standpoint to the topic of Machine learning for subsurface mappingusing CPT parameters. Even with a simplified perception of the subsurface, the Bayesian model has been effective in generating an accurate map of cone resistance, sleeve friction and layer typesof the subsoil.Lastly, it is noted that the method has proven to be computationally efficient, achieving runtimes that can be facilitated by project settings.

Acknowledgments

The authors would like to acknowledge the contribution of Prof. Jianye Ching of the National Taiwan University, for his support in the development of BaySiC. Moreover, the authors acknowledge the scientific and organizing effort of Kok-Kwang Phoon, Takayuki Shuku, Jianye Ching and Ikumasa Yoshida for preparing the benchmark exercise and distributing it to the community.

References

Gelman, A. et al. (2013) Bayesian Data Analysis. *Chapman and Hall/CRC*. Available at: https://doi.org/10.1201/b16018.

Geyer, S., Papaioannou, I. and Straub, D. (2021) 'Bayesian analysis of hierarchical random fields for material modeling', *Probabilistic Engineering Mechanics, 66*. Available at: https://doi.org/10.1016/j.probengmech.2021.103167.

Marriott, F.H.C. and Eaton, M.L. (1984) 'Multivariate Statistics: A Vector Space Approach.', *Applied Statistics*, 33(3), p. 319. Available at: https://doi.org/10.2307/2347710.

Neal, R.M. (2011) MCMC Using Hamiltonian Dynamics.

Phoon, K.-K. et al. (2022) 'Benchmark examples for data-driven site characterisation', *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, pp. 1–23. Available at: https://doi.org/10.1080/17499518.2022.2025541.

Robertson, P.K. (2016) 'Cone penetration test (CPT)-based soil behaviour type (SBT) classification system — an update', *Canadian Geotechnical Journal*, 53(12), pp. 1910–1927. Available at: https://doi.org/10.1139/cgj-2016-0044.

Vanmarcke, E. (2010) Random Fields: analysis and synthesis. *WORLD SCIENTIFIC*. Available at: https://doi.org/10.1142/5807.