Multivariate Gaussian Process for 3D subsurface stratigraphy prediction from CPT and labelled borehole data

Orestis Zinas[1*], *Iason* Papaioannou[2], *Ronald* Schneider[1] and *Pablo* Cuéllar[1]

[1] *Bundesanstalt für Materialforschung und -prüfung (BAM), Buildings and Structures, Unter den Eichen 87, 12205 Berlin, Germany*
[2] *Technische Universität München, Engineering Risk Analysis Group, Arcisstraße 21, 80333 München, Germany*
[*] *Corresponding author: orestis.zinas@bam.de*

ABSTRACT

Quantifying uncertainties in subsurface properties and stratigraphy can lead to better understanding of the ground conditions and enhance the design and assessment of geotechnical structures. Several studies have utilized Cone Penetration Test (CPT) data and employed Bayesian and Machine Learning methods to quantify the geological uncertainty, based on the Robertson's soil classification charts and the Soil Behaviour Type Index (I_c). The incorporation of borehole data can reduce the stratigraphic uncertainty. Significant challenges can arise, however, mainly due to the intrinsic differences between field and laboratory-based soil classification systems, which can potentially lead to inconsistent soil classification. To this end, this study proposes a multivariate Gaussian Process model that utilizes site-specific data and: i) jointly models multiple categorical (USCS labels) and continuous (I_c) variables, ii) learns a (shared) spatial correlation structure and the between-outputs covariance, and iii) produces two types of dependent classification outputs. The results indicate that the integration of geotechnical and geological information into a unified model can provide more reliable predictions of the subsurface stratification, by allowing simultaneous interpretation of USCS and I_c profiles. Importantly, the model demonstrates the potential to integrate multiple variables of different types, aiming to contribute to the development of a methodology for joint modeling of geotechnical, geological and geophysical data.

Keywords: soil classification; categorical borehole variables; CPT data; Multivariate Gaussian Process.

1. Introduction

Reliable prediction of subsurface stratigraphy is essential for informed decision-making in geotechnical engineering. Cone Penetration Tests (CPTs) have been proved valuable tools for characterizing soil behavior and stratigraphy, providing high resolution measurements of the soil mechanical response along depth. Among the various soil classification systems linking cone parameters to soil type, arguably the most popular are the Robertson's Soil Behavior Type (SBT) charts. The soil boundary delineation is based on the cone tip resistance (q_c) and the sleeve friction (f_s) measurements, and can be well approximated by the bounds of the empirically derived SBT Index (I_c) (Robertson 2009, 2016). Several studies employed continuous CPT data and statistical or Machine Learning (ML) methods for geological modeling. Some of these methods include, spatial regression and Gaussian Process (GP) models (Ching 2021; Ching and Yoshida 2023), Geotechnical lasso (Shuku and Phoon 2021), Compressive Sampling (Hu and Wang 2020), clustering and Gausian Mixture Models (e.g., Shakir 2023), Markov Random Fields (e.g., Wang 2019).

Apart from CPTs, borehole drillings are commonly performed as part of site investigations, extracting soil samples that undergo laboratory testing. The Unified Soil Classification System (USCS) serves as a primary classification scheme, categorizing soils into distinct groups based on their textural and plasticity properties. Some of the statistical and ML approaches that have been investigated for stratigraphic modeling from sparse borehole drillings include Markov Random Fields (MRFs) (e.g., Li et al. 2016; Wang et al. 2017; Shuku and Phoon 2023), coupled Markov chain model (CMC, Elfeki and Dekking 2001; Qi et al. 2016, Zhang et al. 2022), convolutional neural networks (CNN, Shi and Wang 2021), XGBoost (Chen and Guestrin 2016) and combinations thereof (Wei and Wang 2022).

By integrating CPT and borehole data one seeks to enhance the delineation of soil boundaries and achieve more reliable predictions of stratigraphy, compared to relying solely on singular classification systems. Recent studies in this direction have primarily explored clustering approaches based on MRFs and CMC (Wang et al. 2019, Xiao et al. 2021), and conditional Random field models

(Farahbakhsh and Ching 2023).

In this study, we propose a multivariate Gaussian Process Regression model for joint stratigraphy prediction from CPT and borehole data. Our approach is solely based on site-specific data and does not require the use of soil databases. More importantly, the model has the capability of incorporating multiple categorical USCS labels, which are readily available from the site-specific reported borehole logs.

Our modeling approach assumes that the observations of each USCS class and of I_c result from correlated Gaussian Processes, i.e., from a multivariate Gaussian Process. These processes share the same spatial correlation structure, with hyperparameters (e.g., correlation lengths) estimated based on all available data. In this way, we seek not only to overcome identifiability issues of scales of fluctuation, arising from the horizontal spatial sparsity of the data, but also to reduce the noise and mitigate the presence of multiple thin soil layers commonly observed in I_c profiles. Additionally, we estimate the cross-covariance between categorical USCS labels and I_c, using all site-specific data at our disposal. To reduce the computational effort, we employ the Maximum Likelihood Estimation (MLE) for the model hyperparameters, instead of MCMC sampling. The incorporation of the categorical USCS variables to our model made possible through their approximate transformation to Gaussian distributed variables, following (Milios et al. 2018). In the following, the mathematical details of the proposed methodology are presented, which is subsequently applied to a real dataset from a New Zealand site.

2. Multivariate Gaussian Process

Let $\boldsymbol{f}(\boldsymbol{s}) = (f_1(\boldsymbol{s}), ..., f_q(\boldsymbol{s}))^T$ with $\boldsymbol{s} \in D \subset \mathbb{R}^d$ be a vector-valued process with outcome space $\mathbb{R}^q$. If the finite dimensional joint distribution of $\boldsymbol{f}(\boldsymbol{s})$ is multivariate Normal, then $\boldsymbol{f}$ is a multivariate Gaussian Process:

$$\boldsymbol{f}(\boldsymbol{s}) \sim GP(\boldsymbol{m}(\boldsymbol{s}), \boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}')) \tag{1}$$

where $\boldsymbol{m}(\boldsymbol{s}) = \mathrm{E}[\boldsymbol{f}(\boldsymbol{s})])$ is a $q \times 1$ vector valued mean function, $\boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}')$ a $q \times q$ covariance matrix function, with (k, l) element $C_{k,l}(\boldsymbol{s}, \boldsymbol{s}') = \mathrm{Cov}[f_l(\boldsymbol{s}), f_k(\boldsymbol{s}')]$, that returns the covariance of two components of the process at two distinct locations.

Two main assumptions are adopted in this study. First, the mean function $\boldsymbol{m}(\boldsymbol{s})$ is a-priori, i.e., prior to performing the measurements, constant for each random process. Second, the covariance function $\boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}')$ is separable, such that it can be specified as the product between a spatial correlation function (kernel) $\rho_\theta(\boldsymbol{s}, \boldsymbol{s}')$ with hyperparameters $\boldsymbol{\theta}$, and a non-spatial, positive definite cross-covariance matrix $\boldsymbol{\Sigma}$ (e.g., Bonilla 2007),

$$\boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}') = \boldsymbol{\Sigma} \cdot \rho_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{s}'). \tag{2}$$

The definition of the covariance function in Eq. (2) implies that all GPs share the same spatial correlation structure. Although this assumption may be restrictive (or invalid) in some cases where different types of soil variables are modelled, in this study where we aim to jointly model the USCS labels and the SBT index, we expect that the spatial correlation of the different soil patterns will be similar, given that the physical processes involved are the same (e.g. weathering process). Several models have been proposed for the spatial correlation function. In this study, a single exponential function was selected,

$$\rho_\theta(\boldsymbol{s}, \boldsymbol{s}') = \exp\left[-\left(\frac{\sqrt{{\Delta s_x}^2 + {\Delta s_y}^2}}{\theta_h} + \frac{|\Delta s_z|}{\theta_z}\right)\right] \tag{3}$$

which implies isotropy in the horizontal plane, and separability between the horizontal and vertical correlation functions. Apart from the computational gains enabled by the separable structure of $\boldsymbol{C}(\boldsymbol{s}, \boldsymbol{s}')$, this assumption can potentially lead to an improved estimation of the correlation lengths.

In this study, we are particularly interested in the case where not all components are observed at the same locations, such as when boreholes and CPTs are available at different locations. In the geostatistics literature, this data structure is often called *heterotopic*, or *asymmetric* in Machine Learning (e.g., Alvarez 2012). Let $\boldsymbol{S} = \{\boldsymbol{S}_l\}_{l=1}^q$, indicate the collection of all spatial locations in the training data, where $\boldsymbol{S}_l = \{\boldsymbol{s}_{l,i}\}_{i=1}^{n_l}$, denote the spatial locations of each component process. We define the random vector $\boldsymbol{f}(\boldsymbol{S}) = ((f_1(s_{1,1}), ..., f_1(s_{1,n})), ..., (f_q(s_{q,1}), ..., f_1(s_{q,n})))$ of dimensions $nq \times 1$. $\boldsymbol{f}(\boldsymbol{S})$ is a Gaussian random vector with

$$\boldsymbol{f}(\boldsymbol{S}) \sim N(\boldsymbol{m}(\boldsymbol{S}), \boldsymbol{C}(\boldsymbol{S}, \boldsymbol{S})) \tag{4}$$

where $\boldsymbol{m}(\boldsymbol{S})$ the concatenated $nq \times 1$ vector of means, containing distinct constant values at the respective locations of each component, i.e., $\{m_l(\boldsymbol{s}_{l,i}) = \mu_l\}_{i=1}^{n_l}$, $l = 1, ..., q$, and $\boldsymbol{C}(\boldsymbol{S}, \boldsymbol{S})$ the $nq \times nq$ block partitioned covariance matrix given by:

$$\boldsymbol{C} = \begin{bmatrix} (\boldsymbol{C}(\boldsymbol{S}_1, \boldsymbol{S}_1))_{1,1} & \cdots & (\boldsymbol{C}(\boldsymbol{S}_1, \boldsymbol{S}_q))_{1,q} \\ (\boldsymbol{C}(\boldsymbol{S}_2, \boldsymbol{S}_1))_{2,1} & \cdots & (\boldsymbol{C}(\boldsymbol{S}_2, \boldsymbol{S}_q))_{2,q} \\ \vdots & \cdots & \vdots \\ (\boldsymbol{C}(\boldsymbol{S}_q, \boldsymbol{S}_1))_{q,1} & \cdots & (\boldsymbol{C}(\boldsymbol{S}_q, \boldsymbol{S}_q))_{q,q} \end{bmatrix}. \tag{5}$$

For a set of observations $\boldsymbol{y}$ at locations $\boldsymbol{S}$, the *marginal* likelihood emerging from the Gaussian distribution assumption can be written as (Alvarez 2012):

$$p(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma}) = N(\boldsymbol{y}|\boldsymbol{m}(\boldsymbol{S}|\boldsymbol{\mu}),\boldsymbol{C}(\boldsymbol{S},\boldsymbol{S}|\boldsymbol{\theta},\boldsymbol{\Sigma}) + \boldsymbol{D}) \quad (6)$$

where $\boldsymbol{\mu} = \{\mu_l\}_{l=1}^{q}$, $\boldsymbol{D}$ is a $nq \times nq$ diagonal matrix with diagonal blocks, containing the noise variances of the observations. The estimation of the hyperarameters is achieved by maximizing $\ln(p(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\Sigma}))$, e.g. by using a stochastic gradient descent optimizer, such as the Adaptive Moment estimation (Adam) (Kingma and Ba 2015). The predictive distribution of $\boldsymbol{f}_*$ at some new locations $\boldsymbol{S_*}$ is again Gaussian, and is given by:

$$p(\boldsymbol{f}_*|\boldsymbol{S_*},\boldsymbol{S},\boldsymbol{y}) = N(\boldsymbol{f_*}|\boldsymbol{m}_{*|\boldsymbol{S}},\boldsymbol{C}_{*|\boldsymbol{S}}), \quad (7)$$

with,

$$\begin{aligned} \boldsymbol{m}_{*|\boldsymbol{S}} &= \boldsymbol{m}_* + \boldsymbol{C}_{\boldsymbol{S},*}^T(\boldsymbol{C}_{\boldsymbol{S},\boldsymbol{S}} + \boldsymbol{D})^{-1}(\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{S}}) \\ \boldsymbol{C}_{*|\boldsymbol{S}} &= \boldsymbol{C}_{*,*} - \boldsymbol{C}_{\boldsymbol{S},*}^T(\boldsymbol{C}_{\boldsymbol{S},\boldsymbol{S}} + \boldsymbol{D})^{-1}\boldsymbol{C}_{\boldsymbol{S},*} \end{aligned} \quad (8)$$

where $\boldsymbol{m}_{\boldsymbol{S}} = \boldsymbol{m}(\boldsymbol{S})$, $\boldsymbol{m}_* = \boldsymbol{m}(\boldsymbol{S}_*)$, $\boldsymbol{C}_{\boldsymbol{S},\boldsymbol{S}} = \boldsymbol{C}(\boldsymbol{S},\boldsymbol{S})$, $\boldsymbol{C}_{*,*} = \boldsymbol{C}(\boldsymbol{S}_*,\boldsymbol{S}_*)$ and $\boldsymbol{C}_{\boldsymbol{S},*}$ is a $nq \times q$ matrix function with entries $(\boldsymbol{C}(\boldsymbol{S}_i,\boldsymbol{S}_*))_{p,p'}$ for $i = 1,...,n$ and $p,p' = 1,...,q$.

3. Gaussian Processes for joint Regression and Classification

The model introduced in the preceding section is directly suitable as a Gaussian Process Regression model for continuous soil variables. However, since the aim of this study is to jointly model categorical (USCS) and continuous (I_c) variables, additional steps are required. Disregarding for the moment the CPT data, the problem at hand can be viewed as a multi-label classification problem. Each spatial location within the training boreholes, is associated with a specific USCS label, e.g. "ML" corresponding to silt. In general, there are in total 15 soil groups according to the USCS system (e.g. Das 2015). Consider the case where 4 soil groups are observed within the locations of the available boreholes. Assuming that there are 5 available boreholes with similar depth range from which we select a subset of 100 locations along depth, there are in total $5 \times 100 = 500$ locations yielding observations of USCS labels. Each of these locations can be associated to a vector with 4 components, with value "1" assigned to the observed class c and "0" to the other 3 labels. The borehole data can be collected into a 500×4 matrix $\boldsymbol{Y}_{bh}$, with binary values in the entries associated with the locations of "presence" or "absence" of the 4 USCS labels.

Consider now that there are also 10 available CPTs at the site at hand, and that we select again 100 locations along depth from each CPT, yielding data of I_c. Hence, there are $10 \times 100 = 1000$ CPT locations, each one containing an I_c value, collected in a 1000×1 vector. Hence there are in total 1500 locations yielding observations on USCS labels and I_c. These can be viewed as observations of a 5-component stochastic process $\boldsymbol{f} = (f_1,...,f_4,f_{I_c})$. To approach this problem with the multivariate Gaussian Process Regression model presented in the previous section it is necessary to seek for a transformation of the categorical USCS related variables to the Gaussian space, given that I_c can be directly modeled by a Gaussian distribution.

Such a transformation is given in Milios et al. 2018, where the authors propose a methodology for multi-label classification by regressing on transformed labels. Each borehole observation $\boldsymbol{y}_{bh}$ can be viewed as a draw from categorical distribution $Cat(\boldsymbol{\pi})$. Considering C different USCS labels observed in the training borehole data, the class probabilities $\boldsymbol{\pi} = (\pi_1,...,\pi_C)$ are treated as random variables. The joint probability distribution of $\boldsymbol{\pi} = (\pi_1,...,\pi_C)$ can be modeled by a *Dirichlet* distribution $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_C)$ are the concentration parameters. Given a binary observation vector $\boldsymbol{y}_{bh}$, at a location $\boldsymbol{s}$ within a borehole, the components of $\boldsymbol{\alpha}$ take values:

$$\alpha_i = \begin{cases} 1 + \alpha_\epsilon, & y_{bh,i} = 1 \\ \alpha_\epsilon, & y_{bh,i} = 0 \end{cases} \quad (9)$$

for $i = 1,...,C$, where $0 < \alpha_\epsilon \ll 1$ is a small quantity that ensures valid definition of the Dirichlet distribution. For a vector $\boldsymbol{y}_{bh}$ of borehole USCS observed labels, the categorical likelihood is given by:

$$p(\boldsymbol{y}_{bh}|\boldsymbol{\alpha}) = Cat(\boldsymbol{\pi}), \quad \boldsymbol{\pi} \sim Dir(\alpha). \quad (10)$$

A sample π_i from the Dirichlet distribution can be drawn from samples of C independent *Gamma* distributed random variables with shape parameters α_i given in Eq. (9) and rate $\lambda = 1$,

$$\pi_i = \frac{x_i}{\sum_{c=1}^{C} x_c}, \qquad x_i \sim Gamma(\alpha_i, 1). \quad (11)$$

The authors in (Milios et al. 2018) propose the approximation of the Gamma marginals with the *lognormal* distribution,

$$\tilde{x}_i \sim Lognormal(\tilde{y}_i, \tilde{\sigma}_i^2) \quad (12)$$

where $(\tilde{y}_i, \tilde{\sigma}_i^2)$ correspond to the parameters of the underlying normal distribution. Through moment matching $\mathrm{E}[x_i] = \mathrm{E}[\tilde{x}_i]$ and $\mathrm{Var}[x_i] = \mathrm{Var}[\tilde{x}_i]$, the parameters of the underlying normal distribution are derived as:

$$\tilde{y}_i = \ln(\alpha_i) - \tilde{\sigma}_i^2/2, \quad \tilde{\sigma}_i^2 = \ln(1/\alpha_i + 1). \quad (13)$$

Consequently, $\tilde{y}_i$ can now represent the transformed to the Gaussian space observation, and $\tilde{\sigma}_i^2$ the noise variance associated with the respective observation. Note that there is an additional parameter that needs to be determined to complete the transformation of the USCS data to the Gaussian space, and this is the small quantity α_ϵ appearing in Eq. (9). The authors in (Milios et al. 2018) discuss on how to select this parameter, or estimate it as an additional hyperparameter of the model. Here, we fix α_ϵ to a small value. I_c is assumed to be lognormal distributed as well, as it can only take positive values, such that $Y_{I_c} = \ln(I_c)$ follows the Normal distribution.

The important advantage of the approximate transformation to the likelihood of the USCS data is that it can be now assumed that $\boldsymbol{f} = (f_1, ..., f_C, f_{I_c})$ are jointly Gaussian and can be modeled with the multivariate Gaussian Process Regression model described in the previous section. The estimates of the hyperparameters obtained by maximizing the marginal likelihood given in Eq. (6) are now informed by both borehole and CPT data. The estimate of the cross covariance matrix $\boldsymbol{\Sigma}$ can define the covariance between the USCS labels themselves and between I_c.

Turning to the predictions at some new locations within the site, the predictive distribution $\boldsymbol{f}_*$ given in Eq. (7) produces in total $C + 1$ outputs, i.e., C predictions related to the USCS classes, plus predictions of $\ln(I_c)$. It is possible to derive analytically the predictive mean and variance of I_c. The predictive mean of the class probabilities is given by:

$$\mathrm{E}[\pi_{i,*}|\boldsymbol{y}] = \int \frac{\exp(f_{i,*})}{\sum_j^C \exp(f_{j,*})} p(f_{i,*}|\boldsymbol{y})\, d\boldsymbol{f}_*, \qquad (14)$$

where for convenience the dependence on $\boldsymbol{S}_*$ and $\boldsymbol{S}$ has been omitted. The integral in Eq. (14) can be approximated with samples from the posterior predictive distribution of each class. We first generate N Normal distributed samples from $p(f_*|\boldsymbol{S}_*, \boldsymbol{S}, \boldsymbol{y})$, and we approximate the posterior probability of the classes through:

$$\mathrm{E}[\pi_{i,*}|\boldsymbol{y}] \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\exp(f_{i,j,*})}{\sum_j^C \exp(f_{i,j,*})}. \qquad (15)$$

4. Application to a New Zealand site (Christchurch)

A site at Christchurch, New Zealand (Wang and Zhu 2023) is adopted to demonstrate the proposed approach. The site plan including the borehole logs and CPT soundings is illustrated in Fig. 1.

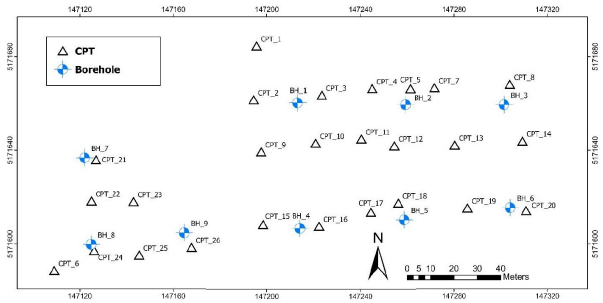

Figure 1. Christchurch site (extracted from Wang and Zhu 2023).

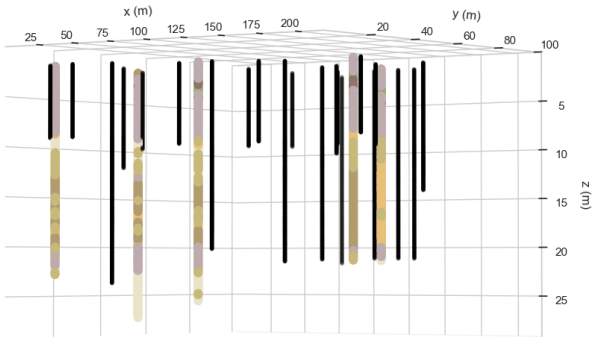

Figure 2. 3D illustration of the selected boreholes and CPTs; black lines: CPT soundings; colored lines: borehole data.

The site investigation plan includes 26 CPT soundings with original depth increment $0.01\mathrm{m}$ and 9 boreholes. 5 out of 9 boreholes were used for training the model (BH_2, BH_4, BH_6, BH_7, BH_8), while the remaining 4 boreholes (BH_1, BH_3, BH_5, BH_9) were reserved for validation. The maximum depth range of the training borehole data is $21.45 - 27.45\mathrm{m}$. Out of the 26 available CPTs, the deepest 21 were used for training, with maximum depth range of approximately $7 - 24\mathrm{m}$. Furthermore, the upper $1.86\mathrm{m}$ were disregarded due to very unstable CPT measurements. Fig. 2 provides a 3D illustration in space of the selected boreholes and CPTs. It is important to note that only 8 out of the 21 selected CPT soundings reach depths similar to those of the borehole logs. The data preprocessing steps can be summarized in the following:

- Transformation of the coordinates to the unit cube $[0, 1]^3$
- Derivation of I_c from q_c, f_s and pore pressure u_2 measurements (see Robertson 2009, 2016)
- Selection of 250 equidistant points along depth of the deepest borehole. Note that the same depth increment was considered for the boreholes and CPTs,

which means that there are different numbers of collected observations from each of the CPT soundings and borehole logs. There are in total 2614 locations yielding observations of I_c in the training data. The I_c was assumed to follow the lognormal distribution, such that the log-transformed $Y_{I_c} = \ln(I_c)$ is normal distributed. Furthermore, the observations were standardized, to ensure that the variables have similar scales.

- Among the selected borehole locations, there are 7 labels observed: GW-"well-graded gravel" (139), SP-"poorly-graded sand" (145), SW-"well graded sand" (147), SM-"silty sand" (231), ML-"silt of low plasticity" (350), OL-"organics of low plasticity" (25) and Pt-"peat" (34), with the numbers in the parentheses corresponding to the number of locations where the respective soil class was observed.
- One-hot encoding: each location within the boreholes is associated with a vector of 7 components, with values of "1" denoting presence and "0" denoting absence of each USCS class. There are in total 1071×7 binary observations from all boreholes. A small value was assigned to the quantity α_ϵ(=0.002) and from Eq. (9), (13) the categorical data were transformed to noisy Gaussian observations $\tilde{y}_i$, with each observation associated with noise variance $\tilde{\sigma}_i^2$

The Gaussian transformed data were collected in a 10111×1 column vector $\boldsymbol{y}$. The single exponential model given in Eq. (3) was selected to model the spatial correlation. The 10111×10111 diagonal matrix $\boldsymbol{D}$ was constructed from the variances $\tilde{\sigma}_i^2$ of the borehole observations, and a noise variance for each y_{I_c} observation. For the latter, a fixed small value ($\sigma^2_{y_{I_c}} = 10^{-5}$) was assumed for each observation. All computations were performed on a commercial laptop CPU (Intel Core i5; 16GB; 2.40GHz).

Estimates of the hyperparameters $\{\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma}\}$ were obtained by minimizing the negative marginal log-likelihood of Eq. (6). The estimates of the vertical and horizontal correlation lengths were $\boldsymbol{\theta} = [1.63\text{m}, 35.45\text{m}]$. The running time for the optimization, with 150 steps of the Adam optimizer was $\approx 40\text{min}$.

Predictions at the reserved boreholes were performed using Eq. (7), (8) and (15). To assess the performance of the model, prediction results for borehole 9 are illustrated in Fig. 3. Starting from the left, the first figure displays the reported classification from borehole log 9. The second shows the most probable classification profile derived from the multivariate GP model and the third exhibits the corresponding predicted probabilities. Different colors and styles are assigned to each USCS soil class to allow easier interpretation. The fourth figure illustrates prediction results for the Soil Behavior Type Index I_c including the 95% confidence intervals, while the last one presents the (standardized) Information Entropy, serving as a measure for uncertainty in the USCS classification prediction. The running time for predictions at one borehole was $\approx 30\text{s}$ (for 2000 locations, 1000 samples for USCS probabilities).

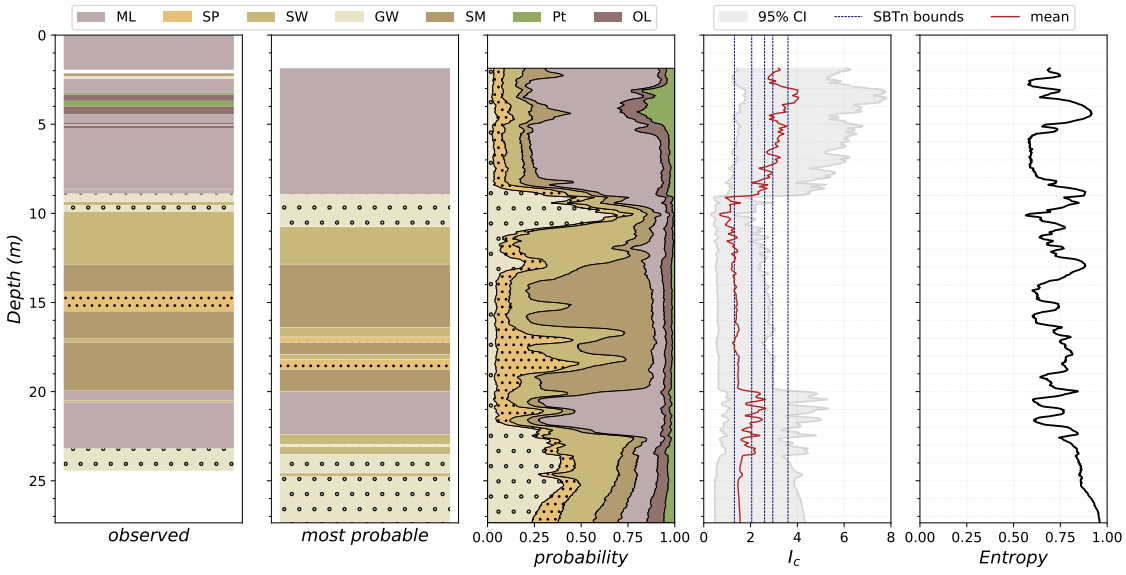

Figure 3. Borehole 9: predictions.

Examining from the top and descending to 9 meters, the reported borehole data highlight ML as the dominant USCS soil class. This dominance is captured in the predicted classification results, evident in the second figure, with a higher probability of ML in the probability profiles of the third figure. The predicted mean of I_c at these depths falls mainly within the SBT categories of clay and silt mixtures, although the prediction uncertainty is considerably high, with the 95% CIs covering a large portion of the regions defined by the SBT categories bounds. The borehole data reveal thin crusts, primarily around depths of 3.5-4.2 meters, comprising Pt and OL soils, which are not visible in the second graph. However, inspecting the probability profiles reveals increased corresponding class probabilities at these depths, accompanied by higher entropy values, indicating higher uncertainty in the USCS predictions. Notably, the mean of the I_c predictions successfully detects the thin layers of organic soils and peat, although the 95% CI is broader at these depths.

The USCS predictions successfully capture the transition to GW at depth equal to 9 meters, with high probability associated to GW and lower entropy. This transition is also depicted in the I_c prediction graph. According to the borehole reported data the GW layer thickness is around 1m, and at 10 meters the profile changes to SW soil which extends up to 13m. The predicted GW layer extends to around 11m with a smoother transition to SW soil as indicated by the probability profiles.

The dominant USCS class down to 20m depth is SM, which is also suggested by the USCS prediction results, with higher probability of this class in most of these depths. The borehole reported data indicate a thin layer of SP soil at depths around 15m. This is not evident in the second figure. Looking at the probability plots there is an indication of a rise in the probability of the SW class, as well as a rise in the entropy, although the highest probability is associated again with the SM class. The prediction results indicate that there are some thin SP layers at lower depths, not evident at the observed borehole data. One can argue that the soils from 10 to 20m are mixed. Although the predominant soil is sand, the fines content affect the classification. It is worth mentioning that according to the USCS some soils may be assigned double classification, such as SP-SM, which may be the case here. This behavior may be also explained by the I_c predictions plot. The predicted mean falls on the border between SBT classes 6 and 7, although the the upper confidence bound extends to SBT class 5.

Beyond 20 meters depth, and down to 23 meters, the reported borehole data signify a transition to ML soil, a shift effectively identified by the USCS prediction. The predicted mean of I_c primarily aligns with SBT categories 5 and 6, although the confidence interval is wider, suggesting potential inclusion of mixed silty and sandy soils. The USCS prediction reveals crusts of SW soil, absent in the reported borehole data. Below 23 meters, the soil transforms to GW both in the observed and predicted profiles, with the borehole reaching depths of approximately 24.4 meters. Below these depths, the predictions exhibit lower probabilities even for the dominant class, accompanied by high entropy and a wide confidence interval in the I_c predictions.

Interpretation of the predicted profiles was performed to all reserved boreholes. Transitions to ML and GW layers were successfully detected in all testing boreholes, as well as embedded soils within the predominant sand layer, consistently observed across all reserved boreholes from approximately 10 meters to 20 meters. A challenge was encountered in detecting thin layers of SP when the dominant layer was SM, and in few cases thin layers of SW when the main soil layer was SM. It is reminded that this difficulty may be attributed to dual classification soils (SP-SM, SW-SM). Overall, the proposed methodology demonstrated accurate prediction of the USCS classification even for boreholes located at a considerable distance from other boreholes and CPTs.

5. Concluding remarks

A methodology for joint stratigraphy prediction from sparse borehole and CPT data was proposed in this study. The key aspect of this methodology lies in its capability to incorporate both numerical CPT and categorical USCS data. This integration is achieved by viewing the USCS labels as Dirichlet-distributed random variables, and approximately transforming these variables to the Gaussian space. This allows to jointly model the transformed Gaussian variables and CPT parameters, such as I_c, with a multivariate Gaussian Process Regression model.

The main assumption of the GP model is the separable structure between spatial and cross-correlation, which implies that the USCS related variables and the modeled CPT parameter, share the same spatial correlation structure. The hyperparameters of the model were estimated by maximization of the marginal likelihood. Predictions of I_c and USCS classification profiles were performed at the locations of the verification boreholes. The results suggested that the USCS predicted classification was in good agreement with the borehole reported data, with major transitions and embedded layers successfully identified.

Despite the emphasis on I_c in this study, the proposed methodology enables the joint modeling of multiple CPT

parameters and categorical geological variables. Consequently, it becomes feasible to obtain predictions for both mechanical parameters and reliable stratification within a unified regression-classification framework, at a reasonable computational cost, with an approach that relies solely on site-specific data. Importantly, the proposed approach can be viewed as a baseline model that can be further extended into a fully Bayesian Gaussian Process model. This extension would enable quantifying the uncertainty in the hyperparameters.

References

Alvarez, M. A., Rosasco, L., Lawrence, N. D. "Kernels for Vector-Valued Functions: A Review.", Foundations and Trends in Machine Learning, 4(3), pp. 195-266, 2012. http://dx.doi.org/10.1561/2200000036

Bonilla, E. V., Chai, K., Williams, C. "Multi-task Gaussian Process Prediction.", In: *Advances in Neural Information Processing Systems*, 2007.

Chen, T., Guestrin, C. "XGBoost: A scalable tree boosting system." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. https://doi.org/10.1145/2939672.2939785

Ching, J., Yang , Z., Phoon, K.K "Dealing with Nonlattice Data in Three-Dimensional Probabilistic Site Characterization" Journal of Engineering Mechanics, 147(5), pp. 06021003, 2021. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001907

Ching, J., Yoshida, I. "Data-drive site characterization for benchmark examples: Sparse Bayesian learning versus Gaussian process regression." ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 9(1), 2023. https://doi.org/10.1061/AJRUA6.RUENG-983

Ching, J., Phoon, K.K., Yang, Z., Stuedlein, A. W. "Quasi-site-specific multivariate probability distribution model for sparse, incomplete, and three-dimensional spatially varying soil data." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), pp. 53-76, 2022. https://doi.org/10.1061/AJRUA6.RUENG-983

Das, B. M. "Principles of Foundation Engineering", 8th ed., Cengage Learning, Boston, USA, 2015.

Elfeki, AMM., Dekking, FM. "A Markov chain model for subsurface characterization: theory and applications." Mathematical Geology, 33(5), pp. 569-589, 2001. https://doi.org/10.1023/A:1011044812133

Farahbakhsh , H. K., Ching, J. "Inferring Spatial Variation of Soil Classification by Both CPT and Borehole Data" In: *Geo-Risk 2023*, pp. 142-151, 2023. https://doi.org/10.1061/9780784484975.016

Hu, Y., Wang, Y. "Probabilistic soil classification and stratification in a vertical cross-section from limited cone penetration tests using random field and Monte Carlo simulation." Comput. Geotech. 124, pp. 103634, 2020. https://doi.org/10.1016/j.compgeo.2020.103634

Kingma , D. P., Ba, J. "Adam: A Method for Stochastic Optimization" In: *3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015. https://doi.org/10.48550/arXiv.1412.6980

Li, Z., Wang, X., Wang, H. and Liang, R. Y. "Quantifying stratigraphic uncertainties by stochastic simulation techniques based on Markov random field." Engineering Geology, 201, pp. 106–122, 2016. https://doi.org/10.1016/j.enggeo.2015.12.017

Milios, D., R. Camoriano, P. Michiardi, R. Lorenzo, and M. Filippone. "Dirichlet-Based Gaussian Processes for Large-Scale Calibrated Classification.", In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, Montréal, Canada, pp. 6008-6018, 2018.

Qi, X.H., Li, D.Q., Phoon, K.K., Cao, Z.J., Tang, X.S. "Simulation of geologic uncertainty using coupled Markov chain." Engineering Geology, 307, pp. 129-140, 2016. https://doi.org/10.1016/j.enggeo.2016.04.017

Robertson, P.K. "Interpretation of cone penetration tests - a unified approach." Can. Geotech. J. 46 (11), pp. 1337–1355, 2009. https://doi.org/10.1139/t09-065

Robertson, P.K. "Cone penetration test (CPT)-based soil behaviour type (SBT) classification system-an update." Can. Geotech. J. 53 (12), pp. 1910–1927, 2016.

Shakir, R.R., Wang, H. "Estimation of probabilistic CPT-based soil profile using an unsupervised Gaussian mixture model.", Arab J Geosci 16, 218, 2023. https://doi.org/10.1007/s12517-023-11283-7

Shi, C., Wang, Y. "Development of subsurface geological cross-section from limited site-specific boreholes and prior geological knowledge using iterative convolution XGBoost". Journal of Geotechnical and Geoenvironmental Engineering, 149(9), pp. 04021082, 2021. https://doi.org/10.1061/(ASCE)GT.1943-5606.0002583

Shuku, T. and Phoon, K.-K. "Data-driven subsurface modelling using a Markov random field model." Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 17(1), pp. 41–63, 2023. https://doi.org/10.1080/17499518.2023.2181973

Shuku, T. and Phoon, K.-K. "Three-dimensional subsurface modelling using Geotechnical Lasso." Computers and Geotechnics, 133, pp. 1034068, 2021. https://doi.org/10.1016/j.compgeo.2021.104068

Wang, H., Wang, X., Wellmann, J. F., Liang. R.Y. "A Bayesian unsupervised learning approach for identifying soil stratification using cone penetration data." Can. Geotech. J. 56 (8): pp. 1184–1205, 2019. https://doi.org/10.1139/

cgj-2017-0709

Wang, H., Wellmann, J. F., Li, Z., Wang, X. and Liang, R. Y. "A segmentation approach for stochastic geological modeling using hidden Markov random fields." Mathematical Geology, 49(2), pp. 145–177, 2017. https://doi.org/10.1007/s11004-016-9663-9

Wang, H. and Zhu, Y. Georisk2023: "3D geological modeling using CPT data" Kaggle, 2023. [online] Available at: https://kaggle.com/competitions/georisk2023-3d-geological-modeling-using-cpt-data, accessed: 01/11/2023

Wei, X. and Wang, H. "Stochastic stratigraphic modeling using Bayesian machine learning." Engineering Geology, 307, pp. 106789, 2022. https://doi.org/10.1016/j.enggeo.2022.106789

Xiao, T., Zou, HF., Yin, KS., Du, Y. and Zhang, LM. "Machine learning-enhanced soil classification by integrating borehole and CPTU data with noise filtering." Bull Eng Geol Environ, 80, pp. 9157–9171, 2021. https://doi.org/10.1007/s10064-021-02478-x

Zhang, J.-Z, Liu, Z.-Q, Zhang, D.-M, Huang, H.-W, Phoon, K.-K, Xue, Y.-D. "Improved coupled Markov chain method for simulating geological uncertainty." Engineering Geology, 298, pp. 106539, 2022. https://doi.org/10.1016/j.enggeo.2022.106539