Stratification identification and prediction of missing CPT data by Mixture of Gaussian Processes

Durmaz, Muhammet; van den Eijnden, Abraham P.; Hicks, Michael A.

DOI
10.23967/isc.2024.020

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 7th International Conference on Geotechnical and Geophysical Site Characterization

Stratification identification and prediction of missing CPT data by Mixture of Gaussian Processes

Muhammet Durmaz[1#], *Abraham P.* van den Eijnden[1], and *Michael A.* Hicks[1]

[1]*Department of Geoscience and Engineering, Delft University of Technology, Delft, The Netherlands*
[#]*Corresponding author: M.durmaz@tudelft.nl*

ABSTRACT

Stratification identification and spatial interpolation play a fundamental role in geotechnical site characterization. A unified approach is needed to perform these two tasks simultaneously to reduce overall uncertainty in site characterization. This paper explores the applicability of the Mixture of Gaussian Processes (MoGP) to address this gap, with a specific focus on characterizing and completing missing CPT data. The investigation encompasses both synthetic and real-world field CPT datasets and includes a comparison of the MoGP's interpolation accuracy with the use of a single GP for entire datasets. Additionally, the study examines the sensitivity of the model's performance with respect to the number of training data points. Although the interpolation performance of the MoGP model is promising with synthetic data, limitations appear in its application to real-site CPT data.

Keywords: Mixture of Gaussian Processes; spatial variability; stratification; uncertainty

1. Introduction

Stratification identification and interpolation are essential components of data-driven site characterization. Identifying stratification involves recognizing layers with distinct behaviors and all associated geometric features, including boundaries, discontinuities, and anomalies (Phoon, Ching, and Shuku 2022). To address this challenge, several methods have been proposed (Depina et al. 2016; Cao and Wang 2014; Cao et al. 2019; Wang et al. 2018). Spatial interpolation, on the other hand, aims to fill in missing data arising from sensor failure or sparse geotechnical data. The point and spatial statistics of underground attributes, such as mean function (or trend), standard deviation, and scale of fluctuation, are used when completing missing data. Therefore, these parameters should also be determined during site characterization.

While soil layers exhibit distinctive behaviors, they also possess unique statistical properties. For example, layers of clay, sand, and silt often vary in their scale of fluctuation (Phoon et al. 2022). Therefore, stratification refers to both behavioral change and non-stationarity. Although some methods can estimate both stratification and point and spatial statistics (Cao and Wang 2014; Cao et al. 2019), they do not provide any method to complete missing data. Completing missing data in the presence of heterogeneity is challenging because both feature values (such as soil behaviour type index, I_c) and layer information are missing at unobserved locations. Therefore, the statistic to be used for prediction is unknown. Some studies have attempted to tackle this problem with two-step approaches, which involve interpolating data with stationary assumptions and then deterministically assigning each unobserved point to the layers (Ching and Yoshida 2023; Mavritsakis et al. 2023).

This paper explores the application of the Mixture of Gaussian Processes to address both stratification identification and the completion of missing geotechnical data. The proposed method also considers the uncertainty in layer boundaries during data interpolation. The accuracy of the approach is evaluated using both synthetic and real Cone Penetration Test (CPT) data.

2. Mixture of Gaussian Processes

The Mixture of Gaussian Processes (MoGP) model proposed by Chen, Ma, and Zhou (2014) is adopted in this paper. The MoGP model comprises two primary components: Gaussian processes (GPs) in feature space (I_c) and Gaussian distributions (GDs) in input space (depth). In our formulation, the GP serves as the regression model, while the GD acts as the probabilistic allocation model determining the selection of GPs at specific locations. In other words, the GD indicates the likelihood of a layer's presence at a specific location.

GPs can be described as the generalization of probability distributions to functions (Rasmussen and Williams 2006). According to this definition, functions estimating the outputs are drawn from a multivariate Gaussian distribution characterized by a covariance matrix C, so that:

$$p(\boldsymbol{y}|\boldsymbol{x}) \sim N[m(\boldsymbol{x}), C(\boldsymbol{x},\boldsymbol{x})] \quad (1)$$

where $\boldsymbol{y} = [y_1, y_2, y_3, \dots y_n]$ is a collection of observations, $\boldsymbol{x} = [x_1, x_2, x_3, \dots x_n]$ is the input vector, $\boldsymbol{m}(\boldsymbol{x}) = [m(x_1), m(x_2), m(x_3), \dots m(x_n)]$ is the vector of mean values of $\boldsymbol{y}$ and $C(\boldsymbol{x},\boldsymbol{x}) = [c(x_i, x_j)]_{n\times n}$ is the covariance matrix. This study adopts a single

exponential autocorrelation function; therefore, the elements of the covariance matrix are defined as:

$$c(x_i, x_j) = \theta_0 \exp\left(-\frac{\theta_1}{2}|x_i - x_j|\right) \quad (2)$$

where θ_0 and θ_1 represent the variance and correlation length of y, respectively.

When accounting for noise in the observations, the noise variance should be included in the covariance matrix in Eq. (1). However, this paper does not consider such noise. Consequently, each Gaussian process is characterized by two hyperparameters, $\theta^{GP} = \{\theta_0, \theta_1\}$. The maximum likelihood estimation is typically employed for estimating these hyperparameters.

Once the hyperparameters are estimated, the predictive distribution of y at a new input x^* remains Gaussian, and is given by:

$$p(y|x^*) \sim N[y^*, V(y^*)] \quad (3)$$

where y^* is the mean prediction, and $V(y^*)$ is the variance of the prediction. These values are calculated using the following formulae, as outlined by Rasmussen and Williams (2006):

$$y^* = C(x^*, \boldsymbol{x})C(\boldsymbol{x}, \boldsymbol{x})^{-1}[\boldsymbol{y} - m(\boldsymbol{x})] + m(x^*) \quad (4)$$

$$V(y^*) = C(x^*, x^*) - C(x^*, \boldsymbol{x})C(\boldsymbol{x}, \boldsymbol{x})^{-1}C(\boldsymbol{x}, x^*) \quad (5)$$

The Mixture of Gaussian Processes (MoGP) can be conceptualized as a weighted mixture of multiple Gaussian processes. A gating function - represented by a Gaussian distribution - is employed to combine these Gaussian processes across the input region. The overall model is described by Chen, Ma, and Zhou (2014) as follows.

Firstly, given a data set $\{x_n, y_n\}_{n=1}^N$, the latent indicators $\{z_n\}_{n=1}^N$, which indicate from which Gaussian process the data originates, are assumed to follow a multinomial distribution:

$$P(z_n = k) = \pi_k \quad (6)$$

Given the indicator variable, each input has a Gaussian distribution:

$$p(x_n|z_n = k) \sim N(\mu_k, s_k^2) \quad (7)$$

where μ_k is the mean and s_k^2 is the variance. After specifying the input set with the indicator $\{z_n, x_n\}_{n=1}^N$, each observation fulfils the properties of a GP such that:

$$p(\boldsymbol{y^k}|\boldsymbol{x^k}) \sim N[\boldsymbol{m}(\boldsymbol{x^k}), C(\boldsymbol{x^k}, \boldsymbol{x^k}|\theta^{GP,k})] \quad (8)$$

where $\boldsymbol{x^k}, \boldsymbol{y^k}$ denote the input and output vectors that are considered to be part of the k^{th} component.

2.1. Hard-cut EM algorithm

This model requires learning five hyperparameters $\{\pi_k, \mu_k, s_k, \theta_0^k, \theta_1^k\}$ for each component. To achieve this, the hard-cut Expectation Maximization (EM) algorithm, as implemented by Chen, Ma, and Zhou (2014), is utilized. The EM algorithm enables the maximization of the likelihood function for problems involving hidden variables. During training, the parameters are updated iteratively through two steps: expectation and maximization. In the expectation step, the complete likelihood is calculated based on the posterior probability of the latent variable. Subsequently, in the maximization step, the model parameters are updated to maximize the likelihood. The complete likelihood of the current model is expressed as follows:

$$p(z_n = k, x_n, y_n) = \pi_k N(x_n|\mu_k, s_k) N(y_n|m(x_n), \theta_0^k) \quad (9)$$

The posterior probability of the latent variable is calculated by:

$$P(z_n = k|x_n, y_n) = \frac{p(z_n=k, x_n, y_n)}{\sum_{k=1}^{k=K} p(z_n=k, x_n, y_n)} \quad (10)$$

where K is the number of components (i.e. the number of GPs).

Subsequently, the hard-cut EM algorithm is implemented as follows:

1. Firstly, clusters are initialized by using K-means clustering. The indicator variable z_n is initialized by assigning the cluster index of each point.
2. M-step: μ_k and s_k are calculated in the same way as in the Gaussian Mixture model as follows:

   $$\mu_k = \frac{\sum_{n=1}^{n=N} I(z_n=k) x_n}{\sum_{n=1}^{n=N} I(z_n=k)} \quad (11\text{-a})$$

   $$s_k = \frac{\sum_{n=1}^{n=N} (z_n=k)(x_n-\mu_k)(x_n-\mu_k)^T}{\sum_{n=1}^{n=N} I(z_n=k)} \quad (11\text{-b})$$

   where $I(.)$ is the indicator function, which equals 1 if $z_n = k$, otherwise 0. In contrast to the original algorithm (Chen, Ma, and Zhou 2014), π_k is calculated as $1/K$ in this paper. Subsequently, the GP parameters are obtained by maximum likelihood estimation after subtracting the mean of each cluster from the data within that cluster.
3. E-step: Each sample is classified according to the maximum a posteriori probability (MAP) criterion as follows:

   $$z_n = \underset{k}{\text{argmax}}[\pi_k N(x_n|\mu_k, s_k) N(y_n|m(x_n), \theta_0^k)] \quad (12)$$

Steps 2 and 3 are repeated until the indicators no longer change with increasing iteration number.

2.2. Prediction

This paper diverges from the prediction strategy outlined by Chen, Ma, and Zhou (2014). While they suggested making predictions using the Gaussian process with the maximum posterior probability at the prediction point, this paper adopts a complete probabilistic approach for the prediction.

In a single GP, the prediction at a new input location follows a Gaussian distribution as shown in Eq. (5) and (6). However, in the Mixture of Gaussian Processes (MoGP), the predictive distribution becomes a mixture of Gaussians, as it is a weighted combination of the predictions made by each GP. Therefore, the predictive distribution can be expressed as:

$$p(y_n^*|x_n^*) = \sum_{i=k}^{i=K} P(z_n = k|x_n^*) N\left[y_{n,k}^*, V\left(y_{n,k}^*\right)\right] \quad (13)$$

where $y_{n,k}^*$ and $V\left(y_{n,k}^*\right)$ represent the mean and variance predicted by an individual GP (Eq. (4) and (5)). $P(z_n = k|x^*)$ is the posterior probability of the indicator variable of the new input and is calculated as follows:

$$P(z_n = k|x^*) = \frac{\pi_k N(x^*|\mu_k, s_k)\alpha_{k,x^*}}{\sum_{k=1}^{k=K} \pi_k N(x_n|\mu_k, s_k)\alpha_{k,x^*}} \quad (14)$$

The term α_{k,x^*} is introduced to Eq. (14) in this paper. The motivation for incorporating this term is that the calculation of the posterior probability of the latent variable in the training stage (Eq. (12)) includes both similarity in input space, $\{\pi_k N(x_n|\mu_k, s_k)\}$, and similarity in output space, $\{N\left(y_n|m(x_n), \theta_0^k\right)\}$. However, in the original algorithm by Chen, Ma, and Zhou (2014), the probability of the latent variable at a prediction point is based only on the input space, which often leads to unreasonable results. On the other hand, since the prediction y_n^* is unknown, Eq. (12) cannot be directly utilized here. Thus, this paper includes similarity in the output space in a different way through the incorporation of α_k.

For a deeper understanding, readers can first refer to Fig. 2b. When a single GP is utilized for the entire profile, nearby points have a greater influence on the prediction, although the GP is trained by the entire dataset. Consequently, the predictive distribution in the middle of a layer resembles the distribution obtained by a GP trained solely on data from that layer. In contrast, in the transition zone between two layers (Fig. 2b), the GP's prediction distribution evolves to become more similar to the distribution obtained by data from the layer it is approaching (Fig. 2a). The overlapping area is one way to quantitatively measure the similarity between the distributions (Inman and Bradley 1989). Therefore, α_k is calculated to be proportional to the overlapping area of each GP's predictive distribution with the single GP's predictive distribution, as illustrated in Fig. 1.

In addition, the mean prediction of the MoGP, $\bar{y}^*$, is equal to $\sum_{i=k}^{i=K} P(z_n = k|x^*) y_k^*$.

2.3. Implementation

2.3.1. Synthetic Data

The MoGP model is tested using a synthetic dataset under two scenarios. The dataset represents the Soil Behavior Type Index (I_c) (Robertson, 2016) values of a 3-layer soil profile. In the first scenario, the layer thicknesses are considered equal, whereas they differ in the second scenario. The layer properties are provided in Table 1. Correlation lengths for the layers were based on the literature (Phoon et al. 2022; Shuku and Phoon 2023). The variance values were chosen to avoid significant overlap between the I_c distributions of the layers. For the second scenario, the layer depths are shown in Table 1 in parentheses.

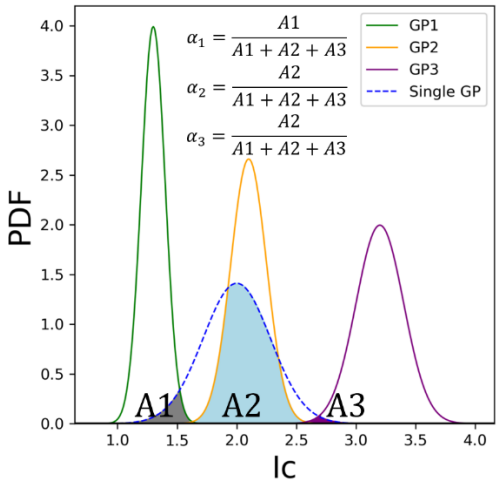

Figure 1. Illustration of the calculation of α_k. The distributions show the predictive distributions at a test point obtained by each GP component of the MoGP and the GP trained with all data.

To assess the accuracy of parameter estimation for the models, 50 datasets were generated for each scenario with a sampling interval of 0.05 m. Subsequently, the Mixture of Gaussian Processes (MoGP) with three components was trained using all available data. During training, θ_0 was constrained between half and twice the variance of the data, while θ_1 was constrained between 0.1 m and 10 m. The means of the estimated parameters for each scenario are presented in Table 2.

The mean values of the estimated parameters closely align with the values used to generate the data. It is anticipated that this difference will decrease with an increase in the number of datasets. Additionally, Table 2 indicates that the accuracy of the parameter estimation is unaffected by the layer thicknesses, as the predicted parameters are nearly identical in both scenarios.

In the equal thickness scenario, all data points are correctly classified into their respective layers, while the mean misclassification ratio is only 0.06% in the case of different thicknesses.

To evaluate the interpolation performance of the MoGP model, each dataset was divided into training and testing groups. The training sizes ranged from 10% to 80% of the total data, to investigate whether the performance varied with the number of training data. Additionally, the number of components of the MoGP was changed from one to five for each simulation.

Fig. 2 compares the predictions made by a single GP (Fig. 2b) and the MoGP with 3 components (Fig. 2c). The first notable observation is that the MoGP exhibits a narrower confidence interval compared to the single GP. This is attributed to the fact that the MoGP utilizes only data from the same layer to make predictions, thereby reducing the variability in the data and resulting

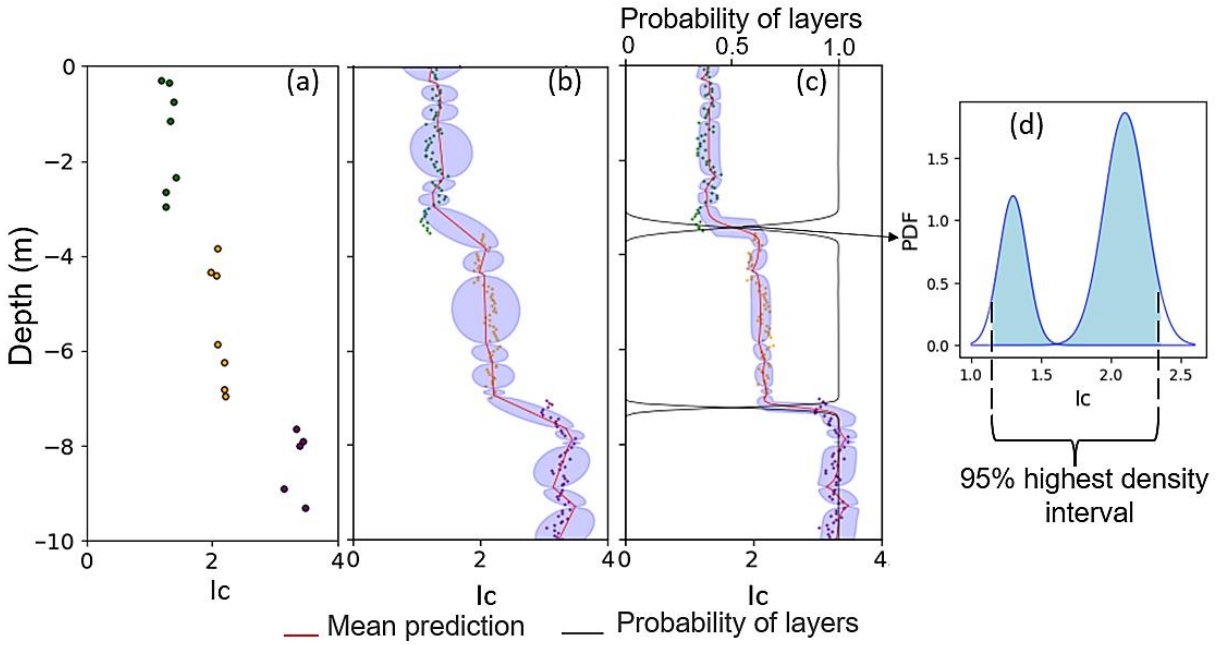

Figure 2. An example of I_c estimates for the equal depth scenario with training data rate equal to 10%. (a) Training data; (b) single GP estimations; (c) 3-component MoGP estimations; (d) 95% maximum intensity interval in the transition regions between layers. Points in (b) and (c) are test points. The shaded area shows the 95% highest intensity range, which narrows at the training points.

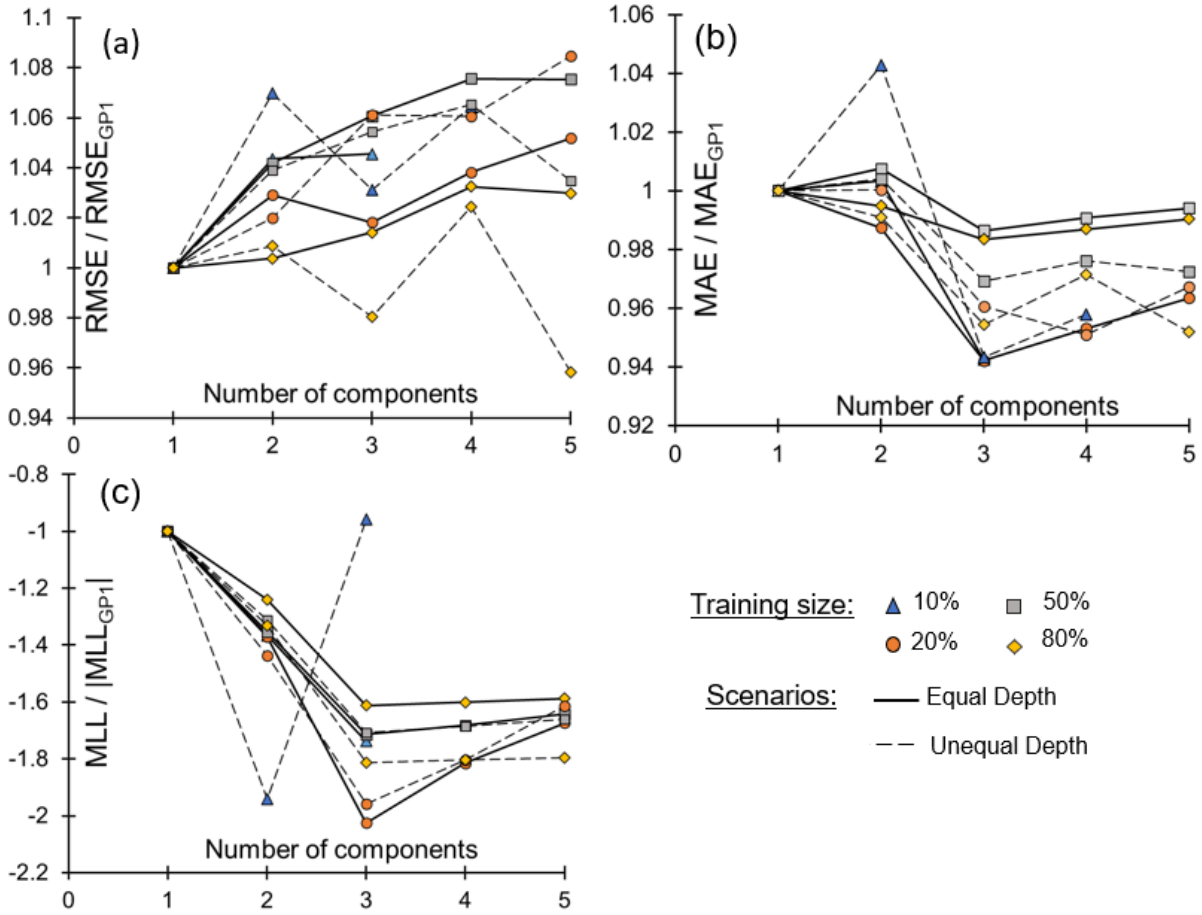

Figure 3. Comparison of interpolation accuracy of the MoGP with different numbers of components.

in less uncertainty in the prediction. However, this reduced uncertainty should be validated with test data.

Table 1. Parameters of synthetic data

	Thickness (m)	Mean (Ic)	θ_0	θ_1 (m)
Layer 1 (sand-gravel)	3.5 (5)	1.3	0.01	0.4
Layer 2 (sand-silt)	3.5 (2)	2.1	0.0225	0.8
Layer 3 (clay)	3.5 (3.5)	3.2	0.04	1.2

Table 2. Means of estimated parameters from 50 simulations. S1 and S2 denote equal thickness and different thickness scenarios respectively.

	Mean (I_c)	θ_0	θ_1 (m)	μ_k(m)	s_k(m)
S1	1.30	0.0094	0.39	1.75	1.065
S2	1.30	0.0094	0.39	2.50	2.141
S1	2.10	0.0169	0.62	5.28	1.035
S2	2.10	0.0210	0.71	6.02	0.347
S1	3.21	0.0344	1.03	8.80	1.065
S2	3.21	0.0344	1.02	8.80	1.064

The interpolation performances of models with different numbers of components are compared using three different metrics. For performance comparison, the root mean square error (RMSE) and the mean absolute error (MAE) are widely used performance metrics and are employed in this study. However, these metrics may not be sufficient for assessing the performance of Gaussian Processes (GPs) and the MoGP models, as the prediction is represented by a distribution. Therefore, the mean log probability loss (MLL), as proposed by Rasmussen and Williams (2006), is also utilized in this study. MLL is calculated by Eq. (14) and Eq. (15):

$$LL_{i,k} = -\log \mathrm{p}\left(y_{i,k}^{*}\middle|x_i^{*}\right) = \frac{1}{2}\log\left(2\pi\sigma_{i,k}\right) + \frac{(y_i^T - y_{i,k}^{*})^2}{2\sigma_{i,k}^2} \quad (15)$$

$$MLL = \frac{\sum_{i=1}^{i=Nt}\sum_{k=1}^{k=K} P(z=k)LL_{i,k}}{N} \quad (16)$$

In Eq. (15) y_i^T is the test data, $y_{i,k}^{*}$ is the mean prediction made by the k^{th} GP and $\sigma_{i,k}$ is the variance of the prediction.

MAE and RMSE are calculated using the mean prediction, y^{*}.

The mean error metrics from 50 simulations are summarized in Fig. 3. For illustrative purposes, all errors are normalized according to the single GP. Generally, RMSE exhibits an increasing trend with an increasing number of components. Conversely, the MoGP with 3 components demonstrates the lowest mean error in almost all cases according to MAE and MLL, with a few exceptions (Fig. 3b-c). The lower MLL in the 3-component case can be interpreted as a validation of the reduced uncertainty in the MoGP.

The differing behaviors of MAE and RMSE can be attributed to the calculation methods. RMSE assigns more weight to points with higher error, while all points have equal weight in MAE. The mean prediction of the MoGP exhibits high curvature near layer boundaries (Fig. 2), resulting in some points having significantly deviated predictions compared to the real values, thereby amplifying the RMSE.

2.3.2. Real-site CPT data

The applicability of the MoGP model is tested on real Cone Penetration Test (CPT) data from the Groningen region in the Netherlands. Initially, the Soil Behavior Type Index (I_c) profile of the CPT data is created with a 0.1 m interval (Fig. 4a). In real CPT data, the number of components is unknown. However, the results from synthetic data suggest that MAE and MLL can potentially be used to determine the number of components in a cross-validation framework. Therefore, 10-fold cross-validation (CV) is utilized to decide on the number of components and to test the interpolation accuracy of the MoGP on real CPT data.

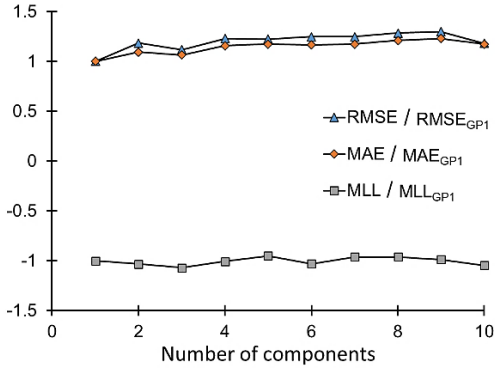

Figure 4. Mean error metrics of real CPT data calculated with 10-fold cross-validation.

The mean values of the error metrics calculated by 10-fold CV are illustrated in Fig. 4. Contrary to the synthetic data experiment, the MLL and MAE results are not consistent with each other. MAE and RMSE are at their lowest when a single GP is used, indicating that the MoGP did not improve the interpolation accuracy. On the other hand, MLL is at its lowest for the three-component model, although there is not much difference for any number of components. Consequently, determining the number of components by CV is not feasible in this case.

Thus, the number of components is determined by visually inspecting the I_c profile and evaluating the final layer distribution obtained by the MoGP. Looking at the profile in Fig. 5a, it can be observed that the I_c value mainly fluctuates around two values. Moreover, while the 2-component model yields 7 layers, with a thin layer at around 10 m depth (Fig. 5b), the model with a higher number of components resulted in many very thin layers as is the case in the three-component model (Fig. 5c). Therefore, a 2-component model seems to be more suitable for this dataset. The parameters obtained by the 2-component model are provided in Table 3. Although other component selection methods have been proposed in the literature (Zhao and Ma 2016; Zhao, Chen, and Ma 2015), they are beyond the scope of this paper.

The limitation of the method implemented in this paper is that the model yields discontinuous layers (Fig.

5b) in the vertical direction when similar values are grouped at different depths, leading to the overlap of components in input space with a high probability. Therefore, misclassification of new input becomes more likely. Consequently, higher interpolation errors are obtained with the MoGP compared to the single GP in real CPT data. The reason for this behaviour is that during the assignment of indicator variables in the E-step, the correlation between data is neglected (Eq. (12)). Currently, a model is being developed to address this limitation.

Table 3. Predicted layer parameters by the two-component MoGP.

	Mean (I_c)	θ_0	θ_1 (m)	μ_k(m)	s_k (m)
Layer 1	1.86	0.043	1.72	15.8	8.66
Layer 2	3.09	0.072	0.96	13.6	8.54

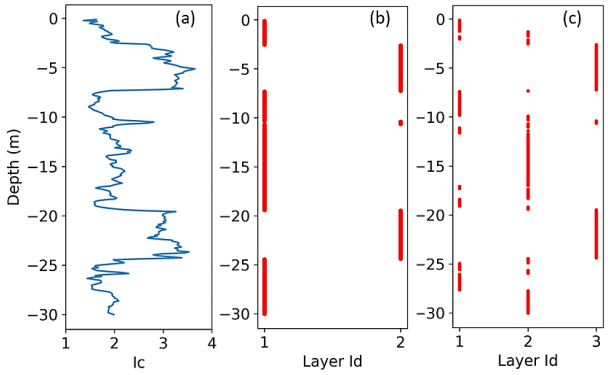

Figure 5. (a) I_c profile of CPT data; (b) layer distribution obtained by the two-component MoGP; (c) layer distribution obtained by the three-component MoGP.

3. Conclusions

A Mixture of Gaussian Processes (MoGP) model was explored in this paper for soil stratification identification. In practice, soil layers can be determined using engineering judgement or Robertson's chart when CPT data are available. However, these methods are not capable of quantifying stratification uncertainty, which can be significant due to the spatial variability of soil properties and the uncertainty of the boundaries proposed in the chart (Hu and Wang 2020). The proposed model can effectively discern various layers, capturing both the point and spatial statistics of each layer, along with assessing uncertainty in layer boundaries. The primary objective of the model is to perform stratification and data interpolation concurrently.

To evaluate the model's accuracy in interpolation and parameter estimation, it was initially tested on synthetic data with varying percentages of training data. Subsequently, the model was applied to real CPT data. Although the method was tested to complete missing data in the vertical direction, it can easily be applied to address the scarcity of CPT data in the horizontal direction by considering the two-dimensional input space.

The results from synthetic data indicate that the MoGP is promising, whereas the interpolation performance is poor with real CPT data. The physical reason for this behaviour is that the same type of soil is observed at different depths in the real CPT data, leading to the overlap of probability distributions of layers in the input space.

Acknowledgements

This work is supported by the research project RESET (Reliable Embankments for the Safe Expansion of rail Traffic), financed by ProRail.

References

Inman, H.F., and Bradley, E. L., "The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities." *Communications in Statistics - Theory and Methods*, 18(10), pp. 3851-3874, 1989. https://doi.org/10.1080/03610928908830127

Cao, Z.J., Zheng, S., Li, D.Q., and Phoon, K.K., "Bayesian identification of soil stratigraphy based on soil behaviour type index." *Canadian Geotechnical Journal*, 56(4), pp. 570–86, 2019. https://doi.org/10.1139/cgj-2017-0714.

Cao, Z.J., and Wang, Y., "Bayesian model comparison and selection of spatial correlation functions for soil parameters." *Structural Safety*, 49, pp. 10–17, 2014. https://doi.org/10.1016/j.strusafe.2013.06.003.

Chen, Z., Ma, Z., and Zhou, Y., "A precise Hard-Cut EM Algorithm for Mixtures of Gaussian Processes." In *Intelligent Computing Methodologies: 10th International Conference,* Taiyuan, China, 2014, pp. 68-75.

Ching, J., and Yoshida, I., "Data-driven 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.

Depina, I., Le, T.M.H., Eiksund, G., and Strøm, P., "Cone penetration data classification with Bayesian Mixture Analysis." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 10(1), pp. 27–41, 2016. https://doi.org/10.1080/17499518.2015.1072637.

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

Mavritsakis, A., Schweckendiek, T., Teixeira, A., Smyrniou, E., and Nuttall, J., "Bayesian analysis of benchmark examples for data-driven site characterization." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 9(2), 2023. https://doi.org/10.1061/ajrua6.rueng-975.

Phoon, K.K., Ching, J., and Shuku, T., "Challenges in data-driven site characterization." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 16(1), pp. 114–26, 2022. https://doi.org/10.1080/17499518.2021.1896005.

Phoon, K.K., Shuku, T., Ching, J., and Yoshida, I., "Benchmark examples for data-driven site characterisation." *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 16(4), pp. 599–621, 2022. https://doi.org/10.1080/17499518.2022.2025541.

Rasmussen, C.E., and Williams, C.K., "*Gaussian Processes for Machine Learning*." MIT Press, Cambridge, 2006.

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

Shuku, T., & Phoon, K. K. “Comparison of data-driven site characterization methods through benchmarking: Methodological and application aspects.” *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering,* 9(2), 2023. https://doi.org/10.1061/AJRUA6.RUENG-977

Wang, X., Wang, H., Liang, R.Y., Zhu, H., and Di, H., “A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data.” *Structural Safety*, 70, pp. 128–38, 2018. https://doi.org/10.1016/j.strusafe.2017.10.011.

Zhao, L., Chen, Z., and Ma, J., “An effective model selection criterion for Mixtures of Gaussian Processes.” In *Advances in Neural Networks–ISNN 2015: 12th International Symposium on Neural Networks,* Jeju, South Korea, 2015, pp. 345-354. https://doi.org/10.1007/978-3-319-25393-0_38.

Zhao, L., and Ma, J., “A dynamic model selection algorithm for mixtures of Gaussian processes.” In *IEEE 13th International Conference on Signal Processing (ICSP),* Chengdu, China, 2016, pp. 1095-1099.