Proc. of the 9th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Zhongqiang Liu, Jian Dai and Kate Robinson
Published by Research Publishing, Singapore. ©2025 The Author(s), CC-BY 4.0 licence.
doi: 10.3850/GRF-25280825_isgsr-027-P068-cd

2D SITE CHARACTERIZATION BY MIXTURE OF GAUSSIAN PROCESSES

Muhammet Durmaz and Michael A. Hicks

Department of Geoscience and Engineering, Delft University of Technology, Delft, The Netherlands
E-mail: M.durmaz@tudelft.nl, M.A.Hicks@tudelft.nl

Gaussian process regression is an effective method for the stochastic interpolation of geotechnical site data. However, a significant drawback of this method is its stationarity assumption, which is unrealistic given the existence of different soil layers. This assumption results in higher uncertainty in interpolation and poorer performance. To address this limitation, a mixture of Gaussian processes model is investigated for simultaneous layer identification and spatial interpolation in 2D. The model is based on the probabilistic assessment of layer boundaries and Gaussian processes for defining the statistical properties of the layers as well as spatial interpolation. The accuracy of the model is tested with real CPT profiles. The performance of the model is evaluated based on interpolation accuracy.

Keywords: CPT, Gaussian process, mixture of Gaussian processes, site characterization

1. Introduction

Gaussian process regression, also known as kriging, has been extensively applied in geotechnical site characterization. It is a robust method for uncertainty estimation in data interpolation and random field modeling. In this approach, geotechnical data are typically considered to comprise two components: (1) a trend (t) and (2) spatial variability (ε), where the latter is assumed to be a stationary random field. Geotechnical literature highlights that different soil layers exhibit distinct trend functions and spatial variability (Phoon et al. 2022). Consequently, models with simple trend functions, such as linear or quadratic trends, combined with a single spatial variability component are insufficient for accurately identifying and representing geotechnical data. To address this limitation, advanced methods such as sparse Bayesian learning (Ching and Phoon 2017), Gaussian process (Ching and Yoshida 2023), and geotechnical LASSO (Shuku and Phoon 2021) have been proposed. These approaches share a common feature: they train more flexible trend functions for the entire dataset while still using a single spatial variability component. Additionally, layering is not clearly identified. Recently, a preliminary version of the mixture of Gaussian processes (MGP) model was proposed by Durmaz et al. (2024) and implemented for stratification identification and completion of CPT data in a one-dimensional context. This paper introduces and applies an enhanced version of the MGP model for two-dimensional site characterization based on the soil behavior type index (I_c).

2. Model Definition

The mixture of Gaussian processes (MGPs) is derived from mixture-of-experts models, which consist of two components: (1) an expert that performs the desired task, such as regression, and (2) a gate function, which serves as a probabilistic allocation model. The gate function determines which expert is utilized in different regions of the input space. In the context of site characterization, the gate function can be interpreted as the mechanism for identifying different soil layers.

In this study, the gate function was specifically designed by adapting a Gaussian mixture model, commonly used as a gate function in MGPs (Meeds and Osindero 2005; Chen et al. 2014), to address the unique requirements of site characterization based on CPT data.

2.1 Mixture of Gaussian processes for the one-dimensional case

Assume a one-dimensional I_c profile consisting of multiple layers. A Gaussian process (GP) for the k^{th} layer is defined as:

$$p\left(\boldsymbol{I_c}^{\boldsymbol{k}}\middle|\boldsymbol{x^k}\right)\sim N[\boldsymbol{m_k}(\boldsymbol{x}), C\left(\boldsymbol{x^k}, \boldsymbol{x^k}\middle|\theta^{GP,k}\right)] \quad (1)$$

where $\boldsymbol{I_c}^{\boldsymbol{k}} = [I_{c,1}^k, I_{c,2}^k, I_{c,3}^k, I_{c,4}^k \dots I_{c,n}^k]$ is the collection of I_c observations for the k^{th} layer, $\boldsymbol{m_k}(\boldsymbol{x})$ is the mean function for the k^{th} layer and $C\left(\boldsymbol{x^k}, \boldsymbol{x^k}\middle|\theta^{GP,k}\right)$ is the covariance matrix conditioned on the hyperparameters $\theta^{GP,k}$ of the GP representing the spatial variability in the k^{th} layer. For the 1D case, $\boldsymbol{x^k} = [x_1^k, x_2^k, \dots x_n^k]$ represents the depths of I_c observations in the k^{th} layer. This study adopts a single exponential autocorrelation function, and the elements of the covariance matrix are, therefore, defined as:

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

where θ_0 and θ_1 represent the variance and scale of fluctuation of I_c, respectively.

The allocation of GPs to specific parts of the data is performed by a Gaussian mixture model (GMM), defined as follows:

First, given a data set $\{I_{c,n}, x_n\}_{n=1}^{N}$, the latent indicators $\{z_n\}_{n=1}^{N}$ indicate the GP responsible for each data point and are assumed to follow a multinomial distribution:

$$P(z_n = k) = \pi_k \quad (3)$$

The likelihood of the k^{th} GP being responsible for a given location is described by a Gaussian distribution:

$$p(x_n|z_n = k) \sim N(\mu_k, s_k^2) \quad (4)$$

The above model, whose details can be found in Chen et al. (2014), was used by Durmaz et al. (2024) for identifying the layers in a single CPT profile. However, this model has a significant limitation when applied to higher dimensions (2D or 3D): the decision boundaries of GMMs are quadratic in nature, and therefore lack the flexibility to represent the complex shapes of soil layer boundaries.

2.2 Mixture of Gaussian processes for the two-dimensional case

The following modification was introduced to overcome the previous model's limitation in a 2D space. The likelihood of the presence of a layer at a specific depth is still assumed to follow a Gaussian distribution. However, the parameters of the GMM are now considered to change in the horizontal direction. For M CPT logs, the log-likelihood of the observations is calculated as follows:

$$L(\boldsymbol{\Theta}, \boldsymbol{Z}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \{\sum_{n=1}^{N} z_{nkm} [ln(\pi_{k,m}) + lnp(x_n|\mu_{k,m}, s_{k,m})] + lnp(I_c^k|\boldsymbol{x}^k; \theta^k)\} \quad (5)$$

where $\boldsymbol{\Theta}$ denotes the whole parameter set, and $\boldsymbol{Z}$ is the indicator tensor with elements z_{nkm}. If a data point belongs to the k^{th} layer at the m^{th} CPT, then z_{nkm} = 1; otherwise, z_{nkm} = 0.

The hard-cut expectation-maximization (EM) algorithm (Chen et al. 2014) is applied to train the model as follows:

(i) K-means clustering is used to initialize the clusters. The indicator variables are initialized to represent the indices of the clusters where the data points are assigned.

(ii) **M-step:** The hyperparameters of the GPs in each layer are calculated using the maximum likelihood (ML) method. The quadratic polynomial trend in Eq. (6) is used as the mean function for each layer. This trend represents the highest-order function commonly employed in the literature for homogeneous layers and has also been reported to perform well for lower-order functions, such as constant and linear trends (Ching et al., 2023). To avoid overfitting, normal prior distributions are assigned to the coefficients of the trend function in Eq. (6). The mean and standard deviation of I_c values in each layer are set to the mean and standard deviation of the prior distribution of w_0. The means of the prior distributions of w_1 and w_2 are set to zero. The standard deviations of the priors are set to $[I_{c,max}^k - I_{c,min}^k]/[x_{max}^k - x_{min}^k]$ for w_1 and $[I_{c,max}^k - I_{c,min}^k]/[(x_{max}^k)^2 - (x_{min}^k)^2]$ for w_2.

$$m_k(x) = w_0^k + w_1^k \cdot (x - x_{min,k}^k) + w_2^k \cdot (x - x_{min,k}^k)^2 \quad (6)$$

For the calculation of the gate parameters, $\pi_{k,m}, \mu_{k,m}$, and $s_{k,m}$, the following equations are given in the original algorithm (Chen et al. 2014):

$$\pi_{k,m} = \frac{N_{k,m}}{\sum_{k=1}^{K} N_{k,m}} \quad (7\text{-a}) \qquad \mu_{k,m} = \frac{\sum_{n=1}^{n=N} x_n^{k,m}}{N_{k,m}} \quad (7\text{-b}) \qquad s_{k,m} = \frac{\sum_{n=1}^{n=N} (x_n^{k,m} - \mu_{k,m})(x_n^{k,m} - \mu_{k,m})^T}{N_{k,m}} \quad (7\text{-c})$$

where $x_n^{k,m}$ represents the n^{th} depth data, and $N_{k,m}$ denotes the total number of data points in the k^{th} layer at the m^{th} CPT. With the assumption that layers are continuous (the same layer cannot appear at two distinct depths for a given CPT), $\mu_{k,m}$ and $s_{k,m}$ can be expressed in terms of $\pi_{k,m}$ as follows:

$$\mu_{k*,m} = \sum_{\substack{k=1 \\ k \neq k*}}^{K,m} \left(\pi_k + \frac{\pi_{k*}}{2}\right) \cdot H_m \quad (8\text{-a}) \qquad s_{k*} = \pi_{k*} \cdot H_m / \sqrt{12} \quad (8\text{-b})$$

where H_m is the total of depth the m^{th} CPT. Eq. (8-a) and Eq. (8-b) require the layer index k to be ordered from the ground surface to the bottom. Therefore, after the clusters are initialized, the layer indexes are ordered based on the mean depth of each cluster.

(iii) **E-step**: Indicator variables are updated based on the maximum a posteriori criterion:

$$z_n = \underset{k}{\operatorname{argmax}} [\pi_k N(x_n|\mu_k, s_k) N(y_n|m_k(x_n), \theta_0^k)] \quad (9)$$

(iv) **Prediction:** I_c values between CPTs are estimated as the weighted mixture of predictions made by each GP according to:

$$p(I_c^*|x_n^*) = \sum_{i=k}^{i=K} P(z_n = k|x_n^*) N\left[I_{c,k}^*, V\left(I_{c,k}^*\right)\right] \quad (10)$$

where $P(z_n = k|x_n^*)$ is the posterior probability of the presence of the k^{th} layer at the prediction point x_n^*, and $I_{c,k}^*$ is the mean and $V\left(I_{c,k}^*\right)$ is the variance of the predictive distribution of k^{th} GP, which are calculated as follows:

$$I_{c,k}^* = C\left(x^*, \boldsymbol{x^k}\right) C\left(\boldsymbol{x^k}, \boldsymbol{x^k}\right)^{-1} \left[\boldsymbol{I_c^k} - \boldsymbol{m_k}\left(\boldsymbol{x^k}\right)\right] + m_k(x^*) \quad (11)$$

$$V\left(I_{c,k}^*\right) = C(x^*, x^*) - C\left(x^*, \boldsymbol{x^k}\right) C\left(\boldsymbol{x^k}, \boldsymbol{x^k}\right)^{-1} C\left(\boldsymbol{x^k}, x^*\right) \quad (12)$$

After the training stage, the likelihood of a layer presence between CPTs needs to be identified. To achieve this, the mixture weights of the layers are interpolated using a multivariate (multi-task) GP (Bonilla et al. 2007) with an RBF kernel. Note that this is equivalent to interpolating layer thicknesses since the product of the mixture weights and the total thickness equals the layer thickness. The formulation of the multi-task GP is similar to Eq. (1), except for the covariance matrix. In the multi-task GP, the covariance matrix is the product of the multi-task covariance matrix and the correlation matrix. The posterior probability of a layer presence in Eq. (10) is calculated using Monte Carlo (MC) simulation by sampling from the predictive distribution of layer thicknesses generated by the multi-task GP. The posterior probability of a layer's existence at a given point, $P(z_n = k|x_n^*)$, is then computed as the ratio of the number of samples in which the point is assigned to the k^{th} layer to the total number of MC samples, N_k/N.

3. Implementation with CPT Data

The proposed method is implemented with data from 5 CPTs from the Groningen region in the Netherlands. The relative positions of the CPTs and the I_c profiles are shown in Figure 1. Five layers are selected to model the subsurface as it is a reasonable number of layers for the illustrated data. The results of the I_c prediction, the most probable layer boundaries, and the uncertainties in the layer boundaries are shown in Fig. 2a, 2b, and 2c respectively. To assess the interpolation accuracy of the model, cross-validation (CV) is performed by excluding CPT2, CPT3, and CPT4 in turn during the training stage. The root mean square error (RMSE) and mean absolute error (MAE) are calculated using the validation data and prediction at the location of the validation CPT. The mean RMSE and MAE from the three CV analyses are calculated as 0.33 and 0.26, respectively. The same error metrics are calculated as 0.38 and 0.28 when single Gaussian processes are used for the entire subsurface, which demonstrates that the proposed model improves the interpolation performance. When the most probable layer boundaries (Fig. 2b) are compared to the data shown in Fig. 1, the layer boundary estimation seems reasonable. Also, the uncertainty in the layer boundaries increases when the points become far away from the CPT locations, which is also realistic.

Table 1. Estimation of the coefficients of the mean function and the hyperparameters for each layer. In the table θ_0, θ_1, θ_2 represent the variance, and the vertical and horizontal scales of fluctuations respectively.

	w_0	w_1	w_2	θ_0	θ_1(m)	θ_2(m)
Layer 1	1.73	-0.136	0.1400	0.161	1.8	45.5
Layer 2	3.03	0.076	-0.0190	0.059	1.0	38.6
Layer 3	1.76	0.026	0.0002	0.058	1.3	77.9
Layer 4	2.64	0.041	-0.0014	0.128	1.0	39.3
Layer 5	1.9	0.028	0.0029	0.089	1.2	42.9
single GP	1.04	0.059	-0.0002	0.407	5.7	394

Table 1 shows the estimated mean and covariance parameters of the five layers in the simulation, whose results are illustrated in Fig. 2. The bottom row of the table lists the values of the hyperparameters obtained by a single GP. As expected, the scale of fluctuation estimation in both directions with a single GP is higher than with the layered approach because of the interlayer heterogeneity.

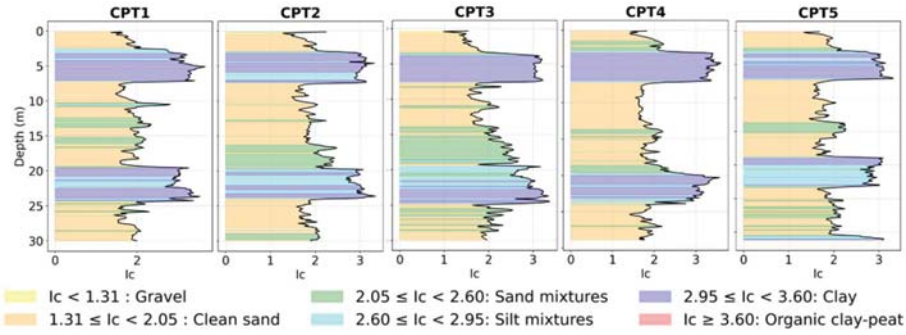

Fig.1 I_c profiles of the 5 CPTs and corresponding soil behaviour type classification used for implementation.

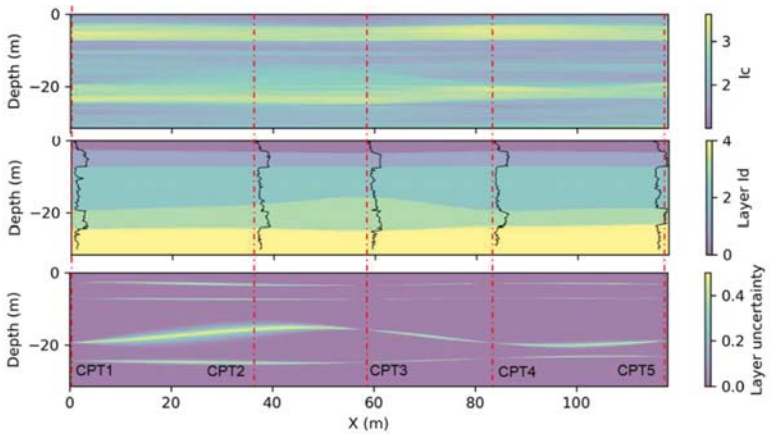

Fig. 2 (a) I_c profiles obtained by the 5-layer MGP model; (b) most probable layer boundaries with I_c profiles at CPT locations (black line); (c) uncertainties in layer boundaries. These results were obtained by excluding CPT2.

4. Conclusion

A mixture of Gaussian processes has been proposed to identify layer boundaries and the spatial statistics of each layer. Geotechnical data are represented with a unique trend function and a spatial variability component in each layer. In this respect, the proposed approach differs from other studies which train complicated mean functions to model non-stationarity. Although this paper has presented a simple trend model for individual layers, the global trend is flexible enough to represent geotechnical data with the inclusion of interlayer heterogeneity.

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

References

Bonilla, E. V., Chai, K., and Williams, C. (2007) Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, *20,* pp. 153-160.

Chen, Z., Ma, Z., and Zhou, Y. (2014) A precise hard-cut EM algorithm for mixtures of Gaussian processes. In *Intelligent Computing Methodologies: 10th International Conference,* Taiyuan, China, pp. 68-75.

Ching, J., and Phoon, K. K. (2017) Characterizing uncertain site-specific trend function by sparse Bayesian learning. *Journal of Engineering Mechanics*, *143(7),* 04017028-1.

Ching, J., and Yoshida, I., (2023) 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),* 04022064.

Durmaz, M., van den Eijnden, A. P., and Hicks, M. A. (2024) Stratification identification and prediction of missing CPT data by Mixture of Gaussian Processes. In *7th International Conference on Geotechnical and Geophysical Site Characterization,* pp. 1786-1792.

Meeds, E., and Osindero, S. (2005) An alternative infinite mixture of Gaussian process experts. *Advances in neural information processing systems*, *18,* pp. 883-890.

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

Shuku, T., and Phoon, K. K. (2021) Three-dimensional subsurface modeling using Geotechnical Lasso. *Computers and Geotechnics*, *133*, 104068.