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
©2022 ISGSR Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-5182-7_02-006-cd

Gaussian Process Regression and Kernel Selection for Missing Geotechnical Data Prediction

JiaweiXie[1],Jinsong Huang[2]and Cheng Zeng[3]

[1]Discipline of Civil, Surveying and Environmental Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.
E-mail: jiawei.xie@uon.edu.au
[2]Discipline of Civil, Surveying and Environmental Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.
E-mail: jinsong.huang@newcastle.edu.au
[3]Discipline of Civil, Surveying and Environmental Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.
E-mail: cheng.zeng@uon.edu.au

Abstract: Geotechnical site investigation data (i.e., CPT data) may be missing sometimes, due to, for example, sensor failure or storage issues. Conventionally, missing data is interpolated based on mean value imputation or linear interpolation, in which spatial correlation within the data is ignored. The spatial correlation can be considered explicitly in the geostatistical interpolation methods such as the kriging methods. However, kriging methods involve challenges for estimating the model parameters such as the scale of fluctuations in the covariance model. Gaussian Process Regression (GPR) method infers the model parameters based on maximizing the marginal likelihood. However, kernel selection will largely influence the model performance of the GPR method. This paper aims to compare the performance of nine widely used base kernels. Ninety new kernels based on combination of the base kernels are generated to enhance the adaptability of the GPR model. Four types of stratum with increasingly complex profiles are tested for each kernel based on multiple CPTs. The most suitable kernels for each type of stratum are suggested based on cross-validation with more than one thousand models. The proposed method has been applied to a real-world CPT dataset to show its applicability and robustness.

Keywords: Missing data;Gaussian process regression;spatial correlation;kernel selection.

1 Introduction

Geotechnical site investigation involves mapping soil properties based on limited measurements. Real-world measurement data are very likely to contain missing values (i.e., null values). The first step in preprocessing these measurement data is to handle these missing data. Various methods are utilized in the literature to predict the missing data in a dataset (e.g., Osman et al., (2018)). Conventional missing data handling techniques include mean value (or median value) imputation and linear interpolation (e.g.,Gómez-Carracedo et al., (2014)). Mean value imputation is simply to replace missing data based on the mean value of the remaining data. However, this method may cause discontinuity and change the distribution of original data. Linear interpolation simply connects two adjacent data points with linear lines. This method fails to consider the variability in the data. The correlations between the adjacent data are ignored in this strategy, which may cause a waste of information.

Machine learning methods make predictions based on learning the data pattern in the available dataset. Nowadays, GPR machine leaning method gets more attention in geotechnical engineering (e.g., Kang et al., 2017; Yoshida et al., 2021). GPR is a generalization of kriging method (e.g., Cui et al., (2021)), which is widely used in geology data interpolation. Different from the kriging method, GPR is a Bayesian perspective, which is a group of kernel methods to provide a conditional statistical description for the target variables. GPR assumes a multidimensional Gaussian process between the observed samples and unknown samples. The missing data is predicted based on conditioning on the observed data. The kernels govern the shape of the multi-dimensional Gaussian process. GPR is flexible with abundant kernels which can adapt to complicated data patterns. The parameters in the kernels can be automatically determined by maximum the log-marginal-likelihood estimation (Rasmussen, 2003). Based on the estimated kernel, the prediction can provide a quantification of uncertainty through a confidence interval.

Kernel selection is essential in applying the GPR method. Different kernel indicates different covariance pattern between adjacent samples. Kernels can be significantly different from each other. The most suitable kernel should be able to best explain the covariance in the observed data. In geotechnical engineering, the real-world data pattern may be complicated (i.e., more than one type of pattern), so a base kernel function maybe not be enough. Nine commonly used kernels in geotechnical engineering are tested in this study. Various new kernels can be generated based on combining base kernels (e.g., Seeger, (2004)). Combining kernels can produce a more accurate and robust tool to describe the data pattern (Duvenaud, 2014). Multiplying two kernels and adding two kernels are two commonly used ways to build a new combined kernel. Adding two kernels will enlarge the values where two base kernels have a high value. Multiplying two kernels will enlarge the values only when both of the kernels have high values.

Automatical kernel selection is one of the main difficulties for GPR (e.g., Duvenaud, 2014; Abdessalem et al., 2017). Currently, automatically kernel selection algorithm still needs manual judgment and intervention (Cui et al., 2021). Conventional kernel selection method is to try different kernels and compare their marginal likelihood on the observed data, which can be treated as an indirect evaluation method. This paper utilizes a direct evaluation method in which the performances of different kernels are compared directly based on cross-validation method. In section 4, four types of stratum with increasingly complex profiles are tested for each kernel based on multiple CPTs. The most suitable kernels for these soil stratum are suggested based on cross-validation with more than one thousand models. Finally, the proposed method has been applied to a real-world CPT dataset as an illustration.

2 Methodology

The definition of a Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. A detailed description of GPR can be found in Schulz et al., (2018).

A GP is a probability measure over space and is described by a mean function$\mu(\mathbf{x})$and a covariance function$k(\mathbf{x},\mathbf{x}')$. $\mathbf{x}$is the spatial coordinate which is the depth of CPT data in this study. Let $f(\mathbf{x})$denote the cone tip resistance value q_t, the GP can be described by $f(\mathbf{x})\sim\mathrm{GP}(\mu(\mathbf{x}),k(\mathbf{x},\mathbf{x}'))$.

Assume$f_i = f(\mathbf{x}_i)$is the observation for $\mathbf{x}_i$. Suppose that a target set $\mathbf{x}_*$is given. The aim is to predict the f_*based on the observations. This can be described as a distribution $p(f_*|\mathbf{x}_*,\mathbf{x},f)$.

The joint Gaussian distribution can be written in the form:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix}\right) \tag{1}$$

where $\mathbf{K} = k(\mathbf{x},\mathbf{x})$ is the covariance matrix $(N\times N)$ for the training points. $\mathbf{K}_* = k(\mathbf{x},\mathbf{x}_*)$ is the covariance matrix $(N\times N_*)$ for the training and target points. $\mathbf{K}_{**} = k(\mathbf{x}_*,\mathbf{x}_*)$ is the covariance matrix$(N_*\times N_*)$ for the test points.

The conditional distribution can be calculated by:

$$\begin{aligned} p(f_*|\mathbf{x}_*,\mathbf{x},f) &= \mathcal{N}(f_*|\mu_*,\textstyle\sum_*) \\ \mu_* &= \mu(\mathbf{x}_*) + \mathbf{K}_*^T\mathbf{K}^{-1}(f-\mu(\mathbf{x})) \\ \textstyle\sum_* &= \mathbf{K}_{**} - \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}_* \end{aligned} \tag{2}$$

If the measurement error needs to be considered in the observations, an error term σ_0^2 can be considered in $\mathbf{K}$. The values in matrix $\mathbf{K}$ are calculated based on the specified kernel function. The hyper-parameters in the kernel function can be determined by maximizing the marginal likelihood $p(f\ |\mathbf{x},\boldsymbol{\theta})$ of the GP.

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}}(p(f\mid\mathbf{x},\boldsymbol{\theta})) \tag{3}$$

$$\log p(f\ |\mathbf{x},\boldsymbol{\theta}) = -\frac{1}{2}f^T\mathbf{K}^{-1}f - \frac{1}{2}\log|\mathbf{K}^{-1}| - \frac{N}{2}log2\pi \tag{4}$$

where $\boldsymbol{\theta}$ is a vector containing the required kernel hyper-parameters.

3 Kernel Visualization

GPR Kernels can be significantly different with each other. The characteristics of different kernel can be visualized based on sampling from specific kernel function. The most commonly used kernel function is the Gaussian kernel (i.e., Radial Basis Function kernel, Squared Exponential kernel). It can be described by following function:

$$k_{\mathrm{GS}}(x,x') = \sigma^2\exp\left(-\frac{(x-x')^2}{2\ell^2}\right) \tag{5}$$

Gaussian kernel is infinitely differentiable, so the associated Gaussian Process has infinitely many derivatives. The parameters for the Gaussian kernel are the scale of fluctuation ℓ and variance σ^2. ℓ is the distance where data are significantly correlated. σ^2 is a scale factor. $x-x'$ can be recognized as the Euclidean distance between two samples. The Gaussian kernel can be visualized by Figure (1).

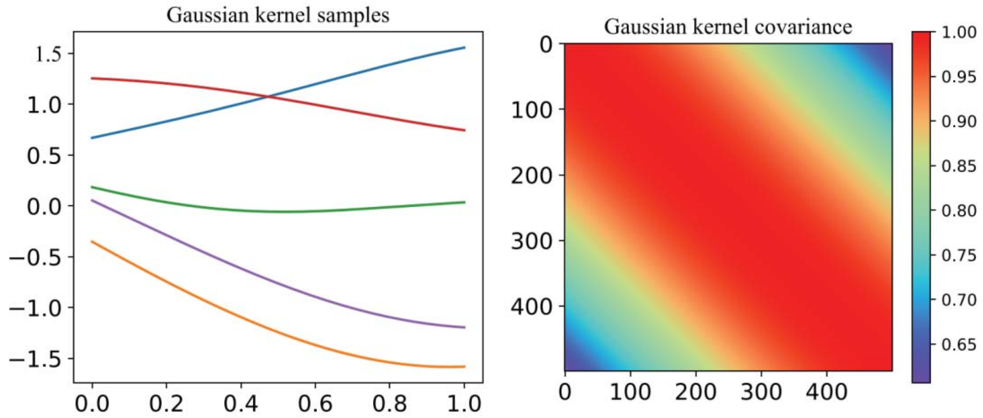


Figure 1. Gaussian kernel visualization

In Figure (1), five randomly sampled profiles based on the Gaussian kernel are generated. The Gaussian kernel sometimes may be too smooth to simulate a correlation structure. Then the exponential kernel function will be a better choice. The exponential kernel is similar to the Gaussian kernel, which can be described by the following function:

$$k_{\text{EP}}(x,x') = \sigma^2 \exp\left(-\frac{|x-x'|}{2\ell}\right) \tag{6}$$

The exponential kernel can be visualized by Figure (2).

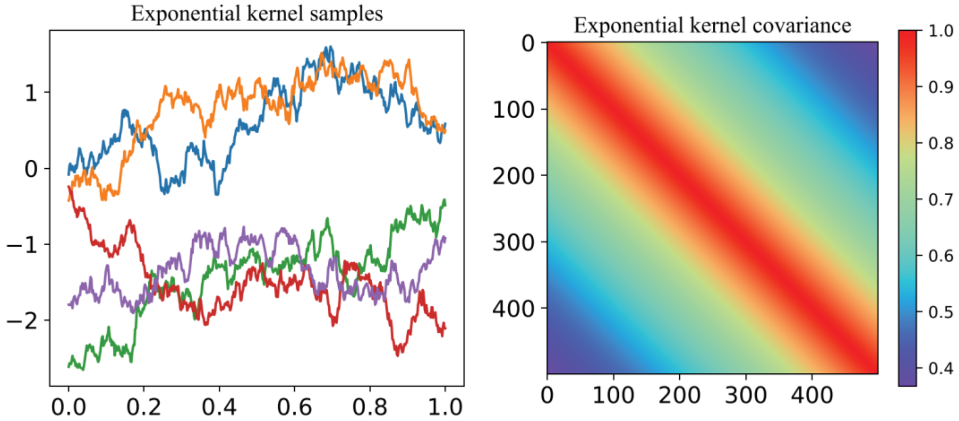


Figure 2. Exponential kernel visualization

There are some more complicated kernels with more hyper-parameters. Such as the Rational Quadratic (RatQuad) Kernel (Eq. (7)) and the Matern kernel (Eq. (8)).

$$k_{\text{RQ}}(x,x') = \sigma^2 \left(1 + \frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{7}$$

$$k_{\text{MA}}(x,x') = \frac{\sigma^2}{\Gamma(v)2^{v-1}} \left(\frac{\sqrt{2v}}{\ell}|x-x'|\right)^{v} K_v\left(\frac{\sqrt{2v}}{\ell}|x-x'|\right) \tag{8}$$

where $K_v$is a modified Bessel function.

These functions are generalizations of the Gaussian kernel and exponential kernel. For example, when $\alpha \rightarrow \infty$, the RatQuad kernel is identical to the Gaussian kernel. If $v = 1$, the Matern function becomes the exponential function. Two important $v = 1.5$ (once differentiable functions) and $v = 2.5$ (twice differentiable functions) will be investigated in this study (Mat32 and Mat52). Other kernels involved in this research are the Linear kernel, Multi layer perceptron (MLP) kernel, Poly kernel and Brownian kernel. More details about these kernels can be found in Seeger, (2004).

As discussed in the introduction, single kernel may not be enough to capture the data pattern in the CPTs. New kernels can be built by combining the base kernels. This study focus on two standard ways of combining kernels: addition and multiplication. These processes can be described by following formulas:

$$k_a = k_1(x, x') + k_2(x, x')$$
$$k_m = k_1(x, x') \times k_2(x, x') \tag{9}$$

As an example, the exponential add Gaussian kernel can be visualized by Figure (3).

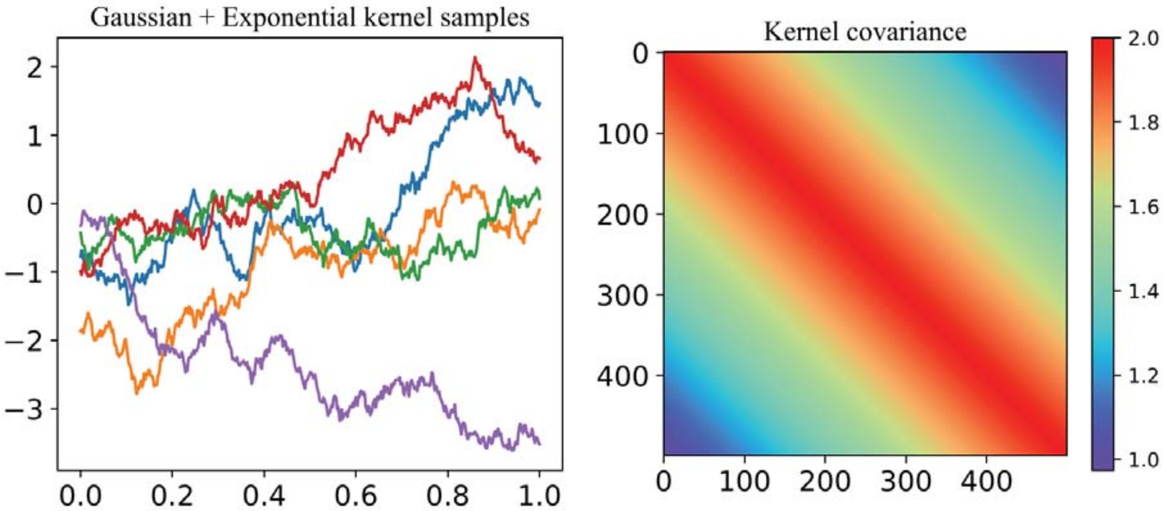


Figure 3. Gaussian add Exponential kernel visualization

In this way, 99 kernels will be generated and tested in this study (9 base kernels, 45 new kernels based on addition and 45 new kernels based on multiplication). It should be noted that considering the measurement error always exists in the geotechnical measurements, a white noisy kernel is always plus to the kernels which assumes a Gaussian noisy with zero mean and standard deviation σ_0. σ_0 can also be determined by maximizing the marginal likelihood function.

4 Buidling GPR Models for Various Stratum Conditions

This paper utilizes the synthetic benchmark soil profiles generated in Phoon et al., (2022). The sampled CPTs come from 3D models with discretized cells of $1 \times 1 \times 0.1$ m. The resolution for the CPT is 0.1m. These 3D models are generated based on predefined trends (constant, linear, or higher-order) and random field methods based on exponential covariance function. Four types of stratum with increasingly complex profiles are tested: Horizontal layers with constant property trend in each layer (S-VG1); Inclined layers with constant property trend in each layer (S-VG2); Inclined layers with linear property trend in each layer (S-VG3); Mixture of continuous and discontinuous layers with a constant trend in each layer (S-VG4).

The term missing rate is defined to describe the ratio between missing data and completed data. This paper focus on the cases with a high missing ratio. One can simply apply linear interpolation for the cases with low missing ratio without a loss of too much accuracy. Missing rate of 0.8 will be used in this study, which means 80 percent of data is assumed to be unknown and will be used to validate the predictions. To make the results more reliable, three CPTs are sampled from each kind of stratum condition. Overall, twelve models will be built for each kind of kernel. Considering there are 99 types of kernels, 1188 models will be built and compared in this study.

The performance of different kernels can be evaluated based on the difference between predictions and validation data. The accuracy can be calculated based on the mean square error (MSE):

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(q_t^i - \hat{q}_t^i)^2 \tag{10}$$

where q_t^i is the ith measured q_t value and$\hat{q}_t^i$ is the ith predicted q_t value.

For GPR method, q_t prediction is given by a normal distribution. The mean value is used to estimate the accuracy. The MSE for each CPT will be normalized and compared to get the final estimation of the kernel performance on each kind of stratum condition. The standard deviation will be used to calculate a 95% confidence interval and be visualized in the predicted profiles.

5 Results

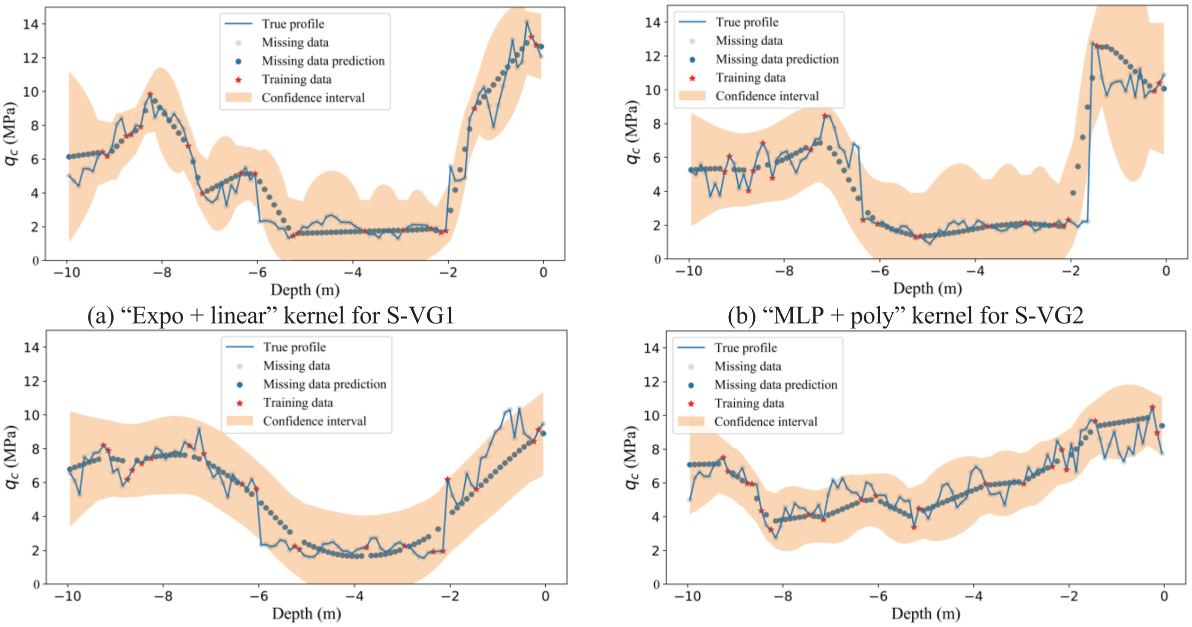
Based on the cross-validation method, the best GPR models with suitable kernels can be determined for each kind of stratum. The top five models for S-VG1 to S-VG4 are listed in Table 1.

Table 1. Top five kernel combinations for different stratum conditions.

	Best model	Second model	Third model	Fourth model	Fifth model
S-VG1	Expo + linear	Expo	Expo × poly	Expo × Expo	Expo + Brownian
S-VG2	MLP + poly	MLP × Mat52	MLP × poly	linear + MLP	MLP + MLP
S-VG3	Expo × Gauss	Expo × Mat52	Expo × Mat32	Expo × poly	Expo + linear
S-VG4	Expo + Expo	Expo + poly	Expo + linear	Expo + Brownian	Expo
ALL	Expo + linear	Expo + poly	Expo + Expo	Expo	Expo × poly

Table 1 shows that all the top one model comes from the combined kernels. This illustrates the importance of using a combined kernel as a more robust kernel in geotechnical engineering. Table 1 can be used as a guideline for GPR kernel selection. For real-world applications, the top five models can be tried first if there is evidence of similarities between one of S-VG1 to S-VG4 and the real stratum of interest. However, the real-world stratum conditions may be more complicated than the predefined S-VG1 to S-VG4 stratum conditions. So the more general top five models for all four kinds of stratum conditions are listed in the last row of Table 1. These five kernels performs good in all soil conditions. So they are more robust than the previous results which should be tried if no additional information is available.

The typical prediction results for S-VG1 to S-VG4 with their best models are shown in Figure (4). Figure (4) shows that the GPR results capture the trend of real profiles well. However, due to the complexity of the profiles and the sparse data available, local fluctuations cannot be well predicted. The good thing is that all the local fluctuations are generally contained by the 95% confidence interval. This is important for risk control. The results are good enough for engineering applications.



(a) "Expo + linear" kernel for S-VG1

(b) "MLP + poly" kernel for S-VG2

(c) "Expo + Gauss" kernel for S-VG3

(d) "Expo + Expo" kernel for S-VG4

Figure 4. Best model results for different stratum conditions

The computation speed for the GPR models is fast. Using a personal computer with Intel(R) Core(TM) i9-9900 CPU @ 3.10GHz, 32GB RAM in a 64-bit Windows 10 operating system as an example, the computation time for a single Exponential kernel is 0.28s and for a single linear kernel is 0.21s. The computation time for an exponential add linear kernel is 0.46s.

To demonstrate the applicability of the proposed method and effeteness of the suggested kernel, GPR methods are used to real-world data that was collected at the South Parklands site in the city of Adelaide, South Australia (Jaksa, 1995). The resolution of the CPT is 0.05m. Though this project has been carefully designed and measured, several CPTs in this dataset contain missing data. One of the CPTs is used as a demonstration. The Exponential add Linear kernel is selected as suggested in Table 1. The results are shown in Figure (5).

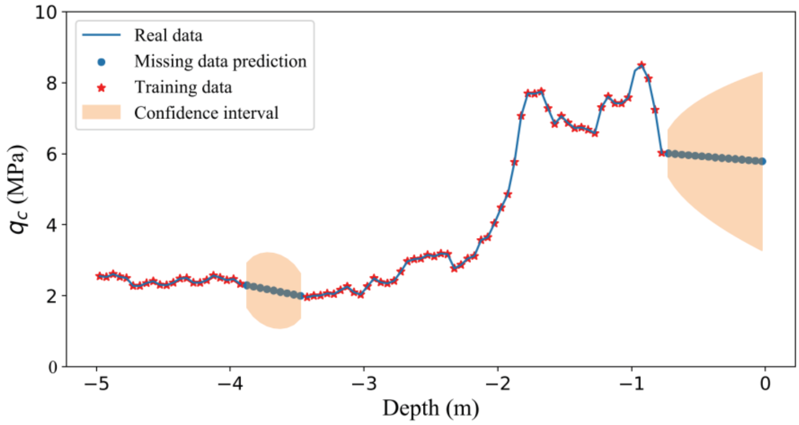


Figure 5. An example of missing data prediction for a real-world CPT profile

The result generally seems to be reliable within depths -4m to -3m as it catches the trend of the CPT data. In the data from 0m to -1m, the model give a linear estimation based on the closest data. Extrapolation is not suggested for most machine learning methods. The confidence interval also shows a bigger uncertainty in this area.

6 Conclusions

GPR method can be used to predict the missing data for geotechnical data. The results show that the combined kernels perform better than the base kernel. The best kernels for different stratum conditions are suggested. The GPR prediction results can capture the main trend of the CPT profile even if the data is sparse. The measurement error can be flexibly considered in the prediction. The confidence interval can be used to estimate the reliability of the prediction.

Acknowledgments

This study was funded by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (Project DP190101592) and the National Natural Science Foundation of China (No.41972280, 52179103).

References

Abdessalem, A.B., Dervilis, N., Wagg, D.J., Worden, K., (2017). Automatic Kernel Selection for Gaussian Processes Regression with Approximate Bayesian Computation and Sequential Monte Carlo. *Frontiers in Built Environment* 3.

Cui, T., Pagendam, D., Gilfedder, M., (2021). Gaussian process machine learning and Kriging for groundwater salinity interpolation. *Environmental Modelling & Software* 144, 105170. https://doi.org/10.1016/j.envsoft.2021.105170

Duvenaud, D., (2014). Automatic model construction with Gaussian processes (PhD Thesis). *University of Cambridge*.

Gómez-Carracedo, M.P., Andrade, J.M., López-Mahía, P., Muniategui, S., Prada, D., (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems* 134, 23–33. https://doi.org/10.1016/j.chemolab.2014.02.007

Jaksa, M.B., (1995). The influence of spatial variability on the geotechnical design properties of a stiff, overconsolidated clay. (Thesis). *The University of Adelaide*.

Kang, F., Xu, B., Li, J., Zhao, S., (2017). Slope stability evaluation using Gaussian processes with various covariance functions. *Applied Soft Computing* 60, 387–396. https://doi.org/10.1016/j.asoc.2017.07.011

Osman, M.S., Abu-Mahfouz, A.M., Page, P.R., (2018). A Survey on Data Imputation Techniques: Water Distribution System as a Use Case. *IEEE Access* 6, 63279–63291. https://doi.org/10.1109/ACCESS.2018.2877269

Phoon, K.-K., Shuku, T., Ching, J., Yoshida, I., (2022). benchmark Examples for Data-Driven Site Characterization. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards* 1–23.

Rasmussen, C.E., (2003). Gaussian processes in machine learning, *Summer School on Machine Learning*. Springer, pp. 63–71.

Schulz, E., Speekenbrink, M., Krause, A., (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology 85, 1–16. https://doi.org/10.1016/j.jmp.2018.03.001

Seeger, M., (2004). Gaussian processes for machine learning. *International Journal of Neural Systems*. 14, 69–106. https://doi.org/10.1142/S0129065704001899

Yoshida, I., Tomizawa, Y., Otake, Y., (2021). Estimation of trend and random components of conditional random field using Gaussian process regression. *Computers and Geotechnics* 136, 104179. https://doi.org/10.1016/j.compgeo.2021.104179