Demystifying the connections between Gaussian Process regression and kriging theories

S. K. Suryasentana
University of Strathclyde, Glasgow, UK

B. B. Sheil
University of Cambridge, Cambridge, UK

ABSTRACT: Gaussian Process (GP) regression is a flexible, non-parametric Bayesian approach towards regression problems that has seen increasing adoption for machine learning (ML) applications. Despite its recent popularity within the ML community, GP regression has a long history in geostatistics, where it is better known as kriging and is commonly used for spatial interpolation. The rapid development of GP regression in ML presents significant opportunities for advanced knowledge transfer to the geotechnical engineering community. However, this knowledge transfer has often been inhibited by the different terminology and conventions adopted in both fields. This obscures the underlying science and introduces much potential for confusion. Therefore, this paper aims to reveal the connections between GP regression and kriging theories, with a view of acting as a bridge to increase the uptake of the latest developments in each field.

1 Introduction

1.1 *Kriging*

In geotechnical engineering, it is common to estimate geotechnical parameters at unsampled locations. For this purpose, spatial interpolation is typically used to estimate the parameter based on data from nearby sampled locations. Among the many interpolation techniques available, kriging is widely used (e.g. Li et al. 2016) as it is a probabilistic approach that can accommodate correlations among the data. Unlike other interpolation techniques that prespecify the interpolating function (e.g. polynomial-based), kriging produces an interpolation function based on a semivariogram/variogram model that is derived from data (Matheron 1963; Cressie 1993).

1.2 *Gaussian Process regression*

In recent years, Gaussian Process (GP) regression has been widely used by the machine learning (ML) community to model nonlinear functional relationships. GP regression is a flexible, non-parametric Bayesian approach towards regression problems. Like kriging, GP regression does not prespecify a parametric form for the regression function but instead lets the data determine the complexity of the function, which allows it to model arbitrarily complex systems. It has been used in many ML applications such as robotics learning (Deisenroth et al. 2013) and time-series modelling (Roberts et al. 2013).

1.3 *Connections*

Despite their apparent differences, the kriging and GP regression approaches can actually be shown to be intimately connected and in many cases, equivalent. This is because central to each approach is the covariance function, which is more commonly called a kernel in the ML literature, and is closely related to the variogram model in the kriging approach. Although the underpinnings of both approaches are very similar, the two approaches differ in how they derive the interpolating/regression function. This may give the impression that differences outweigh the similarities and the connections are only superficial. Consequently, knowledge transfer between these two fields has often been inhibited by the different terminology adopted in both fields, which obscures the underlying science and introduces potential for confusion. Therefore, the main objective of this paper is to reconcile these differences and help researchers in both fields gain mutual understanding, and be able to transfer knowledge across fields. A secondary objective of this paper is to provide a short and accessible overview of both theories to researchers who are new to either field and may not be aware of these connections (as they are typically scattered individually and separately across a variety of non-related literature). This paper will briefly review the theories behind kriging and GP regression Then, their connections are explored in detail, including equivalences and differences.

2 Theories

2.1 *Kriging*

2.1.1 *Introduction*

Kriging originates as a spatial interpolation method in the field of mining geology (Krige 1951), where it uses the spatial correlation between a finite set of sampled data points to estimate the value of a variable over a continuous field. It has since been adopted in geotechnical engineering, primarily for spatial interpolation of soil properties (e.g. Li et al. 2016).

2.1.2 *Semivariogram*

Kriging is an interpolation method based on an assumed correlation with existing data. It first requires the spatial covariance of the sampled data points be determined by fitting a semivariogram γ_h, as follows:

$$\gamma_h = \frac{1}{2}\mathrm{E}[(y_i - y_{i+h})^2] \quad (1)$$

where y_i and y_{i+h} are pairs of sample data points that are separated by a distance h. Equation 1 thus represents the average squared difference in the values between pairs of sample data points (Matheron 1963; Cressie 1993). Associated with the semivariogram are three key properties: (i) 'nugget', which represents the offset value (from zero) of the semivariogram when h is zero; (ii) range, which represents the distance where the semivariogram first flattens out and (iii) 'sill', which represents the value of the semivariogram at the range. Figure 1 illustrates these properties for a typical semivariogram.

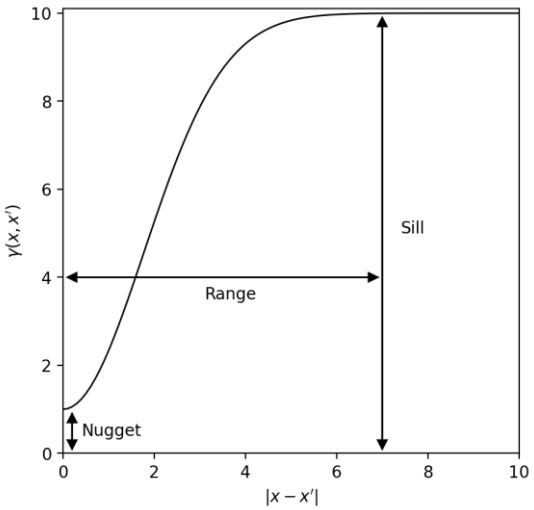


Figure 1. Example Gaussian semivariogram with its nugget, range and sill, where $h = |x - x'|$.

2.1.3 *Prediction*

The spatial covariance structure represented by the semivariogram is used to calculate the weights that are then used to estimate the values at unsampled locations. The estimated value $\hat{y}_*$ at an unsampled location x_* is obtained through a linear weighting of the sampled data y_i:

$$\hat{y}_* = \boldsymbol{\lambda}^T \mathbf{y} \quad (2)$$

$\boldsymbol{\lambda} = [\lambda_1(x_*), \dots, \lambda_N(x_*)]^T$ and $\boldsymbol{y} = [y_1, \dots, y_N]^T$ are vectors of the weights and sampled data respectively (where N is the number of data points), and $\lambda_i(x_*)$ is the weight assigned to the value of y_i for the estimation of $\hat{y}_*$. Generally, the weights are defined such that data near to the point of interest are given more influence than those farther away.

The estimation variance σ_E^2 is used to quantify the accuracy of the estimation and is defined as:

$$\sigma_E^2 = \mathrm{Var}[y_* - \hat{y}_*] \quad (3)$$

y_* represents the true value at location x_*. Note that the actual magnitude of y_*, which is unknown, is not required in the calculation of σ_E^2 as expansion of Equation 3 will lead to requirement of only knowledge of the spatial covariance of the sampled data, which is known from the semivariogram.

To determine the optimal values for the weights, the Best Linear Unbiased Predictor (BLUP) criterion is used, where 'best' means that σ_E^2 is at its minimum and 'unbiased' means that the expected difference between the estimated and true values is zero (this constrains the sum of the weights to be one). Therefore, the weights are determined by solving the following optimisation problem:

$$\underset{\lambda_i}{\text{minimise}}\ \sigma_E^2 \ \text{subject to}\ \textstyle\sum_{i=1}^{N} \lambda_i = 1 \quad (4)$$

Solving Equation 4 for $\boldsymbol{\lambda}$ would lead to a linear system of equations, called 'kriging system', to solve.

After solving for the weights, $\hat{y}_*$ can be obtained from Equation 2. $\hat{y}_*$ is actually the mean of the kriging prediction. Specifically, the kriging prediction is normally distributed with a mean of $\hat{y}_*$ and a variance of σ_E^2.

2.1.4 *Theoretical variogram models*

For kriging, a large amount of effort goes into semivariogram modelling. An experimental (also known as empirical) semivariogram is typically first constructed from the sampled data to explore the correlation structure of the data. However, kriging requires theoretically valid semivariograms, which is not guaranteed for these experimental semivariograms. Thus, experimental semivarigorams are usually approximated using theoretical semivariogram or variogram (defined as twice the semivariogram) models that guarantees validity. Some popular theoretical variogram models include:

$$\gamma_{\mathrm{EXP}}(h) = c_n + c_0\left[1 - \exp\left(-\frac{h}{a}\right)\right] \quad (5)$$

$$\gamma_{\mathrm{GAU}}(h) = c_n + c_0\left[1 - \exp\left(-\left(\frac{h}{a}\right)^2\right)\right] \quad (6)$$

where Equations 5 and 6 are known as the exponential and Gaussian variogram model respectively. Here, c_n is the nugget, c_0 is the sill minus the nugget and a is a parameter that controls the range. In geotechnical engineering, the most widely used theo-

retical variogram model is the exponential model, mainly owing to its simplicity. The optimal values of the variogram parameters such as c_0 are determined by minimising the least squares error between the model and experimental variogram data.

2.1.5 *Kriging types*

Kriging assumes that the data is stationary i.e. the mean and covariance of the data depends only on separation, not on the locations. If the data is non-stationary, the data typically first go through a 'de-trending' transformation, where the trend in the data is removed, and kriging is applied to the residuals.

Various types of kriging have been developed over the years and among the most widely used are: (i) simple kriging for stationary data with known constant mean, (ii) ordinary kriging for stationary data with unknown constant mean, (iii) universal kriging for nonstationary data using a deterministic trend function, and (iv) cokriging for joint kriging of data from multiple correlated sources.

2.1.6 *Random field theory*

Kriging is closely related to random field theory (Vanmarcke 1977), which has been used extensively to characterise the spatial variability of soil properties (e.g. Lloret-Cabot et al., 2014). In random field theory (RFT), the spatial variability of soil properties is described by a random field, which is defined by the mean, variance and a correlation structure with a given scale of fluctuation θ.

The scale of fluctuation is the distance within which the properties are significantly correlated. The most common method to estimate the scale of fluctuation (e.g. Lloret-Cabot et al., 2014) is to best fit a theoretical correlation function to the experimental correlation data on a least squares sense. Common theoretical correlation functions include:

$$\rho_{\mathrm{EXP}}(h) = \exp\left(-2\frac{h}{\theta}\right) \tag{7}$$

$$\rho_{\mathrm{GAU}}(h) = \exp\left(-\pi\left(\frac{h}{\theta}\right)^2\right) \tag{8}$$

where Equations 7 and 8 are known as the Exponential and Gaussian correlation function respectively.

2.2 *Gaussian Process regression*

2.2.1 *Introduction*

A Gaussian Process (GP) is a stochastic process (i.e. a set of random variables) such that any finite number of them have a multivariate Gaussian distribution (Rasmussen and Williams 2006). As there is potentially an infinite number of random variables in a stochastic process, a GP can be intuitively thought of as an infinite-dimensional multivariate Gaussian distribution. Moreover, as a function can be thought of as an infinite-dimensional vector, a GP can also be thought of as a probability distribution over random functions. In GP regression, the output y of a function f at input x can be written as:

$$y = f(x) + \varepsilon \tag{9}$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is a noise term, which represents randomness such as measurement errors.

2.2.2 *GP regression model*

A GP prior distribution is assumed over the unknown function f. The GP is completely defined by its mean function $m(x)$ and covariance function $k(x, x')$ (also known as a kernel in ML literature):

$$f(x) \sim GP\big(m(x), k(x, x')\big) \tag{10}$$

where $m(x) = E[f(x)]$ and $k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$. $m(x)$ is the average of all possible functions in the distribution evaluated at input x, while $k(x, x')$ is the covariance between the function values at different inputs x and x'. If measured data (typically called 'observations' in the ML literature) are available, the GP prior distribution is updated by conditioning on the observations.

2.2.3 *Prediction*

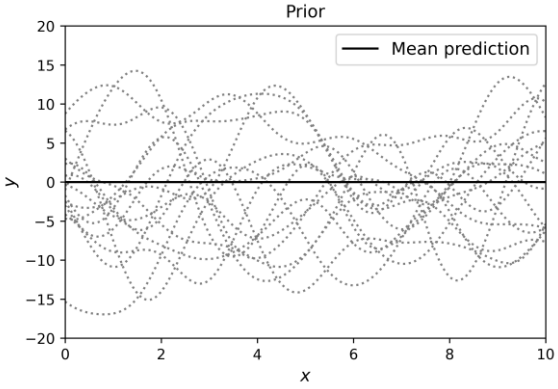
Suppose that some observations $\boldsymbol{y} = [y_1, \dots, y_N]^T$ have been obtained for some inputs $\boldsymbol{x} = [x_1, \dots, x_N]^T$. To predict the output $\hat{y}_*$ for a new input x_*, the GP regression model assumes that y_* is jointly Gaussian distributed with the observations $\boldsymbol{y}$. Thus, the model predicts the output $\hat{y}_*$ for an input x_* (given observations of $\boldsymbol{y}$) by computing the conditional distribution, which can be obtained analytically using the standard conditioning rules for the multivariate Gaussian distribution:

$$p(\hat{y}_* | x_*, \boldsymbol{x}, \boldsymbol{y}) = N(\mu_*, c_*) \tag{11}$$

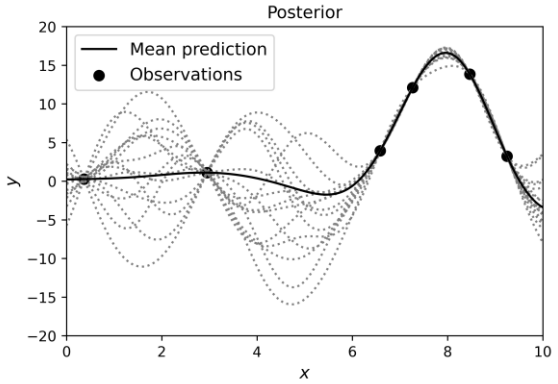
where

$$\mu_* = m(x_*) + \boldsymbol{k}_*{}^T \boldsymbol{K}^{-1}\big(\mathbf{y} - m(\boldsymbol{x})\big)$$
$$c_* = k(x_*, x_*) - \boldsymbol{k}_*{}^T \boldsymbol{K}^{-1} \boldsymbol{k}_*$$
$$\boldsymbol{k}_* = [k(x_1, x_*), \dots, k(x_N, x_*)]^T$$
$$\boldsymbol{K} = N\mathrm{x}N \text{ covariance matrix, } K_{ij} = k\big(x_i, x_j\big)$$

Equation 11 is called the predictive distribution and it showcases the key benefit of the GP regression model: it can provide the full probability distribution of the predictions, instead of merely pointwise predictions. To model noise in the observations, $\boldsymbol{K}$ may be replaced with $\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I}$, where σ_ε^2 is the variance of the noise and $\boldsymbol{I}$ is the identity matrix. Figure 2 shows the sampled functions obtained from a GP regression model, before and after observations are collected.



(a)



(b)

Figure 2. Example of sampled function values from GP (a) Prior (before observations) (b) Posterior (after observations). Grey dotted lines represent samples from the GP, and black solid line represents the mean. Black circles represent observations.

2.2.4 *Semi-parametric GP regression model*

Equation 11 only applies if $m(x)$ is known (e.g. zero for mean-centered data). If the functional form of $m(x)$ is known but the parameters are unknown (e.g. one may know that $m(x)$ varies linearly but the gradient and offset are unknown), the semi-parametric GP regression model (Rasmussen & Williams 2006), which combines a parametric model with a zero-mean GP regression model, may be used.

This model is defined as:

$$f_{SP}(x) = \boldsymbol{h}(x)^T\boldsymbol{\beta} + f_0(x) \tag{12}$$

where $f_0(x) \sim GP\big(0, k(x,x')\big)$, $\boldsymbol{h}(x)$ are some basis functions (e.g. monomials) and $\boldsymbol{\beta}$ are unknown parameters. Equation 12 uses a parametric model $\boldsymbol{h}(x)^T\boldsymbol{\beta}$ to approximate the unknown function and a zero-mean GP regression model $f_0(x)$ to estimate the residuals.

If $\boldsymbol{\beta} \sim N\big(\boldsymbol{\mu_\beta}, \boldsymbol{B}\big)$, the semi-parametric GP regression model is equivalent to (Rasmussen and Williams 2006):

$$f_{sp}(x) \sim GP\begin{pmatrix} \boldsymbol{h}(x)^T\boldsymbol{\mu_\beta}, \\ k(x,x') + \boldsymbol{h}(x)^T\boldsymbol{B}\boldsymbol{h}(x') \end{pmatrix} \tag{13}$$

The corresponding prediction of output $\hat{y}_*$ for an input x_* by this model is:

$$p(\hat{y}_*|x_*, \boldsymbol{x}, \boldsymbol{y}) = N(\mu_{**}, c_{**}) \tag{14}$$

where

$$\mu_{**} = \boldsymbol{h}_*^T\bar{\boldsymbol{b}} + \boldsymbol{k}_*{}^T\boldsymbol{K}^{-1}\big(\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{b}}\big)$$
$$c_{**} = c_* + \boldsymbol{r}^T(\boldsymbol{B}^{-1} + \boldsymbol{H}\boldsymbol{K}^{-1}\boldsymbol{H}^T)^{-1}\boldsymbol{r}$$
$$\bar{\boldsymbol{b}} = (\boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{H} + \boldsymbol{B}^{-1})^{-1}\big(\boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{y} + \boldsymbol{B}^{-1}\boldsymbol{\mu_\beta}\big)$$
$$\boldsymbol{r} = \boldsymbol{h}_* - \boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{k}_*$$

$\mathbf{H}$ is a matrix that collects the $\boldsymbol{h}(\boldsymbol{x})$ vectors for all data points and $\boldsymbol{h}_*$ is a vector for $\boldsymbol{h}(x_*)$. c_*, $\boldsymbol{k}_*$ and $\boldsymbol{K}$ are as defined in Equation 11.

2.2.5 *Kernels*

Several kernels have been developed and the most widely used include:

$$k_{\mathrm{SE}}(x,x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x-x'}{l}\right)^2\right) \tag{15}$$

$$k_{\mathrm{MAT}}(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|x-x'|}{l}\right)^{\nu} B_\nu\left(\frac{\sqrt{2\nu}|x-x'|}{l}\right) \tag{16}$$

where Equations 15 and 16 are known as the Squared Exponential (also known as the Radial Basis or Gaussian) and Matérn kernel respectively. σ_f^2 is a scaling factor that determines the variation of the function values from the mean value, while l is the lengthscale of the process (i.e. distance between inputs within which the outputs are highly correlated).

For multi-dimensional inputs, the lengthscale l can be assumed to be either identical or different for each input dimension. The kernel is called 'isotropic' and 'anisotropic' for the former and latter case respectively. ν is a parameter that governs the smoothness of the functions, B_ν is the modified Bessel function of the second kind and $\Gamma(\nu)$ is the gamma function evaluated at ν. As $\nu \to \infty$, the Matérn kernel becomes the Squared Exponential kernel, which produces very smooth functions.

The kernel is typically chosen to reflect one's prior knowledge or belief about the regression function such as its smoothness. For example, if the function is expected to be very smooth, the Squared Exponential kernel is selected. For rougher functions, the Matérn kernel is selected.

2.2.6 *Kernel hyperparameter learning*

The values of the kernel hyperparameters, such as σ_f^2 and l in Equation 15, are optimised by maximising the marginal log likelihood of the data given the hyperparameters (Rasmussen and Williams 2006).

3 Connections

3.1 *Lengthscale*

Kriging and GP regression are closely connected, although this is obscured by the different terminolo-

gies and conventions adopted in each field. Fundamentally, both theories assume that the data can be modelled using a Gaussian random field (which is synonymous with a Gaussian Process). Conceptually, the kriging range, the RFT scale of fluctuation and GP lengthscale are very similar. This is evident if you compare their definitions or observe the role each term plays in Equations 6, 8 and 15 respectively. In geotechnical engineering, the vertical and horizontal scale of fluctuations, which are usually different, are of particular interest. Using GP regression, the anisotropic kernel may be used to identify different lengthscales along the vertical and horizontal directions.

3.2 *Nugget*

The nugget is usually due to measurement errors or sources of variation at distances less than the sampling interval. The nugget is usually referred to in the ML literature as simply the noise in the data and the magnitude of this effect is the variance of the noise (i.e. σ_ε^2 for ε in Equation 9).

3.3 *Covariance function*

The RFT correlation function and GP kernel are analogous to each other (e.g. compare Equation 8 with Equation 15), although each operates on different levels (correlation versus covariance). The kriging variogram, on the other hand, is like the opposite of the correlation or covariance function. Comparing Equation 6 with Equation 8, as h gets larger, the variogram value increases while the correlation value decreases. Thus, the variogram and correlation/covariance function can be thought of as a dissimilarity and similarity function respectively.

For a second-order stationary process, a covariance function C (sometimes called covariogram or autocovariance function) may be obtained from a variogram (Wackernagel 2003):

$$C(h) = \gamma(\infty) - \gamma(h) \tag{17}$$

Note that $C(h)$ is analogous to the kernel $k(x, x')$ defined in the ML literature, except for the different input parameter (i.e. $h = |x - x'|$).

3.4 *Equivalent models*

For a second-order stationary process (where the covariance function is known to exist), the various types of kriging models can be shown to be special cases of the GP regression model. For simple kriging, the mean is a known constant μ_0 and the estimated output $\hat{y}_*$ for an input x_* can be obtained by applying Equation 2 to the de-trended data and thereafter adding back the known mean:

$$\hat{y}_* = \mu_0 + \boldsymbol{\lambda}_{\mathbf{SK}}^T(\mathbf{y} - \mu_0\mathbf{1}) \tag{18}$$

where $\mathbf{1}$ represents a vector of ones. The unique solution for the weights are $\boldsymbol{\lambda}_{\mathbf{SK}}^T = \boldsymbol{k}_*{}^T\boldsymbol{K}^{-1}$ (Cressie 1993). It can be observed that Equation 18 is equivalent to Equation 11 when $m(\boldsymbol{x}) = \mu_0$.

For ordinary kriging, the mean is an unknown constant μ_1 and the estimated output $\hat{y}_*$ for an input x_* is obtained using Equation 2 with the weights $\boldsymbol{\lambda}_{\mathbf{OK}}^T = [\boldsymbol{k}_* + \mathbf{1}(\mathbf{1}^T\boldsymbol{K}^{-1}\mathbf{1})^{-1}(\mathbf{1} - \mathbf{1}^T\boldsymbol{K}^{-1}\boldsymbol{k}_*)]^T\boldsymbol{K}^{-1}$ (Cressie 1993). This is a special case of the semi-parametric GP regression model in Equation 14, when $\boldsymbol{h}(x) = 1$, $\boldsymbol{\mu_\beta} = \mu_1$ and $\boldsymbol{B} = 0$ (since μ_1 is deterministic); this gives $\mu_{**} = \mathbf{1}^T\bar{\boldsymbol{b}} + \boldsymbol{k}_*{}^T\boldsymbol{K}^{-1}(\mathbf{y} - \mathbf{1}\bar{\boldsymbol{b}})$ and $\bar{\boldsymbol{b}} = (\mathbf{1}^T\boldsymbol{K}^{-1}\mathbf{1})^{-1}(\mathbf{1}^T\boldsymbol{K}^{-1}\boldsymbol{y})$, which can be rearranged to get $\mu_{**} = \boldsymbol{\lambda}_{\mathbf{OK}}^T\boldsymbol{y}$.

For universal kriging, the mean is an unknown deterministic trend function. The estimated output $\hat{y}_*$ for an input x_* can similarly be obtained using Equation 2 with the weights (Cressie 1993) $\boldsymbol{\lambda}_{\mathbf{UK}}^T = [\boldsymbol{k}_* + \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{H})^{-1}(\boldsymbol{h}_* - \boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{k}_*)]^T\boldsymbol{K}^{-1}$. This is also a special case of Equation 14, when $\boldsymbol{B} = 0$ (since the trend function is deterministic); this gives $\mu_{**} = \boldsymbol{h}_*^T\bar{\boldsymbol{b}} + \boldsymbol{k}_*{}^T\boldsymbol{K}^{-1}(\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{b}})$ and $\bar{\boldsymbol{b}} = (\boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{H})^{-1}(\boldsymbol{H}^T\boldsymbol{K}^{-1}\boldsymbol{y})$, which can be rearranged to get $\mu_{**} = \boldsymbol{\lambda}_{\mathbf{UK}}^T\boldsymbol{y}$. Table 1 lists the correspondence (not necessarily equivalence) of various terms in the two theories.

Table 1. Correspondence of terms in kriging and GP regression

Kriging	**GP regression**
Range / Scale of fluctuation	Lengthscale
Nugget	Gaussian noise variance in observations
Covariance / Covariogram / Autocovariance function	Kernel
Exponential correlation function	Laplace / Exponential kernel
Gaussian correlation function	Squared Exponential / Gaussian / Radial Basis kernel
Whittle-Matérn correlation function	Matérn kernel
Simple kriging	GP regression
Ordinary / Universal kriging	Semi-parametric GP regression
Co-kriging	Multi-output GP regression

3.1 *Differences*

While both theories have been shown to be closely related, there are some philosophical and practical differences between kriging and GP regression. First, kriging methods are founded on optimisation notions (e.g. BLUP criterion), while GP methods are

founded on Bayesian notions (e.g. assume a prior distribution over functions and then updating this distribution by conditioning on the observations). This difference in the derivation of the models is usually the main source of confusion.

Moreover, the kriging approach for achieving the best fit to observations involves a two-step process. Firstly, the observations are transformed into an intermediate dataset (experimental variogram data). Then, the parameters of the theoretical variogram are optimised by minimising the least squares errors between the theoretical variogram and the experimental variogram data. On the other hand, the GP regression approach optimises the kernel hyperparameters by maximising the marginal log likelihood of the observations, without the need for an intermediate dataset transformation. Since these two approaches optimise for different objectives, the optimal parameters obtained through kriging and GP regression may differ, leading to different predictions. This distinction is illustrated in Figure 3.

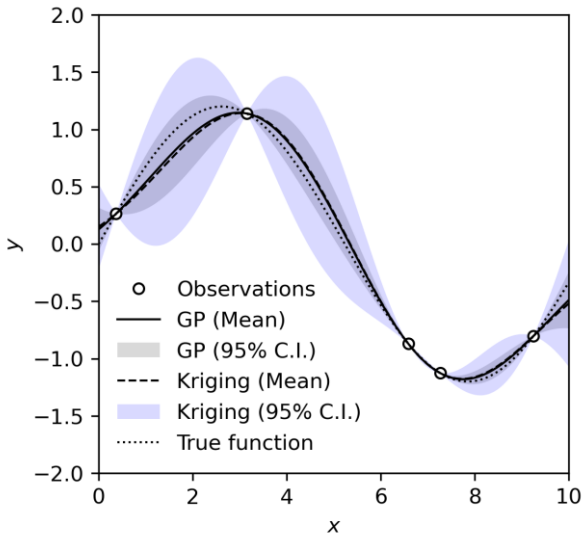


Figure 3. Mean predictions of simple kriging and GP regression, and their corresponding 95% confidence interval.

4 Applications beyond spatial interpolation

Kriging applications in geotechnical engineering are predominantly spatial interpolation. To demonstrate the benefits of knowledge transfer from the ML field, the following describes some common ML applications that could be applied in geotechnics.

4.1 *Automated feature selection*

A GP regression application that is often applied in ML is automated feature selection via Automated Relevance Determination (ARD) (Rasmussen & Williams 2006), where an anisotropic kernel is used to learn the lengthscale for each input based on the data and the learnt lengthscales are regarded as importance measures. Inputs with very large lengthscales (relative to the typical scale of the sampled data e.g. $|\boldsymbol{x}_i - \boldsymbol{x}_i'|$ values) result in the covariance being almost independent of those inputs, making them effectively irrelevant for the prediction of the output. To illustrate this, a zero-mean GP regression model with the anisotropic Matérn ($\nu = 5/2$) covariance function is used to predict the undrained vertical capacity for a suction caisson foundation of different 'length to diameter' ratios (L/D) and 'soil Young's modulus to undrained shear strength' ratios (E/s_u). A dataset that was previously used to develop failure envelopes and Winkler models for suction caisson foundations (Suryasentana et al. 2018, 2019, 2020, 2021, 2022) was used to train the GP regression model, where the inputs for the current study are the two ratios (L/D and E/s_u) and the output is the vertical capacities of suction caisson foundations computed using finite element (FE) analysis.

Figure 3 shows the computed FE capacities for different inputs, where it can be observed visually that E/s_u has a negligible effect on the capacities. Figure 4 shows the mean capacities (and the 95% confidence interval) predicted by the GP regression model for a range of inputs ($0 \leq L/D \leq 2$ and $E/s_u = \{100,\ 500,\ 1000\}$), which shows increased prediction uncertainty for values of L/D with no nearby training data. The learnt lengthscales for L/D and E/s_u are 1.6 and 15000 (which is much larger than the typical scale of the E/s_u training data) respectively; this indicates that the GP regression model identifies E/s_u as an input that is irrelevant for the prediction task. This identification is logical from an engineering perspective, as the ultimate capacity of the foundation is independent of the stiffness of the soil. Automated feature selection is useful for geotechnical engineering problems with a large input dimension space, where it is more difficult to visualise the effect of each input on the output. For example, this could be applied to identify which of the multitude of sensor measurements are relevant for predicting some geotechnical output.

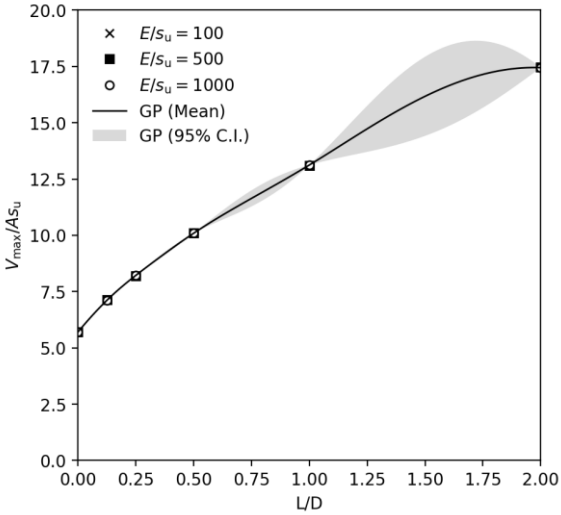


Figure 4. Vertical capacities (normalised by As_u) of a suction caisson (where A is the cross-sectional area of the caisson). Note that the GP regression model predictions for different E/s_u values overlap each other).

4.1 *Forecasting with composite kernels*

A common approach in ML that has not been widely adopted in geotechnical engineering is the composition of kernels to create custom, composite kernels. This is a powerful tool as it provides a set of reusable building blocks (Duvenaud et al. 2013) for complex regression problems. This is particularly useful for the task of forecasting. With an appropriate choice of covariance function, GP regression may also be used reliably for extrapolation or forecasting over a short horizon (in general, extrapolation is reliable only for one unit of lengthscale away from the data). For example, Sheil et al. (2020a, b) has demonstrated the use of a GP regression model (with the Matérn covariance function) to forecast near-term jacking forces as an intermediate step towards detecting anomalies during microtunnelling.

To demonstrate the influence of the kernel composition on the forecasting abilities of the GP regression model, three different covariance functions are used to forecast the accumulated axial strain for Karlsruhe Kaolin clay under undrained cyclic triaxial loading (Wichtmann & Triantafyllidis 2018). A zero-mean function is used for all three GP regression models and the three covariance functions employed are: a Squared Exponential (SE) covariance function, a custom covariance function which is the sum of a linear covariance function and an SE covariance (LE+SE), and another custom covariance function which is the sum of a linear covariance function and the product of an SE and a periodic covariance function (LE+SE*PER).

Figure 5 shows the measured axial strain for an undrained cyclic triaxial test on Karlsruhe Kaolin clay, with initial mean stress $p_0 = 200$ kPa, initial deviatoric stress $q_0 = 100$ kPa and cyclic deviatoric stress amplitude $q_{\text{amplitude}} = 30$ kPa. The figure also shows the forecast predictions of the GP regression models with the three covariance functions. The results indicate that the SE covariance function does not perform well for the extrapolation task as the forecast tends to revert back to the (assumed zero prior) mean; the time interval over which this occurs is governed by the lengthscale learnt from the training dataset. The second model (LE+SE covariance function) performs better than the first by capturing the global linear trend, but it fails to capture the local periodic trend. The third model (LE+SE*PER covariance function) provides the best performance by capturing both the global linear trend and the local periodic trend. The optimised log marginal likelihood values obtained for the three GP regression models are 214.46, 232.66 and 328.03 for the GP regression models with the SE, LE+SE and LE+SE*PER covariance function respectively; comparing the log marginal likelihood values indicate that the covariance function that is most appropriate for this problem is the LE+SE*PER covariance function, as it has the highest likelihood value.

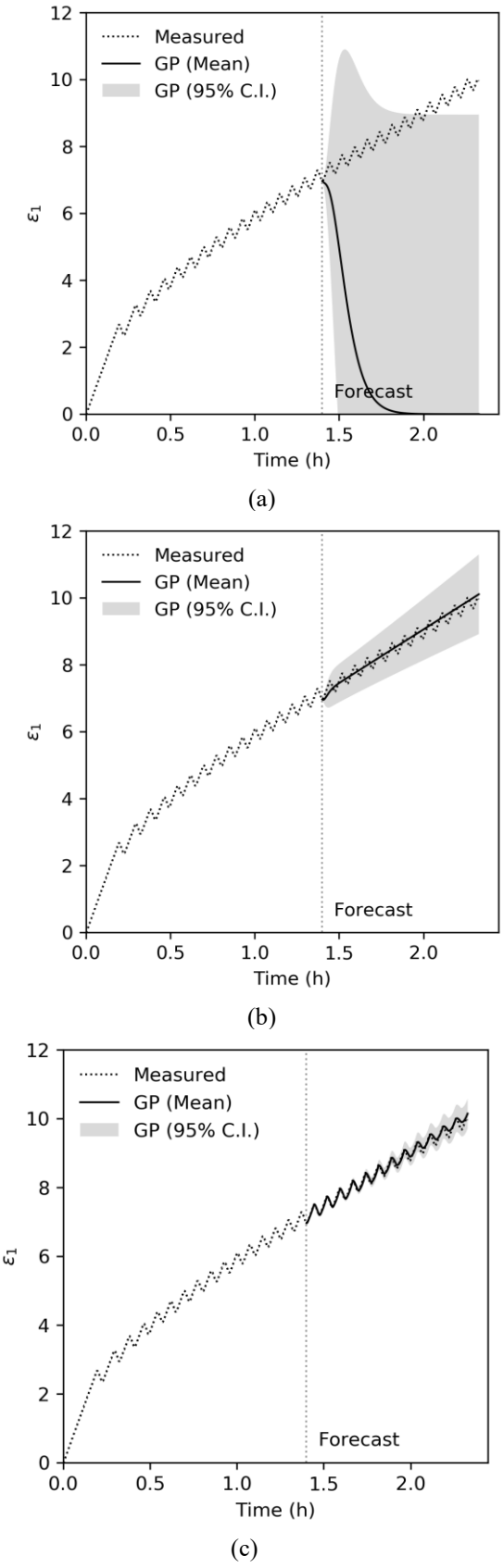


Figure 5. Comparison of the measured axial strain for an undrained cyclic triaxial test on clay and the forecasted axial strain (shaded bounds are the 95% confidence interval of the forecast) by the GP regression model predictions with the (a) SE covariance function, (b) LE+SE covariance function, and (c) LE+SE*PER covariance function.

An important aspect of forecasting is not only predicting future values but also expressing the uncertainty associated with those predicted values. Figure 5 shows that the confidence interval increases over time, which expresses increasing uncertainty for forecasts over a longer time horizon. Figure 6 provides an alternative view of Figure 5c in the deviatoric stress-axial strain space.

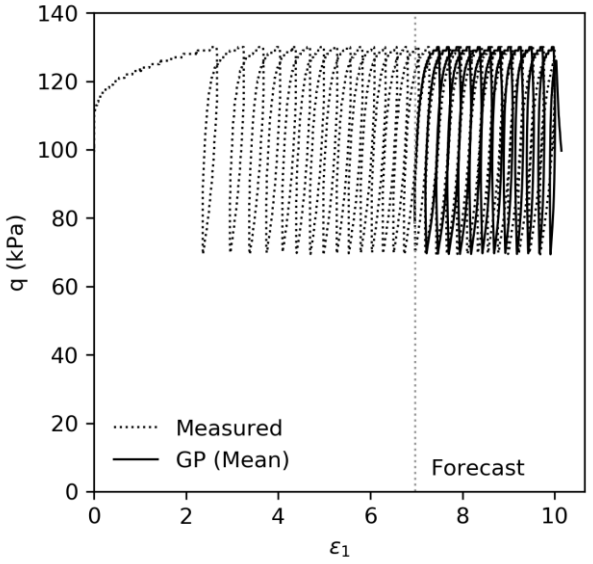


Figure 6. Comparison of the deviatoric stress-axial strain measurements for an undrained cyclic triaxial test on Karlsruhe Kaolin clay with the mean prediction of the GP regression model with the LE+SE*PER covariance function.

5 Conclusions

This paper provides a review of the kriging and GP regression theories and unpacks some of their equivalences and differences. It is important to understand these connections to facilitate the transfer of knowledge from one field to the other. It is hoped that this paper is a useful contribution towards developing a bridge between the two fields, which will hopefully lead to further advances in each.

6 References

Cressie, N. (1993). Statistics for Spatial Data, Wiley, New York, NY.

Deisenroth, M. P., Fox, D., and Rasmussen, C. E. (2013). Gaussian processes for data-efficient learning in robotics and control. IEEE transactions on pattern analysis and machine intelligence, 37(2), 408-423.

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., and Zoubin, G. (2013, May). Structure discovery in nonparametric regression through compositional kernel search. In International Conference on Machine Learning (pp. 1166-1174). PMLR.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. J. Chem. Metall. Min. Soc. South Africa, 52(6), 119–139.

Li, J., Cassidy, M. J., Huang, J., Zhang, L., and Kelly, R. (2016). Probabilistic identification of soil stratification. Géotechnique, 66(1), 16-26.

Lloret-Cabot, M. F. G. A., Fenton, G. A., and Hicks, M. A. (2014). On the estimation of scale of fluctuation in geostatistics. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 8(2), 129-140.

Matheron, G. (1963). Principles of geostatistics. Economic geology, 58(8), 1246-1266.

Rasmussen, C. E. and Williams C. K. I. (2006). Gaussian processes for machine learning. The MIT Press, Cambridge, MA.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. Phil. Trans. of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110550.

Sheil, B. B., Suryasentana, S. K., & Cheng, W. C. (2020a). Assessment of anomaly detection methods applied to microtunneling. Journal of Geotechnical and Geoenvironmental Engineering, 146(9), 04020094.

Sheil, B. B., Suryasentana, S. K., Mooney, M. A., & Zhu, H. (2020b). Machine learning to inform tunnelling operations: recent advances and future trends. Proceedings of the ICE, 173(4), 74-95.

Suryasentana, S. K., Byrne, B. W., Burd, H. J., & Shonberg, A. (2018). An elastoplastic 1D Winkler model for suction caisson foundations under combined loading. In Numerical Methods in Geotechnical Engineering IX (pp. 973-980). CRC Press.

Suryasentana, S. K., Burd, H. J., Byrne, B. W. & Shonberg, A. (2019) A Systematic Framework for Formulating Convex Failure Envelopes in Multiple Loading Dimensions. Géotechnique 70(4), 343-353.

Suryasentana, S. K., Dunne, H. P., Martin, C. M., Burd, H. J., Byrne, B. W., & Shonberg, A. (2020). Assessment of numerical procedures for determining shallow foundation failure envelopes. Géotechnique, 70(1), 60-70.

Suryasentana, S. K., Burd, H. J., Byrne, B. W., & Shonberg, A. (2021). Automated procedure to derive convex failure envelope formulations for circular surface foundations under six degrees of freedom loading. Computers and Geotechnics, 137, 104174.

Suryasentana, S. K., Burd, H. J., Byrne, B. W., & Shonberg, A. (2022). A Winkler model for suction caisson foundations in homogeneous and non-homogeneous linear elastic soil. Géotechnique, 72(5), 407-423.

Vanmarcke, E. H. (1977). Probabilistic Modeling of Soil Profiles. Journal of the Geotechnical Eng. Div. 103 (11): 1227–1246.

Wackernagel, H. (2013). Multivariate geostatistics: an introduction with applications. Springer Science Media, NY.

Wichtmann, T., Triantafyllidis, T. (2018). Monotonic and cyclic tests on Kaolin - a data base for the development, calibration and verification of constitutive models for cohesive soils with focus to cyclic loading. Acta Geotechnica, Vol. 13, No. 5, pp. 1103-1128.