Asian Journal of Civil Engineering (2024) 25:5599–5612
https://doi.org/10.1007/s42107-024-01132-7

RESEARCH

Advanced modeling techniques using hierarchical gaussian process regression in civil engineering

Amani Assolie[1]

Received: 1 July 2024 / Accepted: 19 July 2024 / Published online: 6 August 2024

Abstract

Gaussian process regression (GPR) models, with their desirable mathematical properties and outstanding practical performance, are increasingly favored in statistics, engineering, and other domains. Despite their advantages, challenges arise when applying GPR to extensive datasets with repeated observations. This study aims to develop models for predicting Finland's soft-sensitive clays' undrained shear strength (Su). The study presents the first correlation equations for Su of Finnish clays, derived from a multivariate dataset compiled using field and laboratory measurements from 24 locations across Finland. The dataset includes key parameters such as Su from field vane tests, reconsolidation stress, vertical effective stress, liquid limit, plastic limit, natural water content, and sensitivity. The GPR model demonstrated high accuracy, with a mean squared error (MSE) of 0.11% and a correlation coefficient (R^2) of 0.98, indicating excellent predictive performance. These findings highlight the strong interactions between Su, consolidation stresses, and index parameters, establishing a robust foundation for practical GPR implementation. The GPR model is recommended for forecasting Su due to its high learning performance and ability to display prediction outputs and intervals. This research has significant implications for various civil engineering applications, including transportation, geotechnical, construction, and structural engineering, offering a valuable tool for improving engineering practices and decision-making.

Keywords Multivariate database · Civil Engineering · Machine learning · Gaussian process regression · Prediction

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) applications in civil engineering have revolutionized various subfields, such as structural, transportation, and geotechnical engineering. In structural engineering, AI and ML predict material behaviors and structural responses under different load conditions. In transportation engineering, these technologies aid in forecasting traffic patterns and optimizing transportation networks. Geotechnical engineering benefits from AI and ML through enhanced modeling of soil behaviors and predicting ground interactions. One prominent AI technique in these fields is Gaussian Process Regression (GPR), which has shown high accuracy in predicting material properties and behaviors. GPR's unique ability to provide estimates and prediction intervals makes it a valuable tool for ensuring reliability and trustworthiness in civil engineering applications. The GPR model, with its high accuracy and practical performance, offers a reliable method for predicting the undrained shear strength (Su) of Finland's soft-sensitive clays, providing reassurance to civil engineers, researchers, and practitioners in the fields of transportation, geotechnical, construction, and structural engineering.

Machine learning involves developing methods for automatically acquiring new knowledge from existing data. In this context, algorithms, theory, and expression analysis are paramount, often surpassing traditional statistical methods. Supervised learning, which includes regression (yielding continuous results) and classification (producing discrete outputs), is prevalent among the various machine learning techniques. When applied to complex datasets, supervised machine learning and data analytics become essential tools (Kaveh & Bakhshpoori, 2019; Kaveh & Eslamlou, 2020; Verma, 2023; Zhang et al., 2020, 2021).

Gaussian Process Regression (GPR) has emerged as a highly effective regression solution for engineering problems, gaining favor for its ability to anticipate a range of

✉ Amani Assolie
AM.assolie@anu.edu.jo

[1] Department of Civil Engineering, Faculty of Engineering, Ajloun National University, Ajloun 26810, Jordan

functions with a fundamental structure, quickly interpretable parameters, and Bayesian conclusions (Monisha & Balasubramanian, 2023; Williams & Rasmussen, 2006). GPR models are particularly valued for their high learning performance and inherent capacity to display prediction outputs and intervals. These qualities make GPR an excellent choice for solving various engineering issues through controlled learning and for use in supervised machine learning for classification and regression analysis (Djeziriand & Bendahan, 2021).

GPR has extensive applications in civil engineering, including structural engineering, transportation, construction, and geotechnical engineering. For instance, GPR can predict material behaviors and structural responses under different load conditions in structural engineering. In transportation engineering, GPR aids in forecasting traffic patterns and optimizing transportation networks. Construction engineering can predict project outcomes and assess risks, while geotechnical engineering assists in modeling soil behaviors and predicting ground interactions (Cai et al., 2020; Calandra et al., 2016).

One of the most effective non-parametric regression approaches is the Gaussian Process (GP), which utilizes a covariance function to model high-level function assumptions such as smoothness and periodicity. Selecting the appropriate covariance function is crucial for data analysis, with the smooth and stationary square-exponential (Gaussian) covariance function being commonly assumed. However, general covariance measures may not be suitable for all applications, such as modeling robot locomotive ground interactions (Cai et al., 2020; Calandra et al., 2016).

Gaussian processes are utilized in statistical and machine-learning models to generate distributions without specifying a specific functional form (Schulz et al., 2018). They are employed for time-series data (survival analysis), regression, classification, improved learning, and spatial models. While simple examples suggest that basic ideas are straightforward to apply, more complex versions are often required. The mean and covariance functions of the Gaussian process are intricate in describing a priori and are expressed as hyperparameters.

As Rasmussen and Nickisch (2010) discussed, the probability function presents particular challenges. This research addresses these limitations through solutions implemented using existing interpretations, covariance, probability functions, and inference methodologies. Specifically, this study employs Gaussian regression in conjunction with unknown variance models to handle varied difficulties. The stochastic supervised Gaussian process is extensively used in regression analysis to analyze uncertainty and predict values, even with observable inputs that have unknown uncertainty (D'Ignazio et al., 2016).

This study has two primary objectives. First, it aims to incorporate variance uncertainty into Gaussian process regression. Second, it compares the model's performance using cross-entropy and mean square losses. By addressing these objectives, the research contributes to the existing knowledge and application of GPR in civil engineering, providing a robust framework for future studies and practical implementations. Additionally, this study highlights the application of artificial intelligence and machine learning in various civil engineering fields, such as structural engineering, transportation engineering, and geotechnical engineering. It demonstrates the effectiveness of GPR in predicting material properties and behaviors, emphasizing its ability to provide accurate estimates and prediction intervals. This research enhances the understanding of GPR in handling engineering datasets and opens avenues for future applications and improvements in AI-driven civil engineering solutions.

Methods and materials

Standard gaussian process regression

The Gaussian process (Rasmussen, 2004) is a powerful machine-learning tool. It improves prediction accuracy by using past data. Function adaptation is the most obvious. Robotics and time series forecasting use regression (Lu et al., 2010; Mackay, 1992) for classifying and clustering data. An endless number of models can be fitted to training points. Gaussian processes provide a solution to this problem (Rasmussen, 2004). The normal distribution best describes the data. In the regression model, a probabilistic method increases forecast confidence.

The fundamental stochastic process is a function-based extension of a Gaussian process distribution, and the first premise of process modeling is to identify common event sequences. The model can also be used to forecast or isolate process aspects. Gaussian processes simplify deductively and learn computations. The Gaussian technique can immediately supervise machine-learning issues or case-based learning. This section evaluates and applies the first model, which represents the original Gaussian process regression, by employing a novel process or expanding the GPR toolbox. The researcher investigates novel approaches and strives to increase model accuracy. This chapter discusses the Gaussian process regression model, which will be covered in subsequent chapters. GPR is introduced as a prior over random functions, likelihood, and posterior over functions given observed data, a data modeling tool, a machine learning alternative, and a flexible nonparametric regression. Despite the technical overanalysis and weird verbiage, GP regression is only an extension of linear modeling.

Theoretical framework of gaussian process model (GPR)

Gaussian process regression is a standard, probabilistically controlled machine learning method. Gaussian regression models estimate uncertainty. Nearly a century of research into supervised learning problems led to the "well theory," the present state of the art in statistics. Thanks to cheap and fast computation, machine learning has solved more complex issues (Jordan & Mitchell, 2015). Statistical and machine-learning organizations share theories and approaches. The nature of the issue and the instructional focus may be essential differentiators. Statistics studies data, models, and approximations such as linear and indigenous relationships. Machine learning aims to foresee and comprehend learning algorithms accurately. As black-box function approximations in machine learning, many statisticians find neural network techniques inadequate (Williams & Rasmussen, 2006). This is just one way the two fields have grown differently due to their varied goals.

Due to its theoretical similarities to the Bayesian linear model, neural networks, and large neural networks, the Gaussian process is related to support vector machines (under specific conditions). Gaussian models may be simpler to use and comprehend than neural networks. Although Gaussian processes have received much attention in statistics, their widespread use is probably improper outside computer analysis and meteorological and geological spatial models (Quadrianto et al, 2010). To understand Gaussian processes, one must first understand their underlying arithmetic. Gaussian methods are based on Gaussian (or regular) distributions. The joint distribution of multivariate average data is Gaussian. The mean value is clustered in multivariate normal distributions, multivariate Gaussian distributions, and joint normal distributions (Hoang et al., 2016a, 2016b).

Structure of gaussian process regression model

Regression and classification are two types of supervised learning. Continuous regression is used to predict continuous values, whereas classification is used to predict discrete class labels. Future commodity prices can be predicted in a financial application using interest rates, currency exchange rates, and supply and demand. This section employs a Gaussian process regression on 217 testing models to demonstrate the shear power of undrained soft Finnish clay.

A Gaussian process is a multivariate probability distribution in which all finite-dimensional distributions are Gaussian. A Gaussian process is used to construct a probability distribution between functions: values taken at random locations xi (i = 1)N X have an N-variate Gaussian distribution. The distribution of xi is a Gaussian random field. Supervised learning learns a function that transforms input to output from a dataset of input–output pairs. Figure 1 shows that the training data is used to forecast new inputs. The researcher employed a dataset of known input–output pairings to predict unseen inputs x'.

The first and most crucial step toward Gaussian process regression is to obtain the mean function m(x) and the covariance function k (x, x′) because these two functions define and characterize the Gaussian process entirely. These functions are provided independently, each with a set of parameters called hyperparameters and a functional form definition. Knowing the mean and covariance functions, you can figure out the model's hyperparameters by minimizing the log's marginal chance (Williams & Rasmussen, 2006).

Implementing kernel functions makes the GPR particularly efficient at handling nonlinear data. In addition, GPR has an essential advantage in providing a reliable answer to the input data (Pal & Deswal, 2010). The mean function m(x) and the covariance function k (x, x′) adequately describe a Gaussian process. Because the Gaussian distribution is over vectors and the Gaussian process is over functions, this is a general form of the Gaussian.

Distribution, with mean and covariance represented by vectors and matrices, respectively. While working with infinite-dimensional items may appear challenging at first, it turns out that one only works with finite-dimensional items when calculating amounts. In practice, answering the

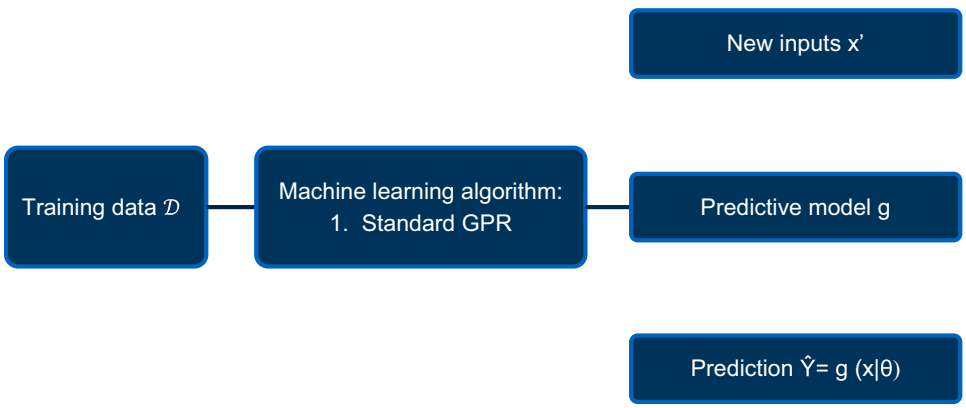

Fig.1 Explanation of machine learning process in GPR

process questions reduces the data sent to the computer. This is the key to enabling Gaussian processes. As shown below, one example where writing is conceivable:

$$f \sim \mathrm{GP}(m, k)$$

The fact that the function is distributed as a GP with a mean function m and a covariance function k is referred to by this phrase. Although the shift from distribution to process is simple, the researcher will go over the intricacies in greater detail because some readers may be unfamiliar with them. Individual Gaussian distribution vector random variables are indexed by their position in the vector. For the Gaussian process, it is the x (of the random function. $f(x)$) argument that performs the role of the index set: the related random variable $f(x)$ is present at each input x, which is the value of f (stochastic) at that point. For notational convenience, the researcher will enumerate the x values of interest by the natural numbers and use these indexes as if they were the process indices–do not be confused: the process index is x_i, which the researcher decided to index by i.

There has been extensive research and implementation of the simple linear regression model, which produces a linear combination of inputs. The main advantages of its use and interpretability are straightforward. The main issue is that only limited flexibility is permitted; if a linear function cannot accurately describe the relationship between input and output, the model will produce inaccurate forecasts. This section discusses the Bayesian treatment of the linear model. Then, the linear model is used to enhance this model class by projecting the entries into a huge functional space. Given a training set $D=\{(x_i, y_i) \mid i=1,..., n\}$, the input data $X \in \mathbb{R}^{D\times n}$ is called the design matrix and $y \in \mathbb{R}^{n}$ is the vector of the desired output. The central assumption of GPR is that the output y is computed as (Ebden, 2015; Williams & Rasmussen, 2006):

$$y = f(x) + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, \sigma_n^2) \in \mathbb{R}$ represents all samples a homoscedastic noise x_i. According to the GPR approach, the n observations in the data set of interest $y=\{y_1, \ldots\ldots, y_n\}$, a single point from a Gaussian multivariate distribution is considered. Besides, the mean of this Gaussian distribution is zeros. The covariance function $k(x, x')$ determines the relationship of one observation to another. The squared exponential covariance function in GPR is frequently chosen for function approximation (Cheng et al., 2013; Stahl, 2006):

$$k(x, x_i) = \sigma_f^2 \times \exp\left(-\frac{(x-x')^2}{2l^2}\right) + \sigma_f^2(x, x') \tag{2}$$

where the maximum allowable covariance is defined as σ_f^2, it is noted that $k(x, x')$ reaches this maximum allowable covariance only when x is very close to x' and therefore (x) is almost perfectly correlated with $f(x')$. Meanwhile, l denotes the length parameter of the kernel function. In addition, $\delta(x, x')$ represents a Kroenke delta function; $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i \neq j$.

The kernel (or covariance function) defines the Gaussian process's covariance. Gaussian processes are covariance-dependent (sometimes called kernels). All assumptions regarding its function are encoded. High covariance suggests proximity or resemblance. Say you have two input points, xi and xj, and two data points, yi and yj. If xi and j are close, so will yi and yj. Covariance (Abdessalem et al., 2017)measures similarity.

Synchrony refers to the interaction of two items in kernels. This could be misleading because the similarity of function values is explicitly stated. The kernel defines which functions will be employed when applying the prior GP, influencing a model's generalizability. The researcher begins by considering squared-exponential (SE), periodic (Per), and linear priors. Figure 2 depicts the kernel definitions.

One option is the squared-exponential covariance function (but not the only one). The advantage of the previous equation is significant because the covariance is expressed as a function of the inputs. It should be observed that the exponential covariance takes nearly unit values among variables when the inputs are relatively close to one another and begins to diminish as the variable distance in the input space increases. The kernel idea predetermines the Gaussian distribution of processes $k(x, x')$.

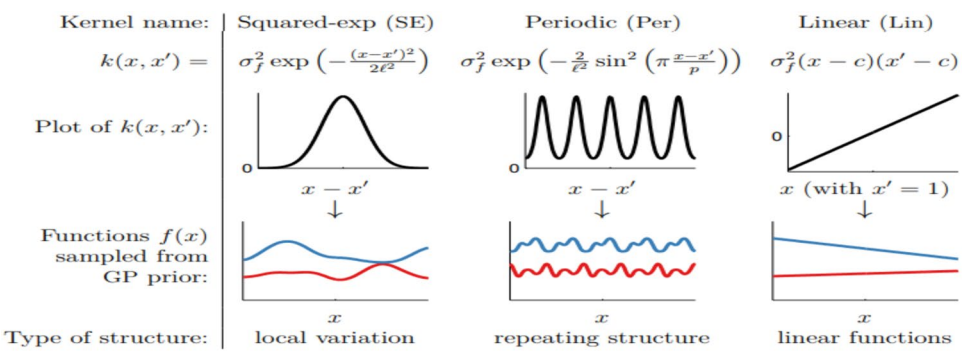

Fig.2 Explanation of kernel types definitions (Duvenaud, 2014)

The study codes are conducted using MATLAB R2020a, which divides codes into five stages, as shown in Fig. 3 below, for the first model (initial Gaussian process regression) and the second model (unknown variance Gaussian process regression). The starting stage in Gaussian regression is creating a prior mean function. $m(x)$ and covariance function $k(x, x')$ alternatively, as GPs are fully defined in the input vector. Often, due to practical reasons (simplicity) and a lack of prior understanding of the general data trend, the average function is considered zero for practical reasons. The Gaussian processes can be characterized as follows:

$$f \sim GP(0, k) \tag{3}$$

In summary, a Gaussian process prior (Joint distribution) is a prior overall sufficiently smooth function. f; data then select the suitable fitting functions from this prior, accessed via a new variable known as the "predictive posterior" or the "predictive distribution." When using noisy observations, the mean function is set to zero. The GP only mentions which means before x. The likelihood is the connection between the GP's random variables and the actual observations Y. Using the prior and the likelihood; the posterior can be computed. The posterior is then used to generate predictions. In a GP, one must estimate the posterior from the prior and the likelihood. The posterior is constructed from the same set of random variables as the prior. The posterior Gaussian process is also known as the posterior predictive or conditional distribution.

Gaussian process prior

The Gaussian process is a broad term that appears in various statistical and probabilistic modeling organizations, with various but highly particular meanings. A finite collection of achievements (i.e., n observations) is modeled using a multivariate normal division (MVN). In consequence, this indicates that their mean function. $m(x)$ and covariance matrix $k(x, x')$ describe the qualities of these realizations completely.

Regarding the mean function, the mean function $m(x)$, provides the mean of the random values. The use of $m(x)$ The expected value is also called to indicate the mean of the random variables (x) at location x. Because x denotes the number of places, $m(x)$ is a vector which contains mean values for random variables. This function m Is in the domain and range $\mathbb{R}$, i.e. m: $\mathbb{R} = \mathbb{R}$. The researcher has a mean value vector when it applies. m In each of the above places:

$$\begin{bmatrix} m(x) \\ m(x') \end{bmatrix} \tag{4}$$

But in this study, the mean is defined as a zero function $m(x) = 0$. It returns 0 regardless of the value of x. One could wonder why a mean null is appropriate when real data may not have a mean nil. The data can always be normalized, for example, if the mean of the data is zero. This works in most cases, but it must be more difficult to occasionally consider a non-zero mean feature to build a better model.

The covariance matrix is a covariation between a random $f(x)$ and a pair of $f(x')$ variables. x & x' can be X or X* from the source set. By providing this covariance among $f(x)$ and $f(x')$, one must identify a function $k(x, x')$.

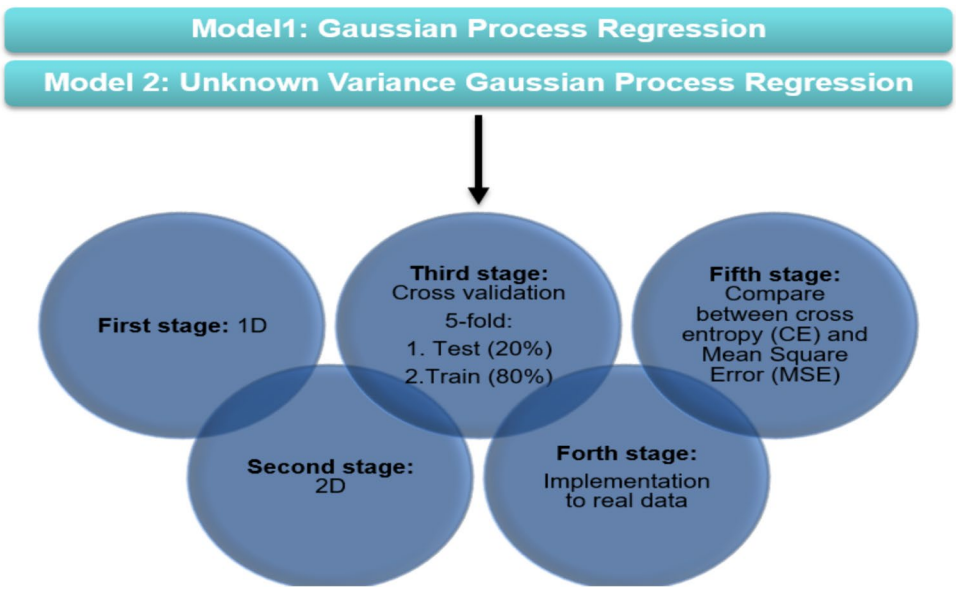

Fig.3 Flow chart of MATLAB work stages

Numerous functions of two x and x' inputs meet this study's criterion that a more excellent value is returned when X and X′ are close to each other and that when they are distant, they yield a smaller value. Moreover, the following is what is focused on:

$$k(x,x_i) = \sigma^2 \exp -\left(\frac{(x-x')^2}{2l^2}\right) + \sigma n^2\, \delta(x,x') \tag{5}$$

An exponential function is in this formula. l Is the long scale, and σ^2 is the variance, which is scalars. The parameters of the model are the values for the parameters of these models that are unknown; it utilizes the learning parameter to identify good values. Due to its structure, this kernel function is dubbed the exponential squared kernel. Note that just x and x' are mentioned in the kernel function. The function value Y is not mentioned at all. This is because having x and x' is sufficient to determine the distance between the two locations. Overall, it can be verified that specific requirements like the exponential function are positive in this kernel function, so covariance has always been positive. In addition, when x equals x, k has a maximum σ^2, so its distance is 0 as low as 0. In addition, the exponential function evaluates to 1. Hence, it evaluates to σ^2 in the kernel function.

The Gaussian rule of marginalization informs that the emphasized sections are also a Gaussian multivariate distribution. This distribution is based on random $f(x)$ and $f(x')$ variables with mean and covariance matrix as shown in the red boxes. So from now on, the researcher can work with the following finite distribution, which represents the Prior Gaussian Process:

$$\begin{bmatrix} f \\ f' \end{bmatrix} \sim N\left(\begin{bmatrix} m(x) \\ m(x') \end{bmatrix}, \begin{bmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{bmatrix}\right) \tag{6}$$

The following represents the Prior GP using noisy observation:

$$\begin{bmatrix} y \\ f' \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x,x)+\sigma_N^2 I_N & k(x,x') \\ k(x',x) & k(x',x') \end{bmatrix}\right) \tag{7}$$

Rather than focusing on the training data when developing the GP, the mathematical conveniences of the distributions are studied and selected by the researcher and the functions they may convey. The distribution was chosen because it has good properties, and the squared exponential kernel was employed to simulate smooth functions. Considering the training data and the probability introduced in the following sections, Bayesian learning employs the Bayes rule to weight the posterior distribution of functions across the same set as the previous functions. However, the probability of these functions is different. The posterior produces functions substantially closer to the training data than functions far from it.

GPR likelihood and marginal log-likelihood

As was mentioned in the previous sections, The GP prior phase only contains X. while the likelihood that linked the random variables from the GP before the observed data Y. To determine and calculate the posterior which it extracts from the prior and the likelihood stage, then from the posterior make predictions. The probabilities are that Y can be observed under the prior random variables $f(x)$ and $f(x')$. Another set of random variables must be introduced to discuss the likelihood of observing Y. The researcher in this study suggests that the sample at position x is a random variable y (x). N Observations are found in Y; hence, there is a random vector y(X) length n.

The y(X) distribution is a multivariate Gaussian with mean f(X) and covariance. $\sigma_N^2 I_N$ is to establish the relationship between y(X) and the random f(X) variable. Where σ_N^2 Is a scalar parameter known as noise variance and I_N is a matrix of identical dimensions n, which represent as follows:

$$y|f \sim N(f, \sigma_N^2 I_N) \tag{8}$$

The Gaussian weighting form is an analogous approach to describing the same random variable y(X):

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

where $\varepsilon = (0, \sigma_N^2 I_N)$

The length scale, signal variance, and noise variance are examples of hyper-parameters often present inside the kernel and must be inferred from the data. Complete Bayesian inferences of the hyper-parameters are frequently not used since obtaining the reverse distribution of the hyper-parameters is not simple. Instead, it is standard procedure to maximize the marginal (log) likelihood to generate hyper-parameter point estimates. The critical feature of GP, which enables automatic model construction, is determining the marginal likelihood of a given model data collection, also known as evidence. The low likelihood of comparing models enables one to balance the model's capabilities and data adaption (Mackay, 1992).

Gaussian process posterior

The Bayesian technique, which learns precise values for each parameter in a single function, is a probability distribution across all possible values in contrast to many machine learning algorithms that are closely watched (Ching et al., 2021, 2023; Han et al., 2023). When discussing a parametric model, start with the parameter of interest and then combine it with the likelihood function to obtain a distribution over the parameter applied to the data:

$$p(w\backslash y, X) = \frac{p(y|X, w)p(w)}{p(y|X)} \tag{10}$$

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{marginal likelihood}} \tag{11}$$

The revised distribution$p(w\backslash y, X)$, Which is termed the posterior distribution and contains both prior distribution information and dataset information. For predictions in unknown areas of interest, x', b$f'\, y \sim N(f',\ \text{cov}(f'))$y weighing all alternative predictive distributions, the predictive distribution has been obtained by calculating the posterior distribution (Williams & Rasmussen, 2006):

$f'y \sim N(f', \text{cov}(f'))$ where.

$$f' = m' + k(x', x)[k(x,\ x') + \sigma_N^2 I_N]^{-1} \tag{12}$$

$$\text{cov}\left(f'\right) = \text{k}\left(x', x'\right) - \text{k}\left(x', x\right)[\ \text{k}(x,\ x') + \sigma_N^2 I_N]^{-1} k\left(x, x'\right) \tag{13}$$

It is typical to assume that the likelihood and prior integration are Gaussian. The predictive distribution hypothesis and resolution create a Gaussian distribution from which a point prediction may be made using the mean, and an uncertain quantification using the variance is used (Han et al., 2022). The Gaussian predictive distribution, as predicted by symmetry considerations, is determined by the weights' posterior means. The predictive variance, a quadratic equation that includes the test input and the postural covariance matrix, demonstrates that the prediction uncertainties grow with the size of the test input, as expected from a linear model (Fig 4).

Cross-validation

The cross-validation model evaluation technique assesses how well a machine learning algorithm performs while developing predictions from fresh, untrained data sets. A subset of the known data set is used to train an algorithm and other data sets. The source data must be arbitrarily divided into a training set and a test set for each cross-validation cycle. The test set can then evaluate the algorithm's performance once trained on the training dataset. With a mean cross-validation error, this method is performed numerous times and used as a performance measure. Controlling or supporting a model during training is unacceptable if it uses complicated or simplified methodologies. Choosing a training set and test set is critical to lessen this risk.

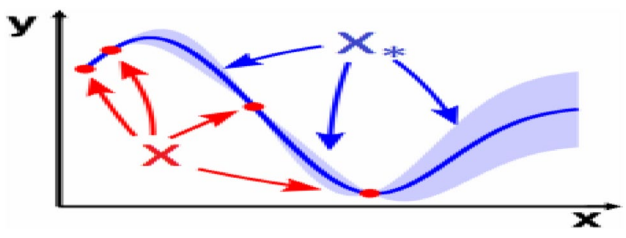

Fig. 4 An illustrative example of a Gaussian process regression (Wang, 2020)

Segmenting the dataset to improve learning and validity can be challenging. In this instance, cross-validation is used. Different methods of data division are presented via cross-validation to choose the best model algorithm. Cross-validation aids in selecting the best model by foreseeing errors on an untrained test dataset. The testing set of data can assess the model's various levels of validity and generalizability with more data (Shalev-Shwartz & Ben-David, 2014).

The data set is randomly divided into a workout set (80%) and a test set (the remaining 20%). Notably, the final 20% was previously covered. However, a fivefold cross-validation approach is used to evaluate model performance accurately and prevent randomness in test sample selection. (Berrar, 2019). Therefore, the entire dataset may be transformed into ten folds, each serving as a tester. By removing fivefold findings, the effectiveness of the suggested models can be measured (GPR & unknown GPR variance). This cross-validation technique can reliably test the GPR model and the other two benchmarking approaches because each sub-sample is mutually exclusive. The GPR model's prediction outputs and two criteria are presented in Table 2, along with a fivefold cross-validation technique. It should be noted that the GPR best predicted all performance evaluation factors. Figure 5 displays the cross-validation process and the CE loss estimation with five-fold cross-validation.

Implementation of real data

This study uses the undrained shear strength of Finnish soft clay data (D'Ignazio et al., 2016). Specific transformation equations for data input factors are used in this work as the following Table 1:

With:

$$PI = LL - PL, \tag{14}$$

$$LI = \frac{w - PL}{LL - PL}, \tag{15}$$

$$s_u(mob) = \frac{1.5}{1 + LL} s_u^{FV} \tag{16}$$

This study involves 216 FV data points from 24 different test sites in Finland. While the data size may be considered too small to capture the characteristics of geomaterial parametric data in Finland fully, efforts have been

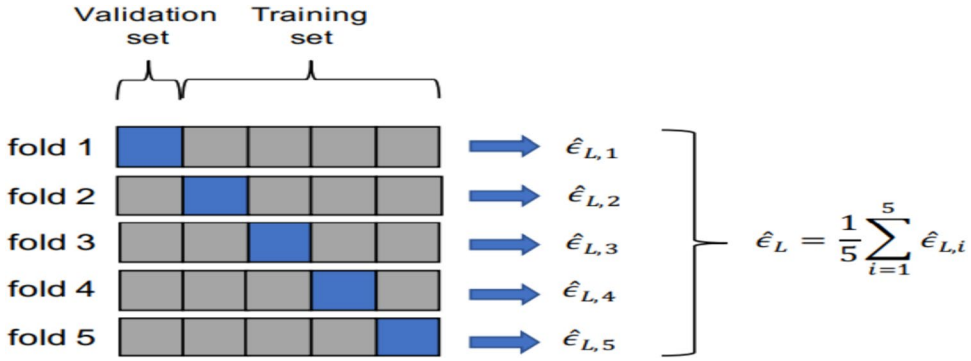

Fig.5 Cross-validation process

Table 1 Parameters Transformation Equations

Parameter	Transformation Equations
Y_1	$\ln(LL)$
Y_2	$\ln(PI)$
Y_3	LI
Y_4	$\ln\left(\sigma'_v/P_a\right)$
Y_5	$\ln\left(\sigma'_p/P_a\right)$
Y_6	$\ln\left(s_u(mob)/\sigma'_v\right)$
Y_7	$\ln\left(S_t\right)$

Table 2 Basic statistics of the study parameters in F-CLAY/7/216

Variable	n	Mean	COV	Min	Max
Y_1	216	21.443	0.501	5	75
Y_2	216	0.464	0.485	0.074	1.609
Y_3	216	0.948	0.515	0.251	2.884
Y_4	216	66.284	0.298	22.0	125.0
Y_5	216	27.740	0.204	10.0	50.0
Y_6	216	76.340	0.268	25.0	150.0
Y_7	216	17.447	0.789	2.0	64.0

made to ensure its representativeness. Each test site was carefully selected to cover various geological conditions across the country. Despite the relatively small sample size, the dataset includes detailed information on critical parameters such as undrained shear strength (*Y1*), effective vertical stress (*Y2*), pre-consolidation pressure (*Y3*, liquid limit (*Y4*), plastic limit (*Y5*), water content (*Y6*), and sensitivity (*Y7*). These parameters were obtained through rigorous field vane tests and laboratory measurements, adhering to stringent standards to ensure accuracy and reliability, as shown in Table 2. Additionally, the data collection spanned multiple years, accounting for temporal variations and providing a comprehensive overview. Detailed documentation of the locations, investigation standards, and data collection periods is available, enhancing the credibility and applicability of the dataset for civil engineering applications.

The first clay database compiled in this study consists of 216 FV data points from 24 different test sites in Finland. Each data "point" contains multivariate information, i.e., information from different tests conducted nearby is available. The collected data points contain information on seven basic parameters measured at comparable depths and sampling locations: su_u^{FV}, σ'_v, σ_p, *w, LL, PL, and* S_t. The standard FV test is usually carried out at a high rotation speed, inducing strain rates in the soil much higher than in conventional laboratory tests (Ching & Phoon, 2014). The main consequence is that su_u^{FV} Is overestimated and, therefore, a correction is needed to convert su_u^{FV} into $su_{(mob)}$ (Silvestri, 1983). Figure 6 illustrates the scatter matrices for the dataset after transformations. From this, one can decide which parameter plays a vital role and ensures high consistency between the parameters, as shown in Fig. 7.

Hyperparameters optimization

Both methods predict using a single hyperparameter set and a single noise variance value. Whereas these processes work well in practice, the Bayesian solution, which predicts the uncertainties in hyperparameters and the difference in noise as follows, is only approximated. In this model, observe that the length scale (l_c), signal variance (σ_p) and the noise variance σ_n can be varied.

$$\begin{aligned}\mathrm{Log}(f(y)) = &-\frac{1}{2}y^T(\mathrm{k}\,(x,x)+\sigma_N^2 I_N)^{-1}\mathrm{y}\\ &-\frac{1}{2}\log(\mathrm{k}\,(x,x)+\sigma_N^2 I_N)^{-1}-\frac{n}{2}\log 2\pi\end{aligned} \tag{17}$$

Overall, three parameters of the Gaussian Process model are introduced, which are length-scale (l_c), signal variance (σ_p) and the noise variance σ_n. The length scale substantially influences the appearance of the functions that the GP prior can model. The Kernel formula shows us how far away two random variables have to be in the

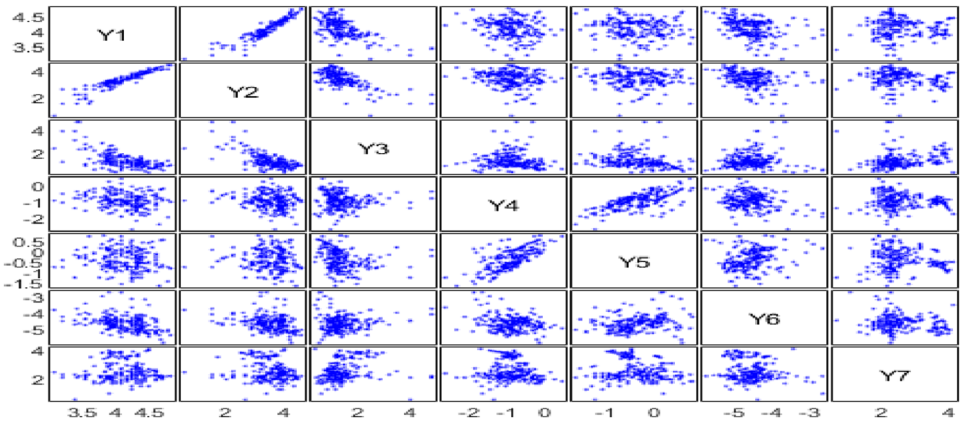

Fig.6 Scatter matrices of parameters after transformations

Parameter Name	Transformation	
Y_1	ln(LL)	Index properties
Y_2	ln(PI)	
Y_3	LI	
Y_4	$\ln(\sigma'_v/P_a)$	Stresses and strengths
Y_5	$\ln(\sigma'_p/P_a)$	
Dependent variable Y_6	$\ln(s_u/\sigma'_v)$	
Y_7	$\ln(S_t)$	

Fig.7 Model validation for clay dataset

sense of $(x, x')^2$, The inverse of the length scale governs how far away they are uncorrelated. It is the opposite of the longitude scale since l appears within the exponential function in the denominator.

The parameter in the model is σ_p, although normally σ_2^p is used in addition to considering that it would be a variance. As shown in the sample in the GP preceding section, the functions that this GP might formulate before in the y-axis are scanned with varied values. It represents the range of functions one wishes to use before using his/her GP. It also reflects that model parameter. $\sigma_n is$ to emphasize the same as the signal variance that specifies a variance. The researcher has the σ_2^p presuming that the data Y are samples with Gaussian noise generated from random variable y(X) The noise of Gaussian has the mean 0 and its variance σ_2^p.

The current model understands training data better when model parameters are specified appropriately. Changes to those parameters are made using the learning parameter until the training data are well explained. The Gaussian Process can be optimized to carry out this model parameter adjustment. The researcher then created an objective function that gauges how effectively this model accounts for the training set of data. This objective function accepts the model's parameters as arguments and outputs an actual value. More meaning can be inferred since the model accurately describes the training data. The learning parameter tries to find specific values for model parameters to optimize the objective function.

Results and discussion

The exploratory data analysis (EDA) of the dataset parameters *Y1* through *Y7* yields a full statistical summary, including the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values, as shown in Table 2. These metrics provide significant information on each parameter's distribution, average, and range. The *Y1* value (representing the logarithm of the liquid limit) has a mean of 4.147, suggesting a central tendency around this number. The standard deviation of 0.315 indicates a minimal amount of variability. The *Y1* values vary from a low of 3.091 to a high of 4.828. Half of the data falls between 3.932 and 4.373.

The *Y2* variable, which represents the logarithm of the Plasticity Index, has an average value of 3.509 and a standard deviation of 0.594, indicating moderate variability. The range of values varies from a low of 0.693 to a high of 4.554. The interquartile range (IQR) is between 3.219 and 3.912. The *Y3* (Liquidity Index) has an average value of 1.443 and a more significant standard deviation of 0.663, suggesting that it has more significant variability than *Y1* and *Y2*. The results exhibit a broad range, spanning from 0.425 to 4.800. The interquartile range, representing the middle 50% of the data, is between 1.060 and 1.629.

The *Y4* (Normalized Effective Vertical Stress) has a mean value of − 0.876 and a standard deviation of 0.471, indicating moderate variability. The data set has a minimum value of − 2.603 and a maximum value of 0.476. Approximately half of the data falls between the range of − 1.156 and − 0.603. The *Y5* (Normalized Preconsolidation Pressure) has a mean of − 0.352 and a standard deviation of 0.482, suggesting moderate variability. The dataset spans from − 1.622 to 0.820, with the interquartile

range ranging from − 0.706 to − 0.013. The average value for *Y6* (Normalized Undrained Shear Strength) is − 4.556, with a standard deviation of 0.479, indicating a minimal variation. The values span from − 5.763 to − 2.620, while the interquartile range of the data is between − 4.885 and − 4.335.

The *Y7* variable, which represents the logarithm of sensitivity, has a mean value of 2.615 and a standard deviation of 0.668. These values indicate that there is considerable variability in the data. The range of values in the dataset spans from 0.693 to 4.159, while the interquartile range, which represents the middle 50% of the data, ranges from 2.197 to 2.996. The exploratory data analysis (EDA) shows that the dataset characteristics display different central tendencies and dispersion levels. *Y3* and *Y4* exhibit more fluctuation, while *Y1* and *Y6* have lesser variability. Comprehending these attributes is essential for further examination and constructing models, guaranteeing that the Gaussian Process Regression (GPR) model precisely represents the fundamental data patterns.

The heatmap above displays the correlation matrix for the dataset parameters Y1 through Y7 as shown in Fig. 8. Correlation coefficients range from − 1 to 1, indicating the strength and direction of the linear relationship between pairs of variables. A coefficient close to 1 suggests a strong positive correlation, while a coefficient close to − 1 indicates a strong negative correlation. A value around 0 implies no linear correlation. There is a robust positive correlation between *Y1* (Logarithm of Liquid Limit) and *Y2 (*Logarithm of Plasticity Index), suggesting that as the Liquid Limit increases, the Plasticity Index also tends to increase significantly. Similarly, the correlation between *Y4* (Normalized Effective Vertical Stress) and *Y5* (Normalized Preconsolidation Pressure) indicates a substantial positive relationship between these parameters.

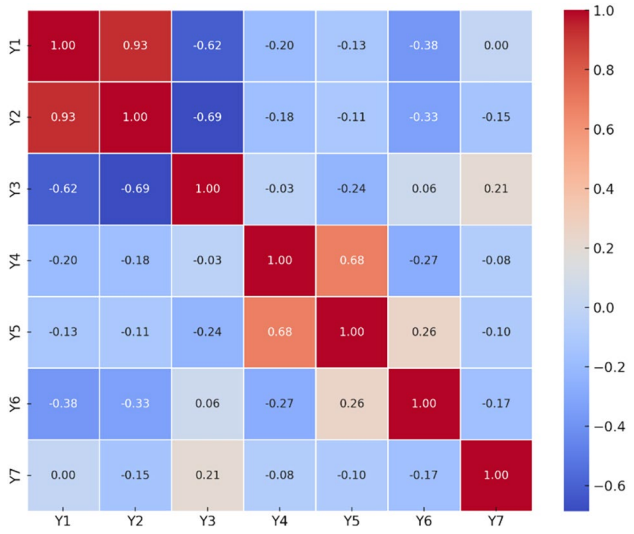

Fig.8 Heatmap correlation of parameters

On the other hand, there is a strong negative correlation between *Y1* and *Y3* (Liquidity Index), suggesting that as the Liquid Limit increases, the Liquidity Index tends to decrease. The same inverse relationship is observed between *Y2* and *Y3*. Additionally, *Y1* and *Y6* (Normalized Undrained Shear Strength) show a moderate negative correlation, indicating that an increase in Liquid Limit is associated with a decrease in Undrained Shear Strength. A similar moderate negative correlation is observed between *Y2* and *Y6*. *Y7* (Logarithm of Sensitivity) shows weak correlations with other parameters, suggesting minimal linear relationships between Sensitivity and the other variables in the dataset. *Y3* has weak correlations with *Y4*, *Y5*, and *Y6*, indicating limited linear relationships with these parameters. The correlations between *Y1* and *Y7*, and between *Y5* and *Y7*, are close to zero, indicating no significant linear relationships between these pairs of parameters.

The heatmap provides a visual representation of the relationships between the different parameters in the dataset. Strong positive and negative correlations highlight potential dependencies and interactions between variables, which are crucial for model development and analysis. Understanding these relationships can help select appropriate features for modeling and interpret the results of the Gaussian Process Regression (GPR) model.

This section shows some results of the initial GPR (Model 1) and an unknown GPR variance (model 2). Because the aim is to get the optimum possible performance on a data set, the model quality is assessed with a cross-entropy (EC) loss and MSE (Table 3).

Artificial data 1D

Figure 9 compares Gaussian processes with varying variances. Large, medium, or minor deviation. _f = (4,5,1) Medium predictive distribution, 90% quantiles (solid blue) (black). The gray area demonstrates that the smaller Sigma prior provides more corner confidence (a narrow collection of angles). Confidence intervals are calculated for a population parameter using sample data. The truth is otherwise. Not everyone can always be checked out. The same facts will be interpreted differently by different samples. Level of assurance Expecting 90 out of 100 random samples to represent the population means a 90% trust level.

In a regression model, the value of the dependent variable is estimated. The prediction interval is the range of model performance for a single new observation with specified variable values. Predictions and confidence intervals are commonly confused. Two interconnected processes have different calculations and goals. The dependability interval

represents the projected range of values associated with statistical data attributes such as the population average. In contrast, the prediction interval predicts how far the following individual observation will fall.

Figure 8 shows the influence of model variance (sigma prior _p). Only the y-axis functions are scaled in the signal variance parameter _p. They look like the following figure, with a more fantastic y-axis range. Increasing grey space

Table 3 Exploratory data analysis (EDA) of Dataset parameters

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Count	216	216	216	216	216	216	216
Mean	4.147045	3.509101	1.442851	−0.87608	−0.35195	−4.55645	2.615468
Std	0.315334	0.593533	0.662575	0.471345	0.482405	0.47895	0.667942
Min	3.091042	0.693147	0.424625	−2.60318	−1.62235	−5.76325	0.693147
25%	3.931826	3.218876	1.060033	−1.15553	−0.70606	−4.88465	2.197225
50%	4.174387	3.570147	1.291005	−0.86588	−0.36959	−4.5746	2.397895
75%	4.372593	3.912023	1.629212	−0.60292	−0.01292	−4.33528	2.995732
Max	4.828314	4.553877	4.8	0.475664	0.819993	−2.62022	4.158883

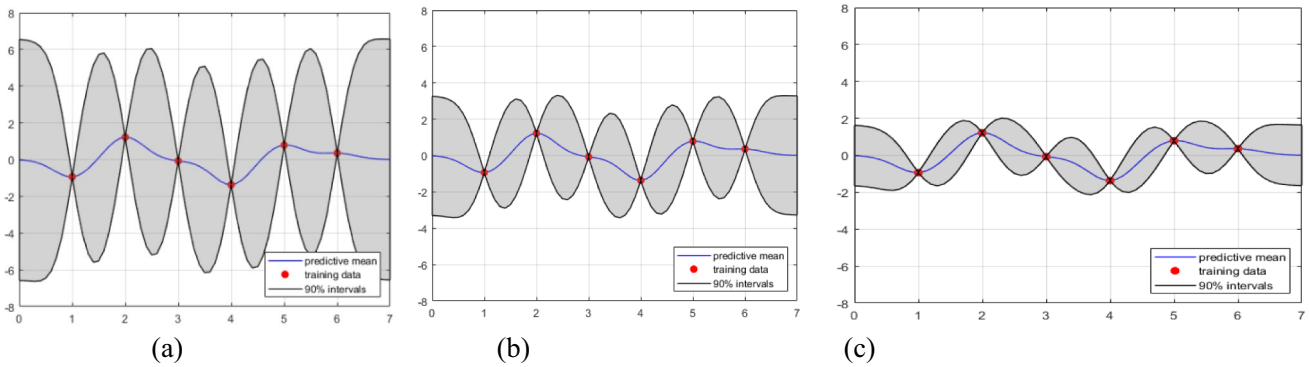

Fig.9 Comparison of Gaussian processes with different values of variance in the 1D model: **a** significant variance, **b** Medium variance, **c** slight variance. σ_f = (4, 2, 1) Respectively

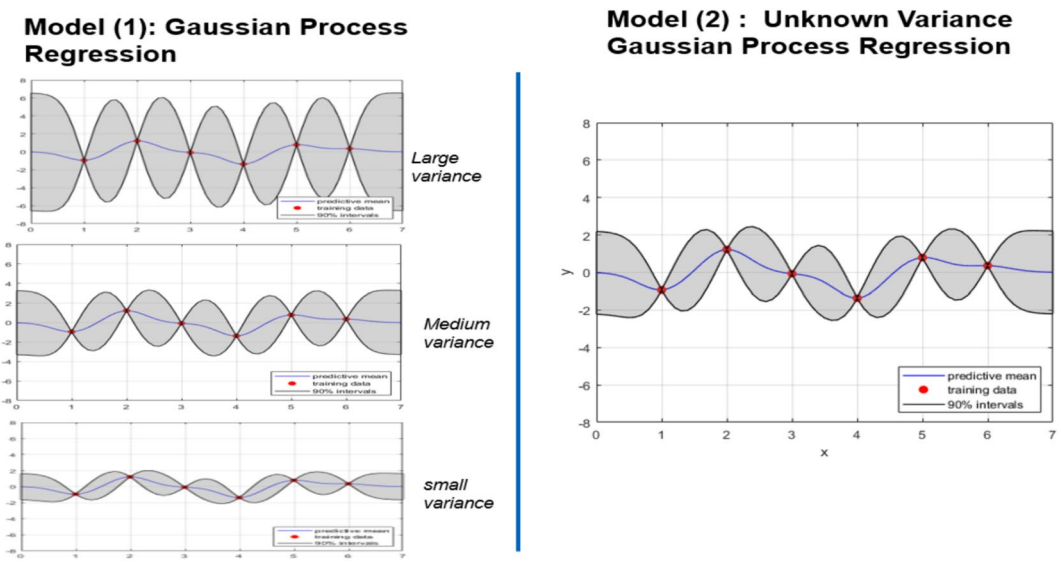

Fig.10 Comparison of Gaussian processes with different values of variance in 1D model between model 1 and model 2

(confidence interval) increases error. The best model has a little gray area around it.

Turning now to 1D artificial data of unknown variance, GPR is presented in Fig. 10, Which illustrates that when the researcher assumes that there is no variance (unknown variance), it gets a better result and gives us the optimum. This can be noted from the area of grey (confidence interval), which indicates that model 2 (unknown variance). Figure 10 shows how the data interpolates the predictive surface. Error bars are in a "football" form, or some say a "sausage" form, which is most widely available from x-values in the dataset. Error bars are mainly outside the data spectrum, a typical feature of standard linear regression. However, the prediction means it is much different from a conventional linear model. It is a reversal for GPs. Predictive variances are also reversed to something, as illustrated by these error bars: a previous variance of 1. For instance, the variance will not increase as x goes beyond X_n. Though one cannot trust extrapolations outside of the range of data, at least their conduct is not surprising in conjunction with these two "reversions" because they may occur in a linear regression context, for example, when based on enlarged (e.g., polynomial) covariates of features.

Because the majority of data is used for fitting, the distance is dramatically reduced, as is the variance, because the majority of data is used for validation. The exchange of training and testing sets improves the method's effectiveness as well. A loss function is defined as how well a prediction model can anticipate the outcome. The most common method for determining the lowest function point is "gradient descent."

Loss functions are equations that provide a curve of loss caused by model predictions. Loss functions are also characterized as objective functions. The aim is to limit the loss function to increase the model's accuracy for better forecasts. This paper used two types of loss functions as mentioned before which are cross entropy (CE) and mean square error (MSE). As shown in Figs. 10 & 11 which indicates that according to the observed measurements, the GPR model has achieved the highest accuracy with relatively low predictions (CE = 0.1, MSE = 0.11 percent in model 1) and (CE = 0.41, MSE = 0.46 percent in model 2) as shown in Fig. 11 & 12.

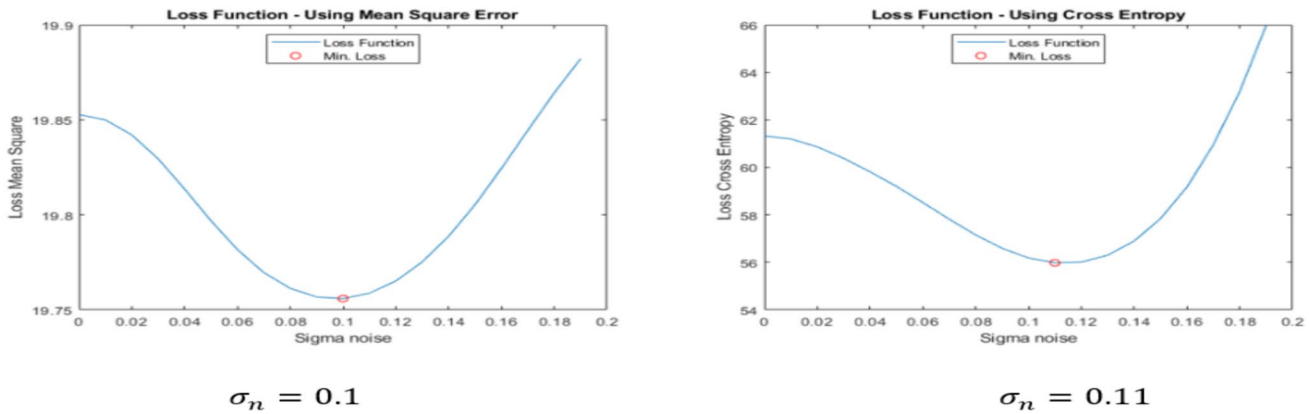

Fig.11 Comparison of loss functions in whole GPR model

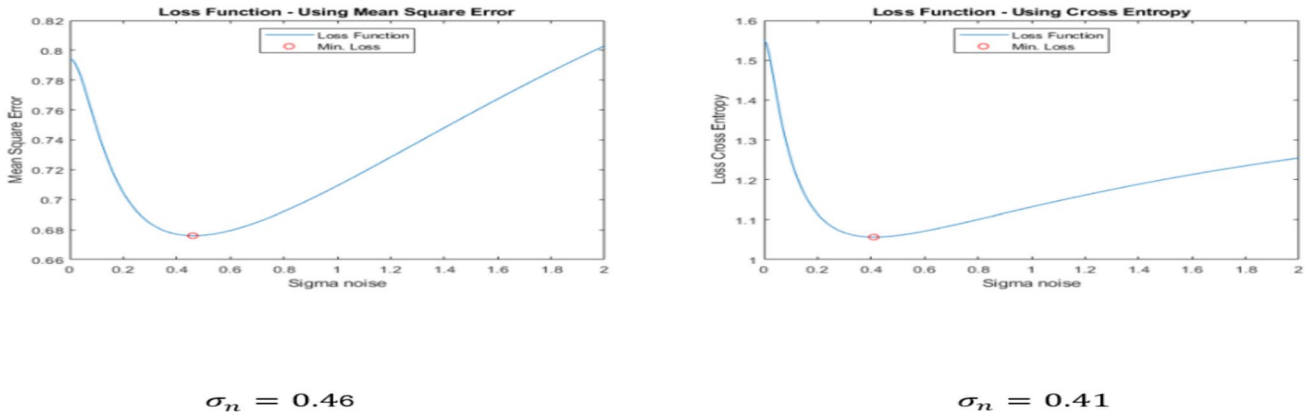

Fig.12 Comparison of loss functions in whole unknown variance GPR model

Conclusion and recommendations

This work highlights the application of Gaussian process regression (GPR) in civil engineering, specifically in fields like structural engineering, transportation engineering, and geotechnical engineering. The study demonstrates the power of artificial intelligence and machine learning in predicting material qualities by proposing a hierarchical approach to modeling clay with spatially scattered data using GPR. Utilizing a Gaussian regression process with a normally multivariate and normal-gamma distribution, the GPR model achieves high accuracy with low prediction errors (CE = 0.1, MSE = 0.11% in model 1, and CE = 0.41, MSE = 0.46% in model 2). Key conclusions include the necessity to account for data variance uncertainty and the effectiveness of GPR in engineering datasets. GPR's ability to provide accurate estimates and prediction intervals for material properties is crucial for construction engineers. Future research should explore GPR applications across various civil engineering domains, analyze novel covariance functions, and investigate other machine learning approaches for better model interpretation. Despite its advantages, the unknown variance model may present challenges for engineers, emphasizing the need for a larger dataset for broader applications. The GPR models effectively idealize complex data systems, with predictions reflecting uncertainty that can be transferred to further estimations.

Acknowledgements I want to convey our sincere gratitude to everyone who helped and contributed to completing this research paper. They have been accommodating and supportive in making this study possible.

Author contribution The author wrote and reviewed the manuscript.

Data availability Additional data and materials are available upon request from the corresponding author.

Declarations

Conflict of interests The authors declare no competing interests.

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*, 52.

Ahmad, M., Keawsawasvong, S., Bin Ibrahim, M. R., Waseem, M., Kashyzadeh, K. R., & Sabri, M. M. S. (2022). Novel approach to predicting soil permeability coefficient using Gaussian process regression. *Sustainability, 14*(14), 8781.

Berrar, D. (2019). *Cross-Validation*. Elsevier.

Cai, H., Jia, X., Feng, J., Li, W., Hsu, Y. M., & Lee, J. (2020). Gaussian process regression for numerical wind speed prediction enhancement. *Renewable Energy, 146*, 2112–2123.

Calandra, R., Peters, J., Rasmussen, C. E., & Deisenroth, M. P. (2016, July). Manifold Gaussian processes for regression. In 2016 International Joint Conference on Neural Networks (IJCNN) (pp. 3338–3345). IEEE..

Cervone, Daniel & Pillai, Natesh. (2015). Gaussian Process Regression with Location Errors.

Cheng, M. Y., Huang, C. C., & Roy, A. F. V. (2013). Predicting project success in construction using an evolutionary Gaussian process inference model. *Journal of Civil Engineering and Management, 19*(sup1), S202–S211.

Ching, J., & Phoon, K. K. (2014). Correlations among some clay parameters—the multivariate distribution. *Canadian Geotechnical Journal, 51*(6), 686–704.

Ching, J., Wu, S., & Phoon, K. K. (2021). Constructing quasi-site-specific multivariate probability distribution using hierarchical Bayesian model. *Journal of Engineering Mechanics, 147*(10), 04021069.

Ching, J., Yoshida, I., & Phoon, K. K. (2023). Comparison of trend models for geotechnical spatial variability: Sparse Bayesian learning vs. *Gaussian Process Regression. Gondwana Research, 123*, 174–183.

D'Ignazio, M., Phoon, K. K., Tan, S. A., & Länsivaara, T. T. (2016). Correlations for undrained shear strength of Finnish soft clays. *Canadian Geotechnical Journal, 53*(10), 1628–1645.

Djeziri, M., & Bendahan, M. (2021). Special Issue "Advances in Machine Learning and Deep Learning Based Machine Fault Diagnosis and Prognosis". Processes, 9(3), 532.

Duvenaud, D. (2014). Automatic model construction with Gaussian processes (Doctoral dissertation, University of Cambridge).

Ebden, M. (2015). Gaussian processes: A quick introduction. arXiv preprint arXiv:1505.02965.

Han, L., Liu, H., Zhang, W., & Wang, L. (2023). A comprehensive comparison of copula models and multivariate normal distribution for geo-material parametric data. *Computers and Geotechnics, 164*, 105777.

Han, L., Wang, L., Zhang, W., & Chen, Z. (2022). Quantification of statistical uncertainties of unconfined compressive strength of rock using Bayesian learning method. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 16(1), 37–52.

Hoang, N. D., Pham, A. D., Nguyen, Q. L., & Pham, Q. N. (2016a). Estimating compressive strength of high performance concrete with Gaussian process regression model. *Advances in Civil Engineering*. https://doi.org/10.1155/2016/2861380

Hoang, N. D., Pham, A. D., Nguyen, Q. L., & Pham, Q. N. (2016b). Estimating compressive strength of high performance concrete with Gaussian process regression model. *Advances in Civil Engineering, 2016*(1), 2861380.

Hu, J., & Wang, J. (2015). Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression. *Energy, 93*, 1456–1466.

Jordan, M. I. and TM Mitchell (2015)2*. ML: Trends, perspectives and prospects.

Karch, J. D., Brandmaier, A. M., & Voelkle, M. C. (2020). Gaussian process panel modeling—machine learning inspired analysis of longitudinal panel data. *Frontiers in Psychology, 11*, 351.

Kaveh, A. (2021). *Advances in metaheuristic algorithms for optimal design of structures*. Switzerland: Springer International Publishing.

Kaveh, A. (2023). *Topological transformations for efficient structural analysis*. Springer.

Kaveh, A. (2024). *Applications of artificial neural networks and machine learning in civil engineering, studies in computational intelligence 1168*. Springer.

Kaveh, A., & Bakhshpoori, T. (2019). *Metaheuristics: outlines*. MATLAB codes and examples: Springer International Publishing, Cham.

Kaveh, A., & Eslamlou, A. D. (2020). *Metaheuristic optimization algorithms in civil engineering: New applications*. Springer International Publishing.

Lu, X., Li, H. X., Duan, J. A., & Sun, D. (2010). Integrated design and control under uncertainty: a fuzzy modeling approach. *Industrial & Engineering Chemistry Research, 49*(3), 1312–1324.

Mackay, D. J. C. (1992). Bayesian methods for adaptive models. California Institute of Technology.

Mahmoodzadeh, A., Mohammadi, M., Ibrahim, H. H., Rashid, T. A., Aldalwie, A. H. M., Ali, H. F. H., & Daraei, A. (2021). Tunnel geomechanical parameters prediction using Gaussian process regression. *Machine Learning with Applications, 3*, 100020.

Momeni, E., Dowlatshahi, M. B., Omidinasab, F., Maizir, H., & Armaghani, D. J. (2020). Gaussian process regression technique to estimate the pile bearing capacity. *Arabian Journal for Science and Engineering, 45*, 8255–8267.

Monisha, R., & Balasubramanian, M. (2023). Energy simulation through design builder and temperature forecasting using multilayer perceptron and Gaussian regression algorithm. *Asian Journal of Civil Engineering, 24*(7), 2089–2101.

Pal, M., & Deswal, S. (2010). Modelling pile capacity using Gaussian process regression. *Computers and Geotechnics, 37*(7–8), 942–947.

Quadrianto, N., Kersting, K., & Xu, Z. (2010). *Gaussian Process*. US: Springer.

Rasmussen, C. E. (2004). *Gaussian processes in machine learning*. Springer, Berlin, Heidelberg: In Summer school on machine learning.

Rasmussen, C. E., & Nickisch, H. (2010). Gaussian processes for machine learning (GPML) toolbox. *The Journal of Machine Learning Research, 11*, 3011–3015.

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.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Emgland: Cambridge University Press.

Silvestri, V. (1983). The bearing capacity of dykes and fills founded on soft soils of limited thickness. *Canadian Geotechnical Journal, 20*(3), 428–436.

Stahl, S. (2006). The evolution of the normal distribution. *Mathematics Magazine, 79*(2), 96–113.

Tong, Y. L. (2012). *The multivariate normal distribution*. Germany: Springer Science & Business Media.

Verma, M. (2023). Prediction of compressive strength of geopolymer concrete by using ANN and GPR. *Asian Journal of Civil Engineering, 24*(8), 2815–2823.

Wang, J. (2020). An intuitive tutorial to Gaussian processes regression. arXiv preprint arXiv:2009.10862.

Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. Cambridge, MA.

Zhang, D., Zhou, Y., Phoon, K. K., & Huang, H. (2020). Multivariate probability distribution of Shanghai clay properties. *Engineering Geology, 273*, 105675.

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers, 12*(1), 469–477.

Zhang, Y., & Xu, X. (2021). Predicting multiple properties of pervious concrete through the Gaussian process regression. *Advances in Civil Engineering Materials, 10*(1), 56–73.

Zhou, L., Chen, J., & Song, Z. (2015). Recursive Gaussian process regression model for adaptive quality monitoring in batch processes. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2015/761280

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com