Probabilistic Modeling of Heteroscedastic Laboratory Experiments Using Gaussian Process Regression

Lucie Tabor[1], James-A. Goulet[1], Jean-Philippe Charron[1], and Clelia Desmettre[1]

[1]Department of Civil, Geologic and Mining Engineering Polytechnique Montreal, CANADA.
Email: tabor.lucie@gmail.com

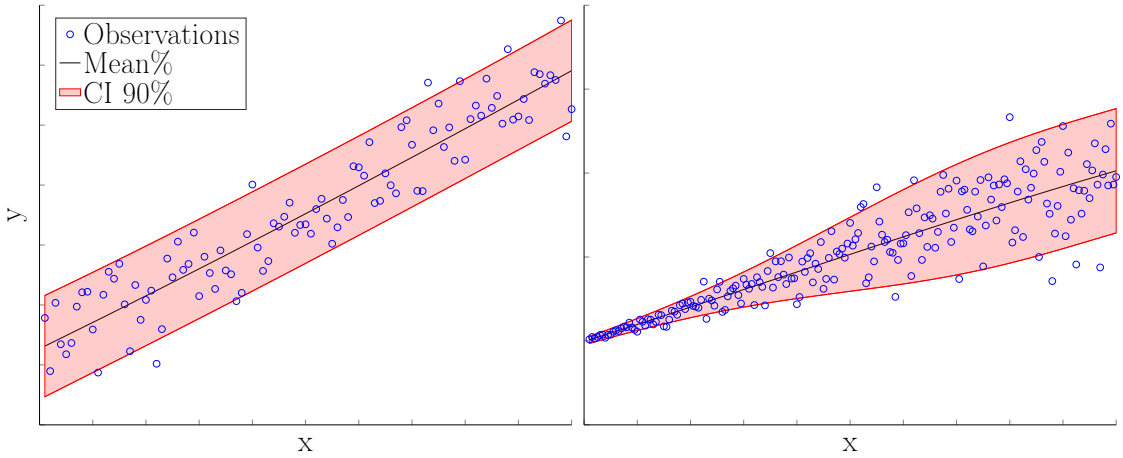
Abstract

This paper proposes an extension to Gaussian Process Regression (GPR) for datasets composed of only few replicated specimens and displaying a heteroscedastic behaviour. As there are several factors that are out of the control of experimenters, it is often impossible to reproduce identical specimens for a same experiment. Moreover, observations from laboratory experiments typically display a heteroscedastic inter-specimens variability. Because experiments and specimens manufacturing is expensive, it is uncommon to have more than three specimens to build a model for the observed responses. The method proposed in this paper uses GPR to predict each tested specimen using a shared prior structure and models the global heteroscedastic behaviour by combining observations using conjugate prior distributions. An application of the method to high performance fiber reinforced concrete experiments highlights fiber addition benefits for reducing water permeability caused by macro-cracks.

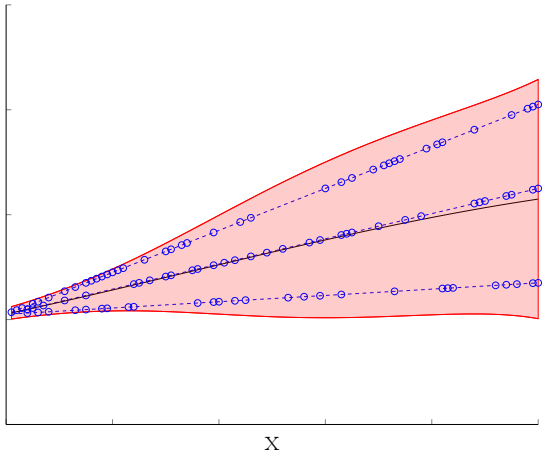
INTRODUCTION

Modeling the variability in the results of laboratory experiments is difficult when only few specimens are available. In civil engineering, this situation is common practice because preparing and testing specimens often incurs high costs. Experimentalists are left with the difficult task of quantifying the inter-specimen variability from a sparse dataset. When tests are performed as a function of covariates, an additional challenge is that experimental results typically display a heteroscedastic behaviour, so that test results variability depends on covariate values. Figure 1a & 1b respectively show an example of a homoscedastic and of a heteroscedastic behaviour. For Figure 1a, the observations variability is independent of the covariate x, which is not the case for Figure 1b. Moreover, both figures represent

the special case where all observations are obtained from different specimens that are independent of each others. This paper focuses on the case where observations obtained for different covariate values, are only available from a limited number of replicated specimens. This case is displayed in Figure 1c.



(a) Homoscedasticity and independent observations (b) Heteroscedasticity and independent observations



(c) Heteroscedasticity and dependent observations

Figure 1: Dataset examples displaying (a) a homoscedastic behaviour, (b) a heteroscedastic behaviour and (c) a heteroscedastic behaviour with dependent observations. In (c) the dashed line links observations obtained from a same specimen.

Many methods to construct a model from laboratory experiments already exist, methods

such as Multiple Linear Regression (MLR), Multivariate Adaptive Regression Spline (MARS) and Symbolic Regression (SR) are able to model complex datasets (Jeon et al. 2014), but are prone to over-fitting as they select the best equation formulation to fit the data. Methods like Support Vector Machine (SVM) or Neural Network (NN) are able to model highly non-linear datasets (Siddique et al. 2008; Pal and Deswal 2008). Słoński (2010) succeeded in modeling compressive strength in high-performance concrete with NN method. Many methods, e.g. Lampinen and Vehtari (2001), Ma et al. (2014) and Zhong et al. (2008) rely on a Bayesian approach to infer the model parameters which can become useful in order to capture uncertainty in physics formula. Based on such a method, Gardoni et al. (2002, 2007) modeled both capacity and fragility in reinforced concrete columns or elastic modulus of concrete. This last method can be applied to heteroscedastic behaviours by transforming data in a homoscedastic space or by adding input dependent-noise to the model (Bansal and Aggarwal 2007; Blau et al. 2008). Also, Yeh (2014) estimated the distribution of compressive strength of high-performance concrete which displayed a heteroscedastic behaviour, through the NN method.

One specific Machine Learning method suited to address these probabilistic models challenges is Gaussian Process regression (GPR) (Rasmussen and Williams 2006; MacKay 1998). Regarding over-fitting, Gaussian Process Regression is known to be more robust to over-fitting than other approaches such as linear regression because it can model complex function using few parameters. For a specific discussion, the reader is invited to consult Rasmussen and Williams (2006). Słoński (2011) recommended this approach instead of NN for the identification of concrete properties. One of the strengths of GPR is that it allows to interpolate and extrapolate experimental values by providing mean values as well as the covariance matrix for its predictions. The accuracy of GPR predictions depends on the distance between predicted and observed covariates. In the case of an extrapolation far from available observations, the prediction uncertainty reflects the absence of empirical observation to rely on. GPR can handle highly nonlinear sets of data. Simple regression problems such as the one presented in Figure 1a can readily be processed with the GPR method which is already implemented in open-source codes such as GPML (Rasmussen and Nickisch 2010). For regression problems involving heteroscedasticity, Goldberg et al. (1997) proposed to employ a hierarchical approach so that the GPR variance is itself modelled by a Gaussian process (Kersting et al. 2007). This model was later extended by Tolvanen et al. (2014) to embody the heteroscedasticity in both process and observation noises. These methods are implemented in the open-source code GPstuff (Vanhatalo et al. 2012). Wand and Neal (2012, 2014) have proposed an alternative approach to model heteroscedasticity in the context of GPR by introducing latent covariates. Although the formulation of both approaches is different, they share the same hypothesis that all observations are independent of each others as illustrated in Figure 1b. Methods based on GPR are efficient to model homoscedasticity ; for example, based on a dataset of only five specimens, Thiyagarajan and Kodagoda (2016) modeled concrete moisture. They are also reliable for modeling heteroscedastic behaviours associated with large datasets obtained

from independent specimens as demonstrated by Kersting et al. (2007), Le et al. (2005) and Titsias and Lázaro-Gredilla (2011).

The shaded region in Figures 1a-c describes the 90% confidence interval computed using in a) the standard homoscedastic GPR approach, and b,c) Tolvanen et al. (2014) heteroscedastic GPR. The confidence interval in Figures 1a & 1b are consistent with the expected result, because the model hypothesis regarding independent observations is satisfied. For Figure 1c, the 90% confidence interval is too narrow to describe the variability across specimens because the underlying hypothesis in Tolvanen et al. (2014) method is that all observation is obtained from a different specimen. In Figure 1b, the effective number of observations is 200, whereas in Figure 1c it is 3. In Figure 1c, if a fourth specimen were to be tested, there is a high probability that the new observations would fall out of the confidence interval. The poor performance displayed in Figure 1c is attributed to the inadequacy of the observation independence hypothesis which overestimates the information in the data, leading to an over-narrow confidence interval. Existing methods are currently not able to model the response from statistically dependent observations caused by same unmeasured variables and displaying a heteroscedastic behaviour.

This paper proposes a new extension to Gaussian Process Regression for creating probabilistic models from few replicated specimens displaying a heteroscedastic behaviour. This situation is common when analyzing laboratory experiments in the context of civil engineering. The key aspect of this paper is to extend the GPR model to heteroscedasticity associated to epistemic uncertainty characterized by a low number of test specimens. In the method presented in this manuscript the heteroscedastic behavior is treated using the conjugate prior distribution. First, the standard GPR method is presented, then its extension to overcome the limitations presented above. Finally, the potential of the method is illustrated using experimental data of the water permeability test conducted on high performance fiber-reinforced concrete tie-specimens (Hubert et al. 2015). The challenge associated with this illustrative example is to provide a probabilistic model for the permeability in order to quantify the effectiveness of different fiber reinforcement ratios.

GAUSSIAN PROCESS REGRESSION

Model definition

Given a dataset $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, including N observations $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ assumed conditionally independent for N values of the covariate $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$, Gaussian Process Regression can predict an unobserved value, given a new covariate value x_*. The vector $\mathbf{x}$ can be extended to a matrix of several covariates. Because GPR quantifies the uncertainty for each of its predictions, it is suited for interpolation and extrapolation. In GPR, observations $\mathbf{y}$ are function of the covariates $\mathbf{x}$ and are described by a multivariate Gaussian distribution $\mathbf{y}|\mathbf{x} : Y \sim \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$. This last distribution is characterized by a

mean column vector $\mathbf{M}$ and a covariance matrix $\mathbf{\Sigma}$ which includes observations errors $\mathbf{v}$,

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{f(\mathbf{x})}_{\text{reality}} + \underbrace{\mathbf{v}}_{\text{observations errors}}, \qquad \text{with} \quad \mathbf{v} : V \sim \mathcal{N}(0, \sigma_v^2) \tag{1}$$

These matrices describing the prior structure or the distribution before any observation need to be chosen carefully in order to suit the studied behaviour. In a first step, a simple prior structure can be set up and afterwards refined if necessary. In practice, it is common to employ $\mathbf{M} = \mathbf{0}$ and a square exponential covariance function,

$$g(x_k, x_l) = \sigma_f^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell^2}\right] + \sigma_v^2 \delta(x_k, x_l)$$

where σ_f^2 is the process noise variance and ℓ is the correlation length which defines the influence of one covariate x_k on another covariate x_l. The longer is the correlation length, the higher will be the correlation between a covariate and another for a same distance $|x_k - x_l|$. The estimation of hyper-parameters σ_f, ℓ and σ_v is described in the next subsection. The observations noise variance σ_v^2 only appears on the covariance matrix diagonal where $x_k = x_l$, as observations error are assumed to be independent from one to another. Also, the covariance matrix $\mathbf{\Sigma}$ needs to be positive semi-definite,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_f^2 + \sigma_v^2 & g(x_1, x_2) & \cdots & g(x_1, x_N) \\ & \sigma_f^2 + \sigma_v^2 & \cdots & g(x_2, x_N) \\ & & \ddots & \cdots \\ \text{Sym.} & & & \sigma_f^2 + \sigma_v^2 \end{bmatrix}$$

GPR can estimate unobserved values $\mathbf{f} = [f(x_{1*}), f(x_{2*}), \ldots, f(x_{P*})]^\intercal$ given the target covariate values $\mathbf{x}_* = [x_{1*}, x_{2*}, ..., x_{P*}]^\intercal$. GPR computes the covariance for all observed and unobserved values and stores it in a new covariance matrix. The mean column vector $\mathbf{M}$ is completed by the prior mean of the predictions $\mathbf{M}_*$ of the target values $\mathbf{x}_*$. And the multivariate Gaussian distribution becomes,

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{M}\\ \mathbf{M}_*\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma} & \mathbf{\Sigma}_*{}^\intercal\\ \mathbf{\Sigma}_* & \mathbf{\Sigma}_{**}\end{bmatrix}\right), \quad \text{with}$$

$$\mathbf{\Sigma}_* = \begin{bmatrix} g(x_{1*},x_1) & g(x_{1*},x_2) & \cdots & g(x_{1*},x_N)\\ & g(x_{2*},x_2) & \cdots & g(x_{2*},x_N)\\ & & \ddots & \cdots\\ \text{Sym.} & & & g(x_{P*},x_N)\end{bmatrix}$$

$$\mathbf{\Sigma}_{**} = \begin{bmatrix} \sigma_f^2 & g(x_{1*},x_{2*}) & \cdots & g(x_{1*},x_{P*})\\ & \sigma_f^2 & \cdots & g(x_{2*},x_{P*})\\ & & \ddots & \cdots\\ \text{Sym.} & & & \sigma_f^2\end{bmatrix}$$

The estimated values $\mathbf{f}$ dependent on the observations $\mathbf{y}$ are described by a multivariate Gaussian distribution,

$$\begin{aligned}\mathbb{E}[\mathbf{f}] &= \mathbf{M}_* + \mathbf{\Sigma}_*\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{y}-\mathbf{M}),\\ \text{cov}(\mathbf{f}) &= \mathbf{\Sigma}_{**} - \mathbf{\Sigma}_*\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{\Sigma}_*{}^\intercal,\\ \mathbf{f}|\mathbf{y} &\sim \mathcal{N}(\mathbb{E}[\mathbf{f}], \text{cov}(\mathbf{f}))\end{aligned}$$

Hyper-parameter estimation

With GPR, the parameters of the prior distribution i.e. the hyper-parameters, are estimated using the dataset $\mathcal{D}$. For the square exponential covariance function, the set of hyper-parameters is $\mathcal{P}_f = \{\sigma_f, \ell, \sigma_n\}$. According to Bayes theorem, the posterior probability for hyper-parameters values is given by,

$$\begin{aligned}p(\mathcal{P}_f|\mathcal{D}) &= \frac{p(\mathcal{D}|\mathcal{P}_f)\cdot p(\mathcal{P}_f)}{p(\mathcal{D})}\\ &\propto p(\mathcal{D}|\mathcal{P}_f)\cdot p(\mathcal{P}_f)\end{aligned}$$

With the hypothesis that prior $p(\mathcal{P}_f)$ is constant, the maximum values for the posterior and the likelihood are reached for the same optimal value $\mathcal{P}_f^*$. In case where large datasets are available, the variance $\text{var}[\mathcal{P}_f|\mathcal{D}] \to 0$ so that it becomes a reasonable assumption to

employ the MLE approximation $\mathcal{P}_f^*$ rather than the full posterior $p(\mathcal{P}_f|\mathcal{D})$. A practical issue is that as the number of terms in the covariance matrix increases, the likelihood is affected by zero underflow. Using the log-likelihood instead solves this issue and the MLE of the hyper-parameters $\mathcal{P}_f$ becomes $\mathcal{P}_f^* = \arg\max_{\mathcal{P}_f}\{\log(p(\mathcal{D}|\mathcal{P}_f))\}$.

GPR FOR SPARSE AND HETEROSCEDASTIC DATASETS

To start with, the attention is restricted to the simplified context of a single covariate value x_i associated with a set of N_s observations $\mathcal{D}_i = \{(x_i, s_j, y_j), j = 1, \ldots, N_s\}$, where each observed response $y_j \in \mathbb{R}$ is obtained from a different specimen $s_j \in \mathcal{S} = \{1, \cdots, N_s\}$. For different specimens, observations are realizations of the process $y_j : Y = T + V$, where the random variable $T \sim p(t; \mathcal{P}_t)$ describes the inter-specimens variability and V is the observations errors. In this case the realization t of the inter-specimens variability T is defined by the variable its parameters $\mathcal{P}_t$. In this context, the aim is the characterization of the posterior predictive probability density function (pdf) when posterior hyper-parameters uncertainty is marginalized,

$$\tilde{T} \sim p(t|\mathcal{D}_i) = \int p(t; \mathcal{P}_t) \cdot p(\mathcal{P}_t|\mathcal{D}_i) d\mathcal{P}_t.$$

For general cases, inferring the posterior pdf for hyper-parameters $p(\mathcal{P}_t|\mathcal{D}_i)$ using Bayes theorem,

$$p(\mathcal{P}_t|\mathcal{D}_i) = \frac{p(\mathcal{D}_i|\mathcal{P}_t) \cdot p(\mathcal{P}_t)}{p(\mathcal{D}_i)}$$

and marginalizing its effect in the posterior predictive are known to be challenging tasks (Murphy 2007; Gelman et al. 2014). If specific conjugate distributions are employed to describe the prior knowlege $p(\mathcal{P}_t)$ and the likelihood of observations $p(\mathcal{D}_i|\mathcal{P}_t)$, both the posterior $p(\mathcal{P}_t|\mathcal{D}_i)$ and the posterior predictive $p(t|\mathcal{D}_i)$ can be exactly calculated with little efforts using analytic formulations (Gelman et al. 2014).

The challenge is that in common experimental setups, such as the example presented in Figure 2, the number of observations y_j available for any given covariate x_k or x_l is most often equal to either zero or one. In such a context, it is not possible to take advantage of the analytic formulations allowed by conjugate priors. This section presents how a combination of GPR and conjugate priors overcomes this limitations.

Combining GPR and conjugate priors

The first aspect of the method proposed consists in employing the GPR method to build a joint model for each of the N_s specimens. The joint model requires increasing the covariate set to include both experiments input $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$ with $x_i \in \mathbb{R}$ and specimens numbers $\mathbf{s} = [s_1, s_2, ..., s_N]^\intercal$ with $s_j \in \mathcal{S}$. Because such a model enables predicting the response for each specimen s_j for any covariate x_i, it will be possible to employ conjugate priors to characterize inter-specimen variability.

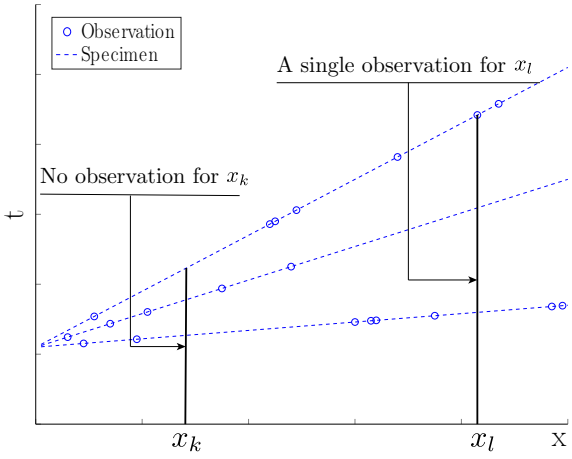


Figure 2: Example of observations obtained on replicated specimens for a set of covariates values. This Figure illustrates the challenge that for most covariate x either only one or no observation is available.

The joint model for multiple specimens employs the modified square exponential covariance function,

$$g(x_k, x_l, s_k, s_l) = \left(\sigma_f^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell^2}\right] + \sigma_v^2 \delta(x_k, x_l)\right) \cdot \delta(s_k, s_l). \tag{2}$$

This new formulation implies that there is no correlation between two distinct specimens, indeed for $s_k \neq s_l$, $g(x_k, x_l, s_k, s_l) = 0$. This allows taking into account dependency within a same specimen. This covariance function enables the creation of a single model sharing the same hyper-parameters $\mathcal{P}_t = \{\sigma_f, \ell, \sigma_v\}$ for all of the N_s specimens and for the N observations. Like for the standard GPR formulation presented in Section 2, the hyper-parameters $\mathcal{P}_t$ are estimated from data using a MLE approach. For each one of the N_s specimens, GPR employs the complete dataset $\mathcal{D} = \{\mathcal{D}_i, i = 1, \ldots, N\}$ to estimate the expected value and covariance for $f(x_{i*}, s_j)$ for any covariate $x_{i*} \in \mathbb{R}$ and any specimen $s_j \in \mathcal{S}$. For example, given a dataset of three specimens, Figure 3 schematizes the predictions of the GPR; note that the between-specimen heteroscedastic uncertainty is not yet considered. Because of the covariance function in Equation 2, which separates the specimens, GPR provides, for a single covariate x_{i*}, the marginal distribution for $f(x_{i*}, s_j)$, and this for each specimen of $\mathcal{S}$. Then, given three tested specimens, GPR results are $\mathbf{f} \sim f(x_{i*}, [s_1, s_2, s_3]^\intercal) = \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$ with the column vectors $\mathbf{M} = \mathbb{E}(\mathbf{f})$ and $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{f})$.

Heteroscedasticity and Conjugate distribution

The methodology proposed to overcome limitations presented in the previous section employs GPR to estimate $f(\mathbf{x}_*, \mathbf{s})$, for any target covariate value x_{i*}, and for any specimen

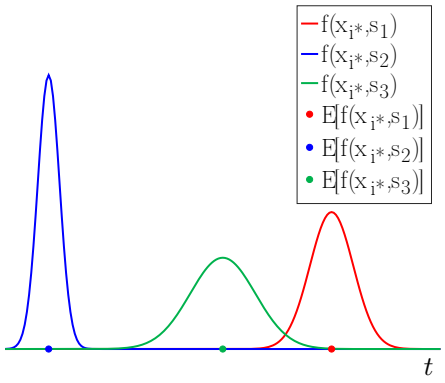


Figure 3: Example of predicted marginal distributions for three specimens for the covariate value x_{i*}.

s_j. Therefore, at a given x_{i*}, even if no actual observation is available for this specific covariate, the N_s missing values are provided by the GPR predictions $f(x_{i*}, \mathbf{s})$. The prediction accuracy depends on the number of observations per specimen, the larger the dataset the more reliable will be the prediction of the specimen behaviour.

As explained in Section 3, the predictive inter-specimens variability $\tilde{T} \sim p(t; \mathcal{D}_i)$ can be modeled with analytic formulations provided by conjugate priors. However, the conjugate prior formulation employed is only suited for perfect observations that are not contaminated by any uncertainties (Murphy 2007). In this paper, GPR outputs are probability density functions (pdfs). Here, a sampling-based approach is employed to marginalize the GPR output uncertainties in order to obtain the posterior predictive pdf $\tilde{T}$. The model explainations will focus on one single covariate value x_{i*} as the method is applicable for any other covariate value x_{i*} with $i = 1, \ldots, N$. That way, for the covariate x_{i*} and given the GPR joint Normal distribution $f(x_{i*}, \mathbf{s})$ with the column vector $\mathbf{s} = [1, 2, \ldots, N_s]^\intercal$, samples $\mathbf{f}_q : \mathbf{f} \sim f(x_{i*}, \mathbf{s}) = \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$ are drawn from the GPR distribution and passed to the conjugate prior so that the new dataset is $\mathcal{D}_i^q = \{(x_{i*}, \mathbf{f}_q)\}$, with q the number of samples.

In order to estimate the posterior predictive pdf $\tilde{T}_q \sim p(t; \mathcal{D}_i^q)$ for the covariate value x_{i*}, the posterior pdf for hyper-parameters $p(\mathcal{P}_t | \mathcal{D}_i^q)$ has to be defined first. Assuming that the variable $T_q \sim p(t; \mathcal{P}_t)$ which describes the inter-specimens variability follows a Normal distribution, the conjugate prior associated to a Normal likelihood $\mathcal{N}(\mu_i, \sigma_i^2)$ with unknown parameters μ_i and σ_i^2 is a Normal-Inverse-Gamma distribution $\mathcal{NIG}(m_0, V_0, a_0, b_0)$ (Murphy 2007). Following Bayes theorem and with the dataset $\mathcal{D}_i^q$, the posterior distribution of the

hyper-parameters $\mathcal{P}_t = \{\mu_i, \sigma_i^2\}$ can be written as,

$$\begin{aligned} p(\mathcal{P}_t|\mathcal{D}_i^q) &= \frac{p(\mathcal{D}_i^q|\mathcal{P}_t)\cdot p(\mathcal{P}_t)}{p(\mathcal{D}_i^q)} \\ &\propto \underbrace{p(\mathcal{D}_i^q|\mathcal{P}_t)}_{\substack{\text{Likelihood}\\ \mathcal{N}(\mu_i,\sigma_i^2)}} \cdot \underbrace{p(\mathcal{P}_t)}_{\substack{\text{Prior}\\ \mathcal{NIG}(m_0,V_0,a_0,b_0)}} . \end{aligned}$$

The hyper-parameters posterior distribution $p(\mu_i, \sigma_i^2|\mathcal{D}_i^q)$ follows a Normal-Inverse-Gamma distribution $\mathcal{NIG}(m_{N_s}, V_{N_s}, a_{N_s}, b_{N_s})$ (Murphy 2007). With the hypothesis that there is no prior knowledge on the hyper-hyper-parameters (parameters of the hyper-parametres), their initial values $\mathcal{P}_c = \{m_0, V_0, a_0, b_0\}$ are chosen to tend to zero (i.e. 10^{-6}) in order to represent the absence of prior knowledge. The hyper-parameters posterior follows,

$$\begin{aligned} p(\mu_i, \sigma_i^2|\mathcal{D}_i^q) &= \mathcal{NIG}(m_{N_s}, V_{N_s}, a_{N_s}, b_{N_s}) \\ V_{N_s}^{-1} &= V_0^{-1} + N_s = N_s, \\ \frac{m_{N_s}}{V_{N_s}} &= V_0^{-1} m_0 + N_s \overline{\mathbf{f}_q} = N_s \overline{\mathbf{f}_q}, \\ a_{N_s} &= a_0 + \frac{N_s}{2} = \frac{N_s}{2}, \\ b_{N_s} &= b_0 + \frac{1}{2}\left[m_0^2 V_0^{-1} + \sum {f_q}^2 - m_{N_s}^2 V_{N_s}^{-1}\right] \\ &= \frac{1}{2}\left[\sum {f_q}^2 - m_{N_s}^2 V_{N_s}^{-1}\right]. \end{aligned}$$

Finally, the posterior predictive is the compound distribution of the Normal prior predictive and the Normal-Inverse-Gamma hyper-parameters posterior. The result is a Student's t-distribution,

$$\begin{aligned} \tilde{T}_q \sim p(t_i|\mathcal{D}_i^q) &= \iint p(t_i|\mu_i, \sigma_i^2)\cdot p(\mu_i, \sigma_i^2|\mathcal{D}_i^q)\, d\mu_i\, d\sigma_i^2 \\ &= t_{2a_{N_s}}\left(m_{N_s}, \frac{b_{N_s}(1+V_{N_s})}{a_{N_s}}\right) \\ &= t_{2a_{N_s}}(\mu_{\tilde{T}}, \sigma_{\tilde{T}_\Sigma}^2). \end{aligned}$$

It must be noted that $\sigma_{\tilde{T}_\Sigma}^2$ is the scale parameter linked to the variance $\sigma_{\tilde{T}}^2$ by

$$\sigma_{\tilde{T}}^2 = \frac{\nu}{\nu-2}\sigma_{\tilde{T}_\Sigma}^2, \quad \text{with } \nu = 2a_{N_s} \text{ posterior degrees of freedom.}$$

This last posterior predictive $\tilde{T}_q$ can be evaluated for every covariate x_{i*} with $i = 1, \ldots, N$.

Prediction of an untested specimen

It is now possible to predict the mean and the variance of an untested specimen $N_s + 1$, relying on the dataset of N_s tested specimens. To do so, the previous method has to be repeated for a large number of samples through a Monte Carlo method. This means sampling Q times the estimates $\mathbf{f}_q$ for the covariate x_{i*}, which will provide Q Student's t-distributions samples $\tilde{t}_q : \tilde{T}_q$. From the posterior predictive distribution $\tilde{T}_q$, the model can predict, for x_{i*}, the response of an untested specimen $N_s + 1$ which is $\tilde{t}_q : \tilde{T}_q \sim t_{2a_{N_s}}(\mu_{\tilde{T}}|\mathbf{f}_q, \sigma^2_{\tilde{T}_\Sigma}|\mathbf{f}_q)$. This prediction is also repeated Q times based on the Q sampled distributions $\tilde{T}_q$ in order to obtain the mean of the specimen $N_s + 1$ for the covariate x_{i*},

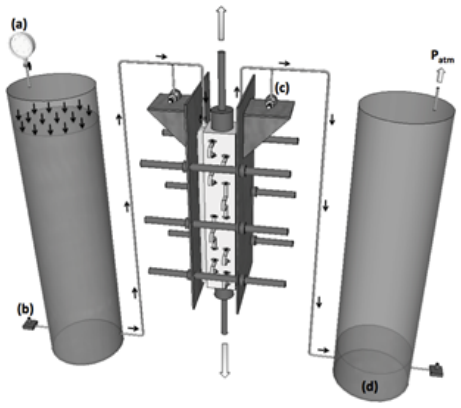
$$\mathbb{E}[\tilde{T}] \approx \frac{1}{Q} \sum_q \tilde{t}_q$$

The empirical confidence interval of the predicted and untested specimen $N_s + 1$ is also evaluated after sampling Q times the posterior predictive Student's t-distributions $\tilde{T}_q$. The Q samples describe the confidence interval for the specimen $N_s + 1$ at a given covariate x_{i*}. The method has been described for a given single covariate value x_{i*}, all of it can be replicated for any other covariate value.

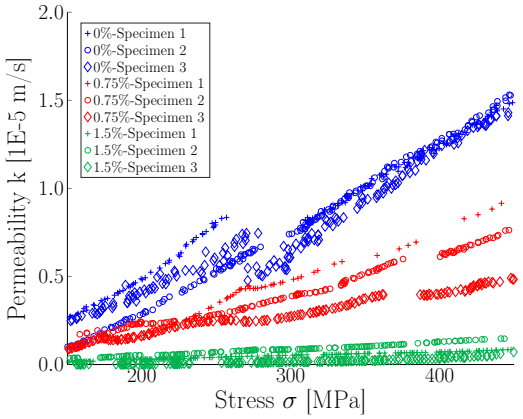
CASE-STUDY: PERMEABILITY OF HIGH PERFORMANCE FIBER-REINFORCED CONCRETE

Test description

The method proposed in this paper is applied to a concrete laboratory experiment. The aim is to model water permeability in high performance fiber-reinforced concrete (HPFRC) in tie-specimens, which is function of the applied stress on the specimen. The model is employed to evaluate the probabilities to obtain a lower water permeability with higher fiber ratios. The dataset studied is the result of experiments performed by Hubert et al. (2015). During these tests, water permeability measurements are performed on reinforced concrete tie-specimens subjected simultaneously to a uniaxial tensile loading. Permeability depends on how long it takes for the water to go through the entire sample. At the same time, average stress is measured in the steel rebar placed inside the tie-specimen as shown in Figure 4a. It can be noted that a tensile loading rate is maintained constant in order to have a progressive cracking spread up to the yielding of the rebar in the tie-specimen.
During the experiment, 9 high performance fiber-reinforced concrete samples were tested, more precisely, 3 samples for 3 different fiber ratios, 0%, 0.75% and 1.5%. Figure 4b presents the set of data obtained from the experimental test. Notice that the range of stress spreads from 150 MPa to 450 MPa which includes both service and ultimate limits. All nine specimens, share a common behaviour ; water permeability increases with stress due to cracks number and cracks width. The graph shows that adding fibers into concrete reduces water permeability by an order of magnitude for fiber ratios 0% and 1.5%. This can be explained by the large number of macro-cracks created in fiber reinforced concrete, whereas



(a) a) Manometer and inlet tank. b) Differential pressure transmitter. c) Pressure sensor. d) Outlet tank



(b) Representation of permeability observations drawn from nine specimens subjected to tensile stress

Figure 4: Permeability measurement setup and representation of the test results dataset.

in standard concrete, cracks are fewer but wider, increasing water permeability known to be proportional to the cube of crack width.

From this set of data, it can be concluded qualitatively that fiber addition reduces significantly water permeability and increases durability. However, raw data does not quantify this benefit. Modeling water permeability using the method proposed in this paper allows estimating the probability to obtain a lower permeability between each pair of fiber ratios.

Probabilistic models

Hypotheses

In this case-study, heteroscedasticity is observable in permeability measurements for $N_s = 3$ replicated specimens which are function of one covariate, the stress measured in rebars. Indeed, the tensile loading on three replicated specimens under the same covariate (stress) did not give the same permeability values, because it is, in practice impossible to manufacture 3 identical prisms of fiber-reinforced concrete. Each fiber ratio, 0%, 0.75%, 1.5%, is studied individually, more precisely three datasets each containing three replicated specimens are examined. Using the method proposed in this paper, the hypotheses assumed regarding the Gaussian Process prior are related to the mean and covariance functions. For the dataset $\mathcal{D} = \{\mathcal{D}_i, i = 1, \ldots, N\}$ with $\mathcal{D}_i = \{(x_i, s_j, y_j), j = 1, \ldots, N_s\}$, the Gaussian Process prior structure is built with the following mean function,

$$\mu(x_i) \quad = \quad a \cdot x_i + b \qquad \text{with } a, b \in \mathbb{R}$$

and covariance function,

$$g(x_k, x_l, s_k, s_l) \quad = \quad \left(\sum_{r=1}^{3} \sigma_{f_r}^2 \exp \left[\frac{-(x_k - x_l)^2}{2\ell_r^2} \right] + \sigma_v^2 \delta(x_k, x_l) \right) \cdot \delta(s_k, s_l)$$

The main structure of the covariance function is the square exponential (SE) but adds two other covariance functions, each having a different correlation length. It allows to consider the impact of one covariate on another at three different scales. This way, the variability on the permeability behaviour may be better approached and represented, whereas one correlation length would only model the average variability. For this study, these new covariance functions have been implemented in the open-source code Gaussian Processes for Machine Learning (GPML) (Rasmussen and Nickisch 2010). To ensure a strictly positive water permeability values, the method is applied in the log-space which means using the log-permeability, the raw dataset becomes $\mathcal{D} = \{(x_i, \log(y_i)), i = 1, \cdots, N\}$. Figure 5 presents the dataset plot in log-space. Note that the scale of the vertical axis has been modified by the log-transformation.

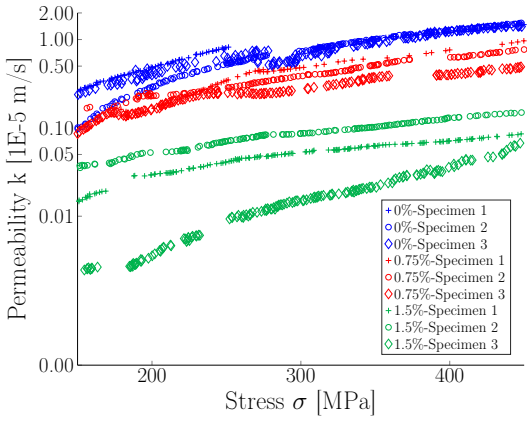


Figure 5: Representation of permeability observations function of the covariate stress in log-space.

Hyper-parameters calibration

Once GPR prior structure is defined, hyper-parameters are identified by MLE. In the model, one set of hyper-parameters, per fiber ratio, has to be estimated, $\mathcal{P}_f = \{a, b, \sigma_{f_1}, \ell_1, \sigma_{f_2}, \ell_2, \sigma_{f_3}, \ell_3, \sigma_v\}$. The difficulty lies in finding the parameters needed to initialize the MLE method. Without a proper choice for initial values, this last estimation could be stuck in a local maximum and miss the global one. Initial parameters are defined by observing the data behaviour at first, and then adjusted to assure the selection of a valid starting point. Table 1 gathers the results of the highest Log-Likelihood (LL) and the

hyper-parameters associated, for the three fiber ratios 0%, 0.75%, 1.5%.

Table 1: MLE Results.

Fiber Ratio	**Hyper-parameters**									**LL**
	a	b	σ_{f_1}	ℓ_1	σ_{f_2}	ℓ_2	σ_{f_3}	ℓ_3	σ_n	$\cdot 10^3$
0%	0.0049	−13.33	0.213	288.42	0.094	76.22	0.012	9.46	0.032	1.70
0.75%	0.0060	−14.53	0.124	222.45	0.017	31.84	0.0024	10.25	0.031	1.30
1.5%	0.0080	−17.50	0.337	271.76	0.041	53.15	0.013	8.01	0.029	2.10

Results

Prediction of a new specimen

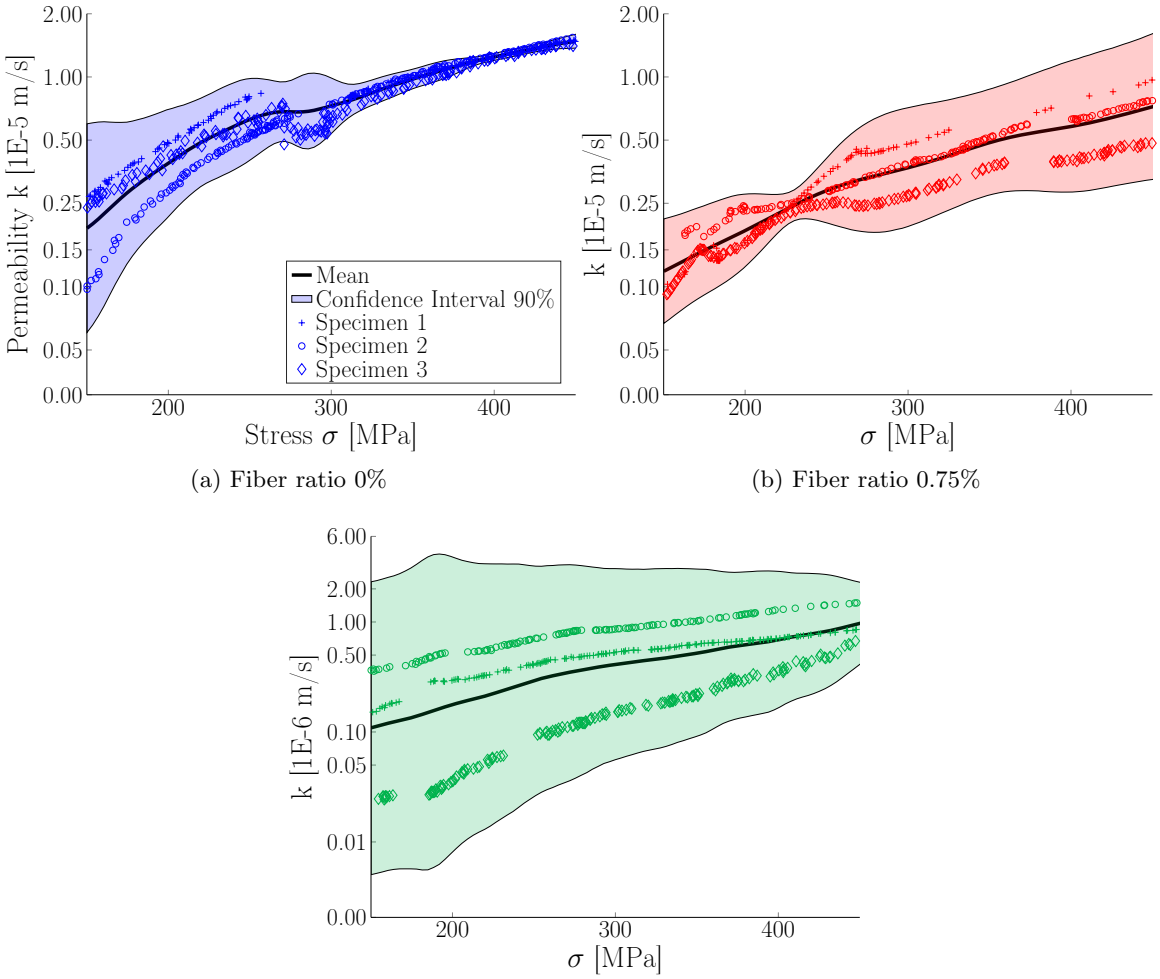
Figures 6 and 7 present the results of the method proposed applied to high performance fiber-reinforced concrete experiments. They show the prediction of a fourth untested specimen relying on the data from three tested ones, and that for three fiber reinforcement ratios. Figure 6 presents the results in the log-transformed space, each graph matching a fiber ratio. The possible response of a fourth specimen is modeled over the stress interval $\sigma = [150, 450]$ MPa, by a mean (thick black line) and its 90% confidence interval. In every case, notice that, where water permeability observations spread, the confidence interval is larger and where data points are concentrated the confidence interval tightens. The probabilistic model in Figure 6c is displayed with a different scale for the vertical axis in order to obtain a better visualization of the behaviour.

The set of graphs in Figure 7 present the results in the original space. The transformation from the log to the original space tends to increase the confidence interval's upper bond. As permeability values are close to zero but at the same time remain positive, uncertainty is skewed towards positive values. This explains the wide confidence interval for the reinforcement ratio 1.5% since for this one, water permeabilities are closer to zero than for the two other lower fiber ratios.

Comparison of fiber reinforcement ratios

The main goal of this case-study is to quantify the fiber addition benefits. This Section compares the three tested reinforcement ratios ; the two first graphs in Figure 8 juxtapose the three predictive models in the log and original spaces. Figure 8 shows that there are overlaps in confidence intervals for several stress values. The conditional probability $\Pr(k_j < k_{j'}|\sigma)$, $j \neq j' \in [1, N_s]$, is estimated by comparing samples from pairs of fiber ratios. A number of $Q = 50000$ samples are drawn from the Student's t-distributions $\tilde{t}_q : \tilde{T}_q \sim t_{2a_{N_s}}(\mu_{\tilde{T}}|\mathbf{f}_q, \sigma^2_{\tilde{T}_\Sigma}|\mathbf{f}_q)$, estimated for a range of stress values σ.

In order to compute the joint probability over a stress interval $\Pr(k_j < k_{j'}|150\,\text{MPa} \leq \sigma \leq 450\,\text{MPa})$ it is essential to consider the correlation between k_j and $k_{j'}$ as a function of stress values σ described in Section 4 for the GPR model. The results of samples



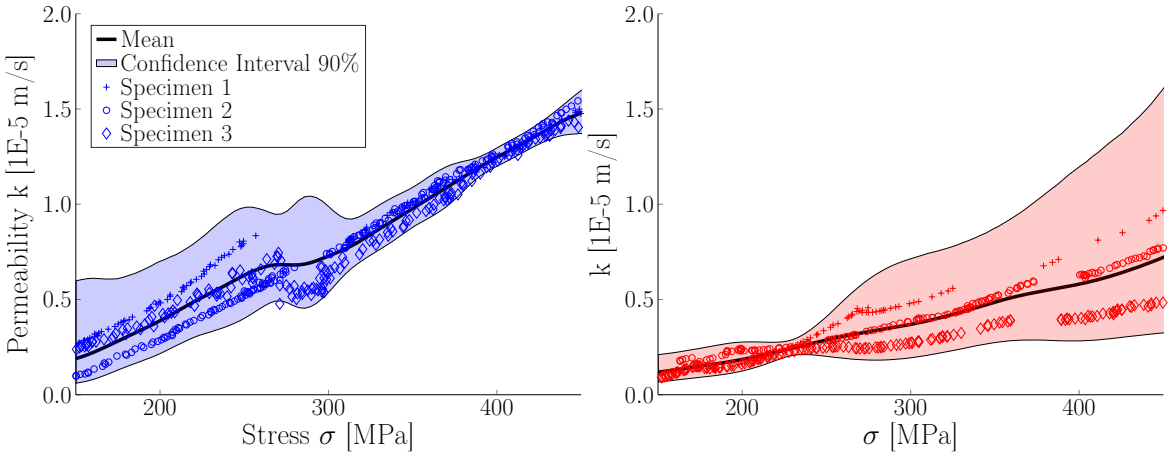
(a) Fiber ratio 0%

(b) Fiber ratio 0.75%

(c) Fiber ratio 1.5%

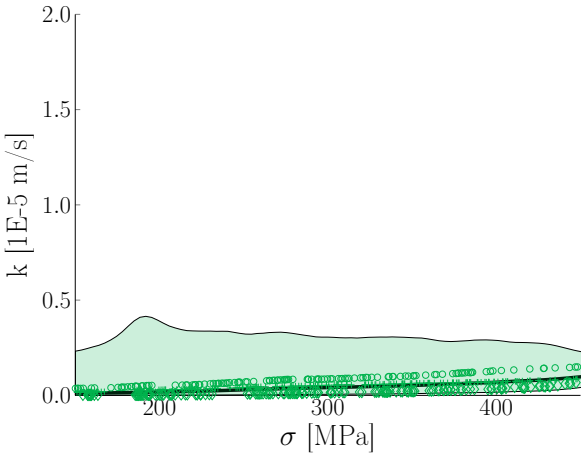
Figure 6: Prediction of water permeability for a fourth untested specimen in the log-space.

comparisons are summarized in Figure 8c. In this case, the variable $\tilde{T}_q$ follows a Multivariate



(a) Fiber ratio 0%

(b) Fiber ratio 0.75%



(c) Fiber ratio 1.5%

Figure 7: Prediction of water permeability for a fourth and untested specimen in the original space.

t-distribution,

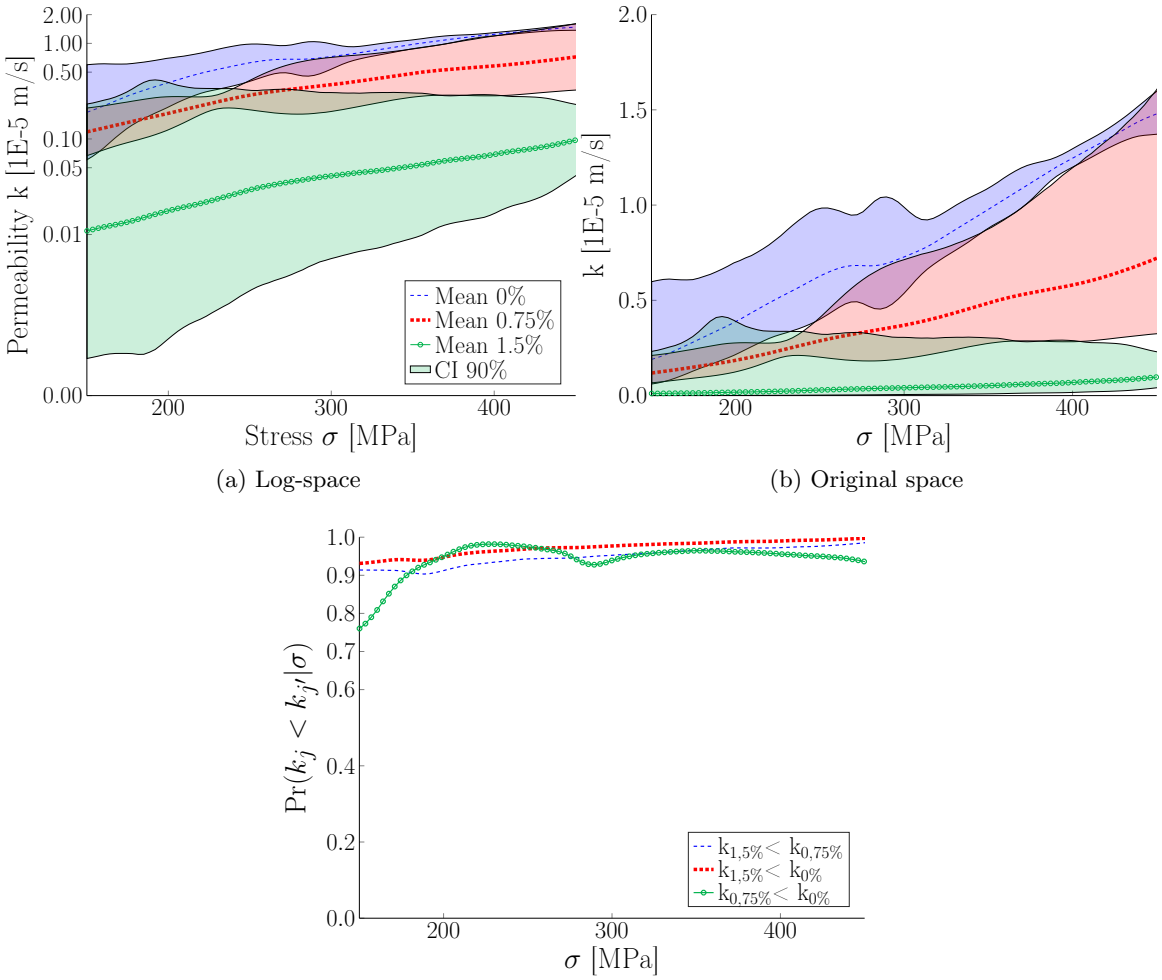
$$\tilde{T}_q \sim T(\mathbf{M}_{\tilde{T}}, \mathbf{\Sigma}_{\tilde{T}}, \nu) \quad \text{with,}$$

$$\mathbf{M}_{\tilde{T}} = \begin{bmatrix} \mu_{\tilde{T}_1} \\ \mu_{\tilde{T}_2} \\ \vdots \\ \mu_{\tilde{T}_N} \end{bmatrix},$$

$$\mathbf{\Sigma}_{\tilde{T}} = \begin{bmatrix} \sigma^2_{\tilde{T}_{\Sigma_1}} & \rho_{1,2}\sigma_{\tilde{T}_{\Sigma_1}}\sigma_{\tilde{T}_{\Sigma_2}} & \cdots & \rho_{1,N}\sigma_{\tilde{T}_{\Sigma_1}}\sigma_{\tilde{T}_{\Sigma_N}} \\ \rho_{2,1}\sigma_{\tilde{T}_{\Sigma_2}}\sigma_{\tilde{T}_{\Sigma_1}} & \sigma^2_{\tilde{T}_{\Sigma_2}} & \cdots & \rho_{2,N}\sigma_{\tilde{T}_{\Sigma_2}}\sigma_{\tilde{T}_{\Sigma_N}} \\ \vdots & \vdots & \ddots & \cdots \\ \rho_{N,1}\sigma_{\tilde{T}_{\Sigma_N}}\sigma_{\tilde{T}_{\Sigma_1}} & \rho_{N,2}\sigma_{\tilde{T}_{\Sigma_N}}\sigma_{\tilde{T}_{\Sigma_2}} & \cdots & \sigma^2_{\tilde{T}_{\Sigma_N}} \end{bmatrix},$$

$$\nu = 2a_{N_s} = \frac{N_s}{2}$$

The sampling method comprises sampling from $\tilde{T}_0 \sim T(0, 1, \nu)$ and then transforming $\tilde{T}_q = M_{\tilde{T}} + R^\intercal \cdot \tilde{T}_0$ with $R = \text{chol}(\Sigma_{\tilde{T}})$, $(R^\intercal R = \Sigma_{\tilde{T}})$.



(a) Log-space

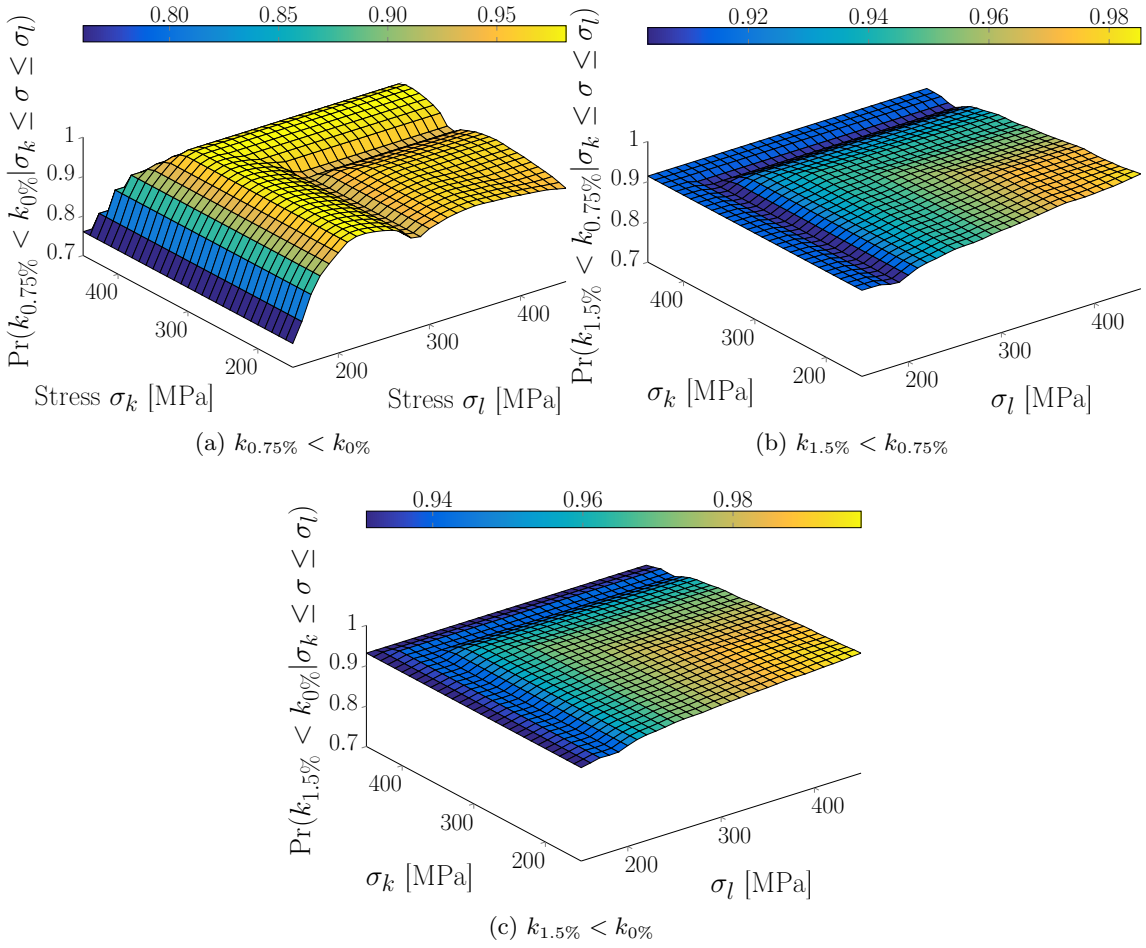
(b) Original space

(c) Conditional probabilities

Figure 8: (a),(b) Predictive models and (c) comparison of conditional probabilities of permeability in high-performance fiber reinforced concrete for the tested fiber reinforcement ratios.

Figure 9 presents $\Pr(k_j < k_{j'} | \sigma_k \leq \sigma \leq \sigma_l)$ between pairs of fiber reinforcement ratios for any stress interval. For a chosen stress value σ_k from axis X and a chosen stress value σ_l from axis Y, the surface provides the global probability of exceedance between two fiber ratios, over the stress interval $[\sigma_k, \sigma_l]$. Notice that the diagonal cross-section for which

$\sigma_k < \sigma_l$ matches conditional probabilities from Figure 8c.



(a) $k_{0.75\%} < k_{0\%}$ (b) $k_{1.5\%} < k_{0.75\%}$

(c) $k_{1.5\%} < k_{0\%}$

Figure 9: Global probabilities of $\Pr(k_j < k_{j'} | \sigma_k \leq \sigma \leq \sigma_l)$ for permeability across each pair of fiber ratios.

The probabilities $\Pr(k_j < k_{j'} | 150\,\text{MPa} \leq \sigma \leq 450\,\text{MPa})$ are presented in Table 2 for the stress interval $\sigma = [150, 450]$ MPa. The probability to obtain a lower water permeability by adding 1.5% of fiber reaches 93% and 91% when comparing with fiber ratios 0% and 0.75% respectively. Likewise, the probability to get a lower water permeability in concrete with a reinforced ratio of 0.75% over 0% is 76%. These probabilities support the qualitative assessment that fiber addition decreases permeability.

DISCUSSION

Probabilistic analyses demonstrate that an introduction of 0.75% or 1.5% of fibers

Table 2: Probabilities over the entire stress interval to obtain lower water permeabilities with higher fiber ratios.

	$\Pr(k_j < k_{j'} \mid 150\,\text{MPa} \leq \sigma \leq 450\,\text{MPa})$
$k_{0.75\%} < k_{0\%}$	0.76
$k_{1.5\%} < k_{0.75\%}$	0.91
$k_{1.5\%} < k_{0\%}$	0.93

in concrete will, with a very high level of confidence, decrease water permeability in concrete structures under load. As permeability is the main indicator of the durability of cracked concrete, the incorporation of fibers will provide an extended durability to concrete structures in service conditions. This statement obtained by the treatment of experimental results by a probabilistic approach increases the value to be given to the results. It provides a quantitative or a better certainty of the trends measured experimentally.

Despite the efficiency of the method, some limits remain ; the main difficulty lies in the Gaussian Process hyper-parameters estimation. The MLE method can lead to a local maximum likelihood instead of the global one and then providing biased hyper-parameters. A careful choice of initial parameters values is therefore essential. Future work could further study the potential of Bayesian parameter estimation for this purpose.

In the application of the method, three specimens were enough to provide a consistent model of water permeability over stress. Also, it would be interesting to analyse the results with the addition of a fourth studied specimen in the dataset and observe if the confidence interval would decrease significantly in the prediction of a fifth untested specimen.

CONCLUSION

This paper proposes a new extension to Gaussian Process Regression for creating probabilistic models from few laboratory specimens displaying a heteroscedastic behaviour. The key aspect of this method resides in the combination of GPR and conjugate priors. This new method can be applied to replicated specimens observations obtained from any laboratory experiments. The application of this new method to a HPFRC case-study probabilistically quantified how adding fibers to high performance concrete decreases water permeability.

References

Bansal, A. K. and Aggarwal, P. (2007). "Bayes prediction for a heteroscedastic regression superpopulation model using balanced loss function." *Communications in Statistics—Theory and Methods*, 36(8), 1565–1575.

Blau, G., Lasinski, M., Orcun, S., Hsu, S.-H., Caruthers, J., Delgass, N., and Venkatasubramanian, V. (2008). "High fidelity mathematical model building with experimental data: A bayesian approach." *Computers & Chemical Engineering*, 32(4), 971–989.

Gardoni, P., Der Kiureghian, A., and Mosalam, K. (2002). "Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations." *Journal of Engineering Mechanics*, 128(10), 1024–1038.

Gardoni, P., Nemati, K., and Noguchi, T. (2007). "Bayesian statistical framework to construct probabilistic models for the elastic modulus of concrete." *Journal of Materials In Civil Engineering*, 19(10), 898–905.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, Vol. 2. Taylor & Francis.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1997). "Regression with input-dependent noise: A Gaussian process treatment." *Advances in neural information processing systems*, 10, 493–499.

Hubert, M., Desmettre, C., and Charron, J.-P. (2015). "Influence of fiber content and reinforcement ratio on the water permeability of reinforced concrete." *Materials and Structures*, 1–13.

Jeon, J.-S., Shafieezadeh, A., and DesRoches, R. (2014). "Statistical models for shear strength of rc beam-column joints using machine-learning techniques." *Earthquake Engineering & Structural Dynamics*, 43(14), 2075–2095.

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. (2007). "Most likely heteroscedastic Gaussian process regression." *Proceedings of the 24th international conference on Machine learning*, ACM, 393–400.

Lampinen, J. and Vehtari, A. (2001). "Bayesian approach for neural networks—review and case studies." *Neural networks*, 14(3), 257–274.

Le, Q. V., Smola, A. J., and Canu, S. (2005). "Heteroscedastic Gaussian process regression." *Proceedings of the 22nd international conference on Machine learning*, ACM, 489–496.

Ma, Y., Wang, L., Zhang, J., Xiang, Y., and Liu, Y. (2014). "Bridge remaining strength prediction integrated with Bayesian network and in situ load testing." *Journal of Bridge Engineering*, 19(10), 04014037.

MacKay, D. (1998). "Introduction to Gaussian processes." *NATO ASI Series F Computer and Systems Sciences*, 168, 133–166.

Murphy, K. P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution." 16.

Pal, M. and Deswal, S. (2008). "Modeling pile capacity using support vector machines and generalized regression neural network." *Journal of geotechnical and geoenvironmental engineering*, 134(7), 1021–1024.

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

Rasmussen, C. and Williams, C. (2006). "Gaussian processes for machine learning." *the MIT Press*, 2(3), 4.

Siddique, R., Aggarwal, P., Aggarwal, Y., and Gupta, S. (2008). "Modeling properties of self-compacting concrete: support vector machines approach." *Computers and Concrete*, 5(5), 123–129.

Słoński, M. (2010). "A comparison of model selection methods for compressive strength prediction of high-performance concrete using neural networks." *Computers & structures*, 88(21), 1248–1253.

Słoński, M. (2011). "Bayesian neural networks and gaussian processes in identification of concrete properties." *Computer Assisted Mechanics and Engineering Sciences*, 18(4), 291–302.

Thiyagarajan, K. and Kodagoda, S. (2016). "Analytical model and data-driven approach for concrete moisture prediction." *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, Vol. 33, Vilnius Gediminas Technical University, Department of Construction Economics & Property, 1.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). "Variational heteroscedastic Gaussian process regression." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 841–848.

Tolvanen, V., Jylänki, P., and Vehtari, A. (2014). "Expectation propagation for nonstationary heteroscedastic Gaussian process regression." *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 1–6.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2012). "Bayesian modeling with Gaussian processes using the GPstuff toolbox." *arXiv preprint arXiv:1206.5754*.

Wang, C. (2014). "Gaussian process regression with heteroscedastic residuals and fast MCMC methods." Ph.D. thesis, University of Toronto, University of Toronto.

Wang, C. and Neal, R. M. (2012). "Gaussian process regression with heteroscedastic or non-Gaussian residuals." *arXiv preprint arXiv:1212.6246*.

Yeh, I. C. (2014). "Estimating distribution of concrete strength using quantile regression neural networks." *Applied Mechanics and Materials*, 584, 1017.

Zhong, J., Gardoni, P., Rosowsky, D., and Haukaas, T. (2008). "Probabilistic seismic demand models and fragility estimates for reinforced concrete bridges with two-column bents." *Journal of engineering mechanics*, 134(6), 495–504.