UNIVERSITÉ DE MONTRÉAL

MODÉLISATION PROBABILISTE D'ESSAIS EN LABORATOIRE PAR PROCESSUS GAUSSIEN AVEC PEU DE SPÉCIMENS RÉPLIQUÉS

LUCIE TABOR
DÉPARTEMENT DE GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
JUILLET 2017

© Lucie Tabor, 2017.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

MODÉLISATION PROBABILISTE D'ESSAIS EN LABORATOIRE PAR PROCESSUS GAUSSIEN AVEC PEU DE SPÉCIMENS RÉPLIQUÉS

présenté par : TABOR Lucie
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. TREMBLAY Robert, Ph. D., président
M. GOULET James-A., Ph. D., membre et directeur de recherche
M. CHARRON Jean-Philippe, Ph. D., membre et codirecteur de recherche
M. MARCOTTE Denis, Ph. D., membre

DÉDICACE

À Léon et mes parents,
À Chris

REMERCIEMENTS

Je souhaite tout d'abord remercier mon directeur de recherche, le professeur James-A. Goulet, pour l'opportunité qu'il m'a offerte en me permettant de travailler sur ce projet de recherche ainsi que pour ses idées et sa disponibilité. En travaillant à ses côtés j'ai pu découvrir le domaine des modélisations probabilistes et alors m'enrichir de son expertise. Je tiens également à remercier mon codirecteur, le professeur Jean-Philippe Charron qui a rendu ce projet concret en alliant au monde abstrait des probabilités, la spécialité du béton. Son expérience, son imagination, sa rigueur et ses encouragements ont permis de mener à bien ce projet de recherche. Je remercie aussi mes deux directeurs pour leur soutien financier.

Je remercie également Clélia Desmettre pour sa participation au projet. Sa connaissance et son expérience m'ont aidé à mieux appréhender ce projet et à le débuter dans des conditions de travail optimales.

Ensuite, je souhaite remercier les professeurs Robert Tremblay et Denis Marcotte pour avoir accepté de faire partie du jury d'examen de cette maîtrise.

Merci à tous ceux formant le groupe de recherche structure (GRS) qui ont su rendre ces deux années de maîtrise conviviales et agréables.

Finalement, je souhaite remercier mes amis, Olivier, Ha, Nelson et Sayouba rencontrés pendant ma maîtrise et qui ont été présents pour moi. Merci à ma famille, qui bien que loin, a réussi à me soutenir tout au long de mes études. Et enfin, un grand merci à Chris pour son soutien inébranlable.

RÉSUMÉ

Les essais expérimentaux en laboratoire, tels que les essais sur des spécimens de béton, demandent de grandes ressources financières. En effet, fabriquer un spécimen de béton peut coûter quelques milliers de dollars. Ainsi, il est fréquent que seuls quelques spécimens répliqués soient soumis à l'essai étudié. Le jeu de données expérimentales est constitué des observations mesurées qui sont fonction des valeurs de la donnée d'entrée soit de la covariable. Ce jeu de données constitué à partir de seulement quelques spécimens testés est alors assez pauvre en informations. Il devient donc difficile, en s'appuyant sur ce faible nombre de données, d'obtenir un modèle pertinent du comportement observé. Ce projet de recherche vise ainsi à établir une méthodologie permettant de modéliser de façon probabiliste un comportement étudié en laboratoire, à partir d'un jeu de données pauvre en spécimens testés. La modélisation probabiliste fournit un modèle prédictif du comportement qui est associé à un intervalle de confiance. Les méthodes de modélisation probabiliste existantes s'appuient sur des jeux de données riches en spécimens et sont capables de modéliser les tendances observées en laboratoire, mais restent incapables de fournir un modèle pertinent pour des jeux de données composés de quelques spécimens. La méthodologie proposée s'appuie sur une méthode probabiliste de Régression par Processus Gaussien (GPR) dans le but d'étendre son application à des jeux de données limités.

Les quelques spécimens répliqués testés peuvent présenter un comportement hétéroscédastique, il existe une variabilité entre les spécimens puisqu'il est quasi-impossible de reproduire deux spécimens identiques. En effet, le scientifique ne peut pas contrôler chaque paramètre physique. La modélisation probabiliste doit être capable de tenir compte de cette variabilité inter-spécimens et de l'incertitude liée au faible nombre de spécimens formant le jeu de données.

Au cours de cette recherche deux approches ont été développées afin de répondre aux objectifs suivants : élaborer une méthodologie visant à modéliser de manière probabiliste des essais en laboratoire effectués sur quelques spécimens répliqués qui eux-mêmes présentent un comportement hétéroscédastique. La première méthode par processus Gaussien rapide et efficace cherche tout d'abord à prédire, dans un même processus Gaussien, le comportement de chaque spécimen testé à partir du jeu de données total tiré des essais expérimentaux. Au sein de ce processus Gaussien, la structure à priori est construite à travers les fonctions de moyenne et de covariance paramétrisées par des hyper-paramètres qui sont inférés à travers l'estimation par maximum de vraisemblance (MLE). Suite à cela, la variabilité inter-

spécimens et l'incertitude sur le nombre de spécimens sont caractérisées par une distribution définie à partir de distributions a priori conjuguées décrites analytiquement. Ces dernières formulations analytiques rendent la méthode par processus Gaussien peu coûteuse en temps et programmation. Son efficacité a été testée sur des jeux de données réels résultant d'essais de perméabilité à l'eau effectués sur les Bétons à Hautes Performances (BHP) et les Bétons Fibrés à Hautes Performances (BFHP).

L'étude de la perméabilité à l'eau chez les bétons est un indicateur clef sur leur durabilité. Les essais de perméabilité à l'eau mettent en évidence une certaine tendance de la perméabilité mais n'ont pu mener à sa modélisation. Les résultats expérimentaux montrent que l'incorporation de fibres réduit la perméabilité et que la perméabilité diminue lorsque la teneur en fibres est augmentée. La modélisation probabiliste cherche à fournir un modèle prédictif du comportement de la perméabilité à l'eau en fonction de la contrainte de traction appliquée aux tirants en BHP et en BFHP. Trois conditions d'essais ont été testées, soient pour trois taux de fibres, 0%, 0.75% et 1.5%, et chacune de ces conditions a été testée pour trois spécimens répliqués. Ainsi, ce projet a pour objectifs de fournir trois modèles probabilistes, soit un pour chaque taux de fibres, et de quantifier le bénéfice d'incorporer des fibres au béton par comparaison des trois modèles.

L'application de la première méthode par processus Gaussien a permis de répondre à ces dernières attentes et de montrer de manière quantitative que l'ajout de fibres dans le béton réduit grandement la perméabilité à l'eau. La perméabilité à l'eau se voit réduite en insérant 0.75% de volume de fibres dans le BHP et de manière plus significative en doublant ce taux de fibres. La modélisation probabiliste confirme le fort potentiel du BFHP à prolonger sa durabilité.

Bien que la méthode par processus Gaussien se soit montrée pertinente, fournissant des résultats cohérents sur le BFHP, une seconde méthode a été considérée au sein de cette recherche. Une approche alternative toujours basée sur la méthode GPR, plus lourde certes, mais plus précise et offrant plus de possibilités, mais cette fois basée sur une conception spatiale de l'hétéroscédasticité et sur une estimation Bayesienne des paramètres. Chaque spécimen peut être décrit par des coordonnées spatiales virtuelles et ces points sont séparés par des distances modélisant la variabilité inter-spécimens. Ces coordonnées virtuelles à inférer sont appelées variables latentes et s'ajoutent à la liste d'hyper-paramètres décrivant les fonctions de moyenne et covariance de la méthode GPR. Ceux-ci sont tirés par échantillonnage de la formulation Bayesienne empirique. La méthode d'échantillonnage rend la modélisation coûteuse en temps contrairement à la méthode MLE choisie pour la méthode par processus Gaussien, qui est plus rapide mais qui occulte l'incertitude liée aux hyper-paramètres. L'application

de cette seconde méthode par approche Bayesienne aux jeux de données tirés des essais de perméabilité n'a pas pu conduire à des résultats cohérents, en effet la formulation de Bayes de la distribution conjuguée à postériori des hyper-paramètres est fortement dépendante de leur connaissance à priori. Cette méthode par approche Bayesienne reste à développer puisque présente du potentiel dans la formulation des modèles prédictifs. Grâce à cette visualisation spatiale des spécimens, la méthode a le potentiel d'augmenter les dimensions en termes de covariables. A partir des jeux de données de perméabilité, il serait possible de prédire le comportement pour un taux de fibres non testé contrairement à la méthode par processus Gaussian qui a besoin de nombreuses valeurs de taux de fibres pour une telle prédiction.

Finalement, ce projet de recherche a permis de mettre en place une méthodologie basée sur le processus Gaussien capable de modéliser des essais en laboratoire sur peu de spécimens à comportement hétéroscédastique. Son application aux essais de perméabilité a permis de quantifier le bénéfice d'incorporation de fibres dans le béton pour la réduction de la perméabilité. Cette méthode par processus Gaussien s'est montrée performante et est applicable à tout type d'essais et dans divers domaines de génie.

ABSTRACT

Laboratory experiments can be very expensive and so require an important budget, for example manufacturing concrete specimens can cost thousands of dollars. That is why, it is common to only test few replicated specimens. Because of a weak set of data, it is difficult to draw a reliable model from it, only the behaviour tendency can be caught and qualitative conclusions can be made. This project aims at building a probabilistic method able to model laboratory experiments with a set of data including only few replicated specimens. Probabilistic models catch the mean behaviour and its associated confident interval for a predictive untested specimen. Existing probabilistic methods rely on an extended set of data making their common hypothesis over observations independency relevant. However with only few replicated specimens, this last assumption becomes inaccurate and so become these methods. The developed methods in this project are based on the well-known and commonly used Gaussian Process Regression (GPR) which will be then extended to overcome the limit of poor datasets and then take into account the uncertainty coming from the low number of specimens.

Tests on replicated specimens often provide different observations and the resulting data set displays what is called an heteroscedastic behaviour which can be defined as the inter-specimens variability. In fact, it is almost impossible to manufacture two identical specimens as it is extremely difficult to control every single physical parameter. Another objective of the probabilistic modeling method is then to catch this heteroscedasticity.

During this research two methods have been developed to meet the previous objectives. The first one based on Gaussian process can be resumed in two major steps : first, each specimen is modeled through a same Gaussian process. In GPR, the prior structure is built based on the mean and covariance functions which are described by hyper-parameters. These last parameters are infered thanks to the method Maximum Likelihood Estimation (MLE). The second step aims at modeling the heteroscedasticity and the uncertainty resulting from the number of specimens. Over each covariable value, observations can be described by a posterior distribution built thanks to conjugate priors which provide an analytical formulation of it. The relevance and efficiency of this method have been highlighted thanks to its application on water permeability experiments on High Performance Concrete (HPC) and High Performance Fiber Reinforced Concrete (HPFRC). The data set is composed of nine specimens, three replicated specimens for each one of the following pourcentage of fibers 0%, 0.75% and 1.5%. For each fiber ratio the GPR method provides a model of the water

permeability depending on the covariate tensile stress applied to the specimen. The main goal of this probabilistic modeling is to confirm experimental conclusions which suppose that water permeability decreases with the addition of fibers into the concrete. The comparison of the three models has enabled to estimate quantitatively this benefit and has shown that the permeability is being reduced with a fibers addition of 0.75% in volume and even more reduced with a fiber ratio of 1.5%. These consistent results comfort the adequacy of this first GPR method which is quick and efficient thanks to the analytical formulation of the conjugate prior and the use of MLE to infer hyper-parameters. The application allows to confirm the potential of HPFRC to be a lasting concrete as permeability is the major indicator of durability once the concrete shows cracks.

Another method has been developed in this study, also based on GPR, the majors changes live in the heteroscedasticity visualisation and the inference method of hyper-parameters. In this second approach, specimens are represented by points in a virtual space and separate by distances modeling the inter-specimens variability. Virtual coordinates or distances are called latent covariates and have to be inferred as well as the mean and covariance functions hyper-parameters. This time, a Bayesian approach is preferred to MLE as more exact since the hyper-parameters variability is considered. The hyper-parameters posterior distribution is built thanks to a hierarchical Bayes formulation but as no easy analytic formulation exists, hyper-parameters samples have to be drawn. This sampling method makes this Bayesian approach heavy, it requires time and major computing. Moreover the Bayesian approach strongly depends on the prior knowledge and since the dataset of permeability tests is quite poor, resulting models lack consistency. However, this method offers openings for more complex modeling, in fact thanks to the virtual spatial description of the specimens it seems possible to add other covariates such as fiber ratio and then predict the permeability behaviour for an untested pourcentage of fiber which is impossible with the first method which requires more different fiber ratio values.

Finally, this research project has enabled to develop an efficient probabilistic modeling method for laboratory experiments performed on few replicated specimens displaying an heteroscedastic behaviour. The application of the GPR method on a concrete set of data allowed to confirm the experimental tendency of water permeability in HPFRC. Increasing the volume of fibers in HPFRC will reduce significantly water permeability giving HPFRC a great potential regarding durability, indeed permeability is a major indicator of durability in cracked concrete. This probabilistic modeling method GPR, can be adapted to any laboratory experiment in any engineering field.

TABLE DES MATIÈRES

DÉDICACE iii

REMERCIEMENTS iv

RÉSUMÉ v

ABSTRACT viii

TABLE DES MATIÈRES x

LISTE DES TABLEAUX xiii

LISTE DES FIGURES xiv

LISTE DES SIGLES ET ABRÉVIATIONS xvi

CHAPITRE 1 INTRODUCTION 1
1.1 Mise en contexte et problématique 1
1.2 Cas d'étude 2
1.3 Objectifs 3
1.4 Méthodologie 3
1.5 Organisation du mémoire 4

CHAPITRE 2 REVUE DE LA DOCUMENTATION 5
2.1 Méthodes de régression appliquées au génie civil 5
2.1.1 De la méthode simple à la plus complexe 5
2.1.2 Approche Bayesienne 7
2.1.3 Régression par processus Gaussien et estimation par maximum de vraisemblance 10
2.2 Hétéroscédasticité 13
2.2.1 Différentes approches de l'erreur de mesure 13
2.2.2 Régression par processus Gaussien et hétéroscédasticité 15
2.3 BFHP : Béton Fibré à Hautes Performances 17
2.3.1 Formulation du BFHP 17
2.3.2 Comportement en traction 18
2.3.3 Durabilité du béton et perméabilité à l'eau 20

2.3.4 Essai de perméabilité sur des tirants en BFHP 22
2.4 Conclusion 25

CHAPITRE 3 MÉTHODOLOGIE 26
3.1 Essai expérimental à l'étude 26
3.1.1 Jeu de données 26
3.1.2 Objectifs de la modélisation probabiliste 27
3.2 Modélisation : de l'expérimental au numérique 29
3.2.1 Interprétation probabiliste des données 29
3.2.2 Deux approches probabilistes 29
3.3 Méthode par processus Gaussien 30
3.3.1 Description 30
3.3.2 Résultats 30
3.4 Méthode par approche Bayesienne 30
3.4.1 Description 31
3.4.2 Résultats 31

CHAPITRE 4 ARTICLE 1 : PROBABILISTIC MODELING OF HETEROSCEDASTIC LABORATORY EXPERIMENTS USING GAUSSIAN PROCESS REGRESSION 32
4.1 Abstract 32
4.2 Introduction 32
4.3 Gaussian Process Regression 35
4.3.1 Model definition 35
4.3.2 Hyper-parameter estimation 37
4.4 GPR for sparse and heteroscedastic datasets 38
4.4.1 Combining GPR and conjugate priors 39
4.4.2 Heteroscedasticity and Conjugate distribution 40
4.4.3 Prediction of an untested specimen 42
4.5 Case-Study : Permeability of High Performance Fiber-Reinforced Concrete . 43
4.5.1 Test description 43
4.5.2 Probabilistic models 44
4.5.3 Results 46
4.6 Discussion 51
4.7 Conclusion 52

CHAPITRE 5 DISCUSSION GÉNÉRALE ET APPROCHE COMPLÉMENTAIRE 56

5.1 Discussion générale : méthode par processus Gaussien 56
5.1.1 Résultats de l'application de la méthode par processus Gaussien aux essais de perméabilité 56
5.1.2 Une méthode performante pour un grand champ d'applications . . . 57
5.1.3 Nombre de spécimens testés 57
5.1.4 Limites de la méthode par processus Gaussien 58
5.2 Méthode par approche Bayesienne 59
5.2.1 Représentation spatiale de l'hétéroscédasticité 59
5.2.2 Echantillonnage des hyper-paramètres et hyper-hyper-paramètres . . 63
5.2.3 Prédiction 65
5.2.4 Structure d'ensemble de la méthode par approche Bayesienne 66
5.2.5 Conclusion 67
5.3 Discussion générale : méthode par approche Bayesienne 68
5.3.1 Une approche interprétable physiquement 68
5.3.2 Vers de plus grandes dimensions 68
5.3.3 Limite de la méthode par approche Bayesienne 69
5.3.4 Comparaison des deux approches 70

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS 71
6.1 Synthèse des travaux 71
6.2 Conclusions 72
6.3 Recommandations 73
6.4 Améliorations futures 74

RÉFÉRENCES 75

LISTE DES TABLEAUX

Tableau 2.1	Caractéristiques du BHP et du BFHP (adapté de Charron et al. (2016))	18
Tableau 2.2	Programme expérimental (adapté de Hubert et al. (2015))	23
Tableau 2.3	Composition des bétons (adapté de Hubert et al. (2015))	23
Table 4.1	MLE Results - Model 1	46
Table 4.2	MLE Results - Model 2	46
Tableau 4.3	Probabilities over the entire stress interval to obtain lower water permeabilities with higher fiber ratios.	52

LISTE DES FIGURES

Figure 2.1 Jeux de données présentant (a) un comportement homoscédastique, (b) un comportement hétéroscédastique and (c) un comportement hétéroscédastique avec observations dépendantes entre elles. 6

Figure 2.2 Exemple de fonctions de covariance exponentielle carré unidimensionnelles pour différentes longueurs de corrélation ℓ (tiré de Quach et al. (2017)) 11

Figure 2.3 Comportement en traction uniaxiale des bétons (tiré de Charron et al. (2016)) 19

Figure 2.4 Patron de fissuration dans différents bétons pour une ouverture de fissure totale de 0.3 mm (tiré de Charron et al. (2016)) 20

Figure 2.5 Patrons de fissuration des tirants de BO et de BRF (tiré de Berrocal et al. (2013)) 21

Figure 2.6 Instrumentation de l'essai de perméabilité (tiré de Plagué et al. (2017)) 23

Figure 2.7 Perméabilité en fonction de la contrainte dans l'armature pour quatre dosages de fibres (tiré de Hubert et al. (2015)) 24

Figure 3.1 Représentation graphique du jeu de données résultant des essais de perméabilité sur neuf spécimens de BHP et BFHP 27

Figure 3.2 Représentations graphiques par taux de fibres des jeux de données résultant des essais de perméabilité sur trois spécimens de BHP ou BFHP 28

Figure 4.1 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. 33

Figure 4.2 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. 39

Figure 4.3 Example of predicted marginal distributions for three specimens for the attribute value x_{i*}. 40

Figure 4.4 Permeability measurement setup and representation of the test results dataset. 43

Figure 4.5 Representation of permeability observations function of the attribute stress in log-space. 45

Figure 4.6 Prediction of water permeability for a fourth untested specimen in the log-space. 47
Figure 4.7 Prediction of water permeability for a fourth and untested specimen in the original space. 48
Figure 4.8 (a),(b) Predictive models and (c) comparison of permeability in high-performance fiber reinforced concrete for the tested fiber reinforcement ratios. 50
Figure 4.9 Global probabilities of $\Pr(k_j < k_{j'} | \sigma_k \leq \sigma \leq \sigma_l)$ for permeability across each pair of fiber ratios 51
Figure 5.1 Représentation spatiale des covariables latentes pour trois spécimens dans un espace 2-D 60
Figure 5.2 Représentation spatiale des covariables latentes pour quatre spécimens dans un espace 3-D 61
Figure 5.3 Nouveau et quatrième spécimen échantillonnée à partir des trois spécimens testés 62
Figure 5.4 s 67
Figure 5.5 Distribution spatiale des neufs spécimens testés en fonction de la covariable taux de fibres 69

LISTE DES SIGLES ET ABRÉVIATIONS

GPR	Gaussian Process Regression
BFHP	Béton Fibré à Hautes Performances
BFUP	Béton Fibré à Ultra-Hautes Performances
LR	Linear Regression
MARS	Multivariate Adaptive Regression Splines
SR	Symbolic Regression
SVM	Support Vector Machine
NN	Neural Network
MC	Monte Carlo
McMC	Markov chain Monte Carlo
MLE	Maximum Likelihood Estimation
MAP	Maximum À Postériori
EP	Expectation Propagation
BRF	Béton Renforcé de Fibres
BO	Béton Ordinaire
BFO	Béton Fibré Ordinaire
$\mathcal{D}$	Jeu de données
y	Observations
x	Attributs
k	Coefficient de Perméabilité (m/s)
σ	Contrainte (MPa)
$\mathcal{N}$	Distribution Normale
$\mathbf{M}$	Vecteur Moyenne
$\mathbf{\Sigma}$	Matrice de Covariance

CHAPITRE 1 INTRODUCTION

1.1 Mise en contexte et problématique

Les essais en laboratoire nécessaires à l'avancée des recherches peuvent engendrer un coût économique non négligeable. Tel est le cas dans le domaine du génie civil, en effet manufacturer et tester des spécimens, de béton fibré par exemple, peut requérir un budget avoisinant les millers de dollars. Ainsi, seul un faible nombre de spécimens est souvent testé rendant la modélisation du comportement étudié lors de l'essai expérimental plus complexe, puisque la variabilité inter-spécimen devient difficile à capturer. Bien que les essais en laboratoire soient effectués sur des spécimens répliqués, il est en pratique impossible de fabriquer deux spécimens identiques en tout point. Soit l'exemple de deux spécimens de béton, les granulats n'auront pas tous la même taille et se placeront différemment dans le coffrage, engendrant ainsi des divergences dans les résultats d'essais expérimentaux entre les spécimens testés. Cette variabilité inter-spécimens provenant de la difficulté à reproduire deux spécimens à l'identique va ressortir dans les mesures tirées des essais en laboratoire. La variabilité inter-spécimens va alors varier avec le chargement appliqué lors de l'essai. La force de chargement est aussi appelée donnée d'entrée ou attribut et les mesures tirées des essais expérimentaux sont les observations qui présentent donc une variabilité dépendante de l'attribut, soit un comportement hétéroscédastique. Par définition, l'hétéroscédasticité est la non-stationnarité de la variance des observations en fonction d'un certain attribut (force de chargement par exemple).

La modélisation probabiliste par processus Gaussien est couramment utilisée dans le domaine de l'apprentissage machine (Murphy, 2012). Cette méthode de régression permet d'interpoler et d'extrapoler les données résultant des essais expérimentaux afin de fournir un modèle prédictif associé, pour des attributs cibles, aux valeurs moyennes et à leur matrice de covariance. La régression par processus Gaussien standard s'intéresse aux comportements homoscédastiques, lorsque la variabilité des données est uniforme suivant l'attribut. Quant à la modélisation de comportements hétéroscédastiques, elle nécessite des méthodes plus approfondies pouvant utiliser parfois deux processus Gaussiens imbriqués de manière hiérarchique (Goldberg et al., 1997).

Les méthodes de modélisation probabiliste existantes basées sur le processus Gaussien s'appuient habituellement sur une base de données composée de nombreux spécimens. Il en résulte alors de très bons modèles, puisque l'hypothèse commune de supposer les observations résultant des essais indépendantes entre elles est alors valide. En effet, les observations sont tirées

de nombreux spécimens distincts. Cependant, dans le cas d'une base de données pauvre en spécimens, l'hypothèse précédente ne peut plus s'appliquer puisqu'elle occulterait l'incertitude associée au faible nombre de spécimens testés. En effet, l'affiliation des observations aux spécimens ne peut plus être négligée, comme cela est le cas avec une base de données offrant un grand nombre d'observations tirées de spécimens variés.

Les méthodes de modélisation probabiliste connues aujourd'hui ne peuvent pas être appliquées au cas d'une base de données contenant de nombreuses observations tirées d'un faible nombre de spécimens. En effet, leur hypothèse principale d'indépendance des observations est dans cette situation erronée en effet, les observations tirées d'un même spécimen sont alors corrélées et sont dépendantes entre elles suivant l'attribut testé. Ainsi, ce projet de recherche s'intéressera à développer une méthode s'appuyant sur le processus Gaussien capable de modéliser des essais en laboratoire effectués sur peu de spécimens répliqués au comportement hétéroscédastique.

1.2 Cas d'étude

La méthode qui sera développée au sein de ce projet sera appliquée à un essai expérimental spécifique. Il s'agit de l'essai de perméabilité à l'eau sur les bétons fibrés à hautes performances (BFHP) conduit par Maxime Hubert lors de sa maîtrise recherche à Polytechnique Montréal (Hubert et al. (2015)). Lors de cet essai expérimental, neuf tirants fabriqués en béton à hautes performances (BHP) et en BFHP ont été testés, plus précisément trois spécimens associés à un des trois taux de fibres suivants : 0%, 0.75% et 1.5%. Ces essais de perméabilité consistent à appliquer un effort de traction croissant dans la barre d'armature au sein du tirant de béton et de faire passer l'eau à l'intérieur du spécimen pour en mesurer sa perméabilité à l'eau en fonction donc de la contrainte appliquée. Ces essais de perméabilité avaient pour objectif de déterminer l'impact de l'incorporation de fibres d'acier dans les bétons afin de réduire la perméabilité à l'eau et donc de prolonger la durabilité des structures en béton armé. Les résultats expérimentaux ont permis de mettre clairement en évidence le bénéfice de l'ajout de fibres vis-à-vis de la réduction de la perméabilité. Les fibres agissent comme des ponts, les contraintes passent dans les fibres, à travers les fissures. Les fissures dans les bétons renforcés de fibres sont plus nombreuses, mais plus étroites, rendant l'écoulement de l'eau plus difficile. En revanche, les fissures dans les bétons ordinaires bien que moins nombreuses sont plus larges et donc plus favorables au passage de l'eau dans le tirant de béton. Ces résultats quantitatifs ont mené à des conclusions qualitatives telles qu'une réduction des ouvertures de fissures et de la perméabilité à l'eau avec l'introduction de fibres dans le béton. Cependant, la représentativité des résultats obtenus sur un faible nombre de tirants testés

pourrait atténuer la robustesse des conclusions qui ont été tirées. La modélisation probabiliste devient ici pertinente puisqu'elle permet de comparer les trois taux de fibres testés dans le programme expérimental et de déterminer si les effets observés sont réellement significatifs en tenant compte de la variabilité des résultats.

La base de données étudiée dans le cadre de ce projet sera donc constituée de valeurs de contrainte en entrée (attribut) et de valeurs de perméabilité en sortie (observation). L'application de la méthode probabiliste développée au sein de ce projet permettra de mieux souligner les effets des fibres ajoutées aux bétons sur la perméabilité à l'eau et donc sur leur durabilité.

1.3 Objectifs

Les objectifs de ce projet de recherche se découpent en deux parties. Premièrement, cette recherche a pour but de développer une méthode de modélisation probabiliste applicable à tout type d'essai en laboratoire. Dans un second temps, il s'agira d'appliquer cette méthode à un cas concret soit aux essais de perméabilité conduits sur les BHP et BFHP. Les objectifs spécifiques de ce projet sont alors :

1. Développer une méthode de modélisation probabiliste d'essais en laboratoire par processus Gaussien avec peu de spécimens répliqués
   (a) Modéliser l'hétéroscédasticité (variabilité inter-spécimens dépendante de l'attribut)
   (b) Modéliser l'incertitude liée au faible nombre de spécimens
2. Appliquer la méthode développée aux essais de perméabilité sur les BHP et BFHP
   (a) Fournir des modèles probabilistes de la perméabilité
   (b) Quantifier l'efficacité des différents ajouts de fibres testés
3. Identifier les limites de la méthode par processus Gaussien et proposer une méthode complémentaire par approche Bayesienne permettant de contourner ces limites

1.4 Méthodologie

La méthode qui sera développée à travers cette étude s'appuiera sur le processus Gaussien couramment utilisé dans le cadre de la modélisation probabiliste. L'application du processus Gaussien sera étendue aux bases de données contenant peu de spécimens testés tout en prenant compte de l'hétéroscédasticité et de la dépendance des observations entre elles. Suite à

l'étude et à l'élaboration théoriques de la méthode, celle-ci sera exécutée sur les essais de perméabilité précédemment présentés. Les modèles probabilistes prédictifs obtenus pour chaque taux de fibres testé permettront par la suite d'estimer la probabilité d'obtenir une perméabilité à l'eau plus faible dans les BFHP que dans le BHP. Cette probabilité sera estimée à partir de la comparaison d'échantillons tirés des modèles prédictifs. Suite à cela, les limitations de la méthode par processus Gaussien seront examinées et une méthode complémentaire par approche Bayesienne sera alors proposée visant à contourner les limites rencontrées. Cependant, les travaux sur cette seconde approche sont encore en cours de développement en effet, les résultats de son application aux jeux de données tirés des essais de perméabilité à l'eau manque de consistence due à un manque de données. Aussi, l'élaboration méthode par approche Bayesienne permettra de proposer et de valider un nouveau concept de représentation spatiale des données pouvant être fort utiles pour des applications futures.

1.5 Organisation du mémoire

Le mémoire est divisé en cinq chapitres. Le premier et présent chapitre introduit le projet par sa mise en contexte et l'identification des objectifs à atteindre au cours de cette recherche. Le second chapitre est la revue de la documentation qui permettra de mettre en lumière les lacunes existantes dans les méthodes proposées quant à la modélisation probabiliste d'essais en laboratoire et ainsi d'établir la pertinence de ce projet. Cette revue s'attardera à la fois sur l'aspect théorique de la modélisation probabiliste d'essais en laboratoire et sur la durabilité des BFHP. Le troisième chapitre présentera la méthodologie suivie au cours de ce projet de recherche. Puis, le quatrième chapitre, sous la forme d'un article scientifique, couvrira l'élaboration de la méthode par processus Gaussien et les résultats de son application aux essais de perméabilité sur les BHP et BFHP. Il s'en suivra le cinquième chapitre qui présentera une discussion des approches et des résultats, tout en proposant également une méthode complémentaire par approche Bayesienne pour traiter la problématique du projet, une méthode plus lourde et complexe mais qui offrirait davantage de possibilités. Finalement le sixième chapitre clôturera ce projet avec les conclusions et les recommandations.

CHAPITRE 2 REVUE DE LA DOCUMENTATION

Ce chapitre dédié à la revue de documentation permettra d'expliciter quelques notions essentielles à la bonne compréhension du sujet de recherche. Aussi cette revue insérera ce projet au sein des travaux de recherche existants en comparant différentes méthodes et en mettant en évidence les lacunes de celles-ci. Ceci révélera la pertinence de la problématique étudiée.

La revue s'articulera sur trois thèmes, le premier étudiera les diverses méthodes de régression utilisées dans la modélisation d'essais expérimentaux, plus particulièrement dans le domaine du génie civil. Le second thème se concentrera sur l'extension des méthodes précédentes visant à modéliser les comportements hétéroscédastiques, une particularité souvent rencontrée dans les comportements étudiés lors d'essais en laboratoire. Ces deux premiers thèmes mettront en lumière qu'aucune méthode n'a été développée pour la modélisation d'essais sur peu de spécimens. C'est pourquoi cette recherche s'intéresse au développement d'une telle méthode. Finalement, le dernier thème de ce chapitre sera dédié à la mise en contexte des essais utilisés dans le cadre de ce projet pour l'application de la méthode de modélisation probabiliste. Cette dernière section s'intéressera donc au comportement en traction et à la durabilité des BHP et BFHP puis à l'essai de perméabilité à l'eau.

2.1 Méthodes de régression appliquées au génie civil

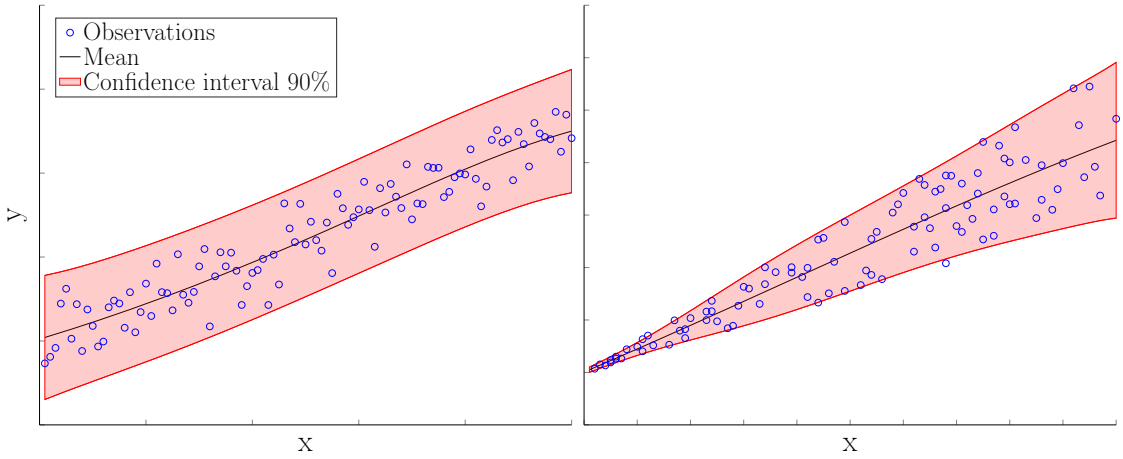
2.1.1 De la méthode simple à la plus complexe

Comme souligné dans l'introduction de ce mémoire, la modélisation d'essais en laboratoire est un point clef dans l'avancée de la recherche. En effet, les scientifiques comptent sur ces modélisations pour normaliser certains comportements, tel est le cas dans le domaine du génie civil où la modélisation d'essais expérimentaux sert à la conception de structures.

Il existe un grand nombre de méthodes de modélisation, ce projet se concentrera sur les modélisations probabilistes. Ces méthodes s'appuient sur les bases de données collectées lors d'essais en laboratoire réalisés pour prédire un comportement global et généralisable. Il s'agit dans ce cas de méthodes de régression, puisqu'à partir de données connues on prédit les résultats pour des attributs non étudiés.

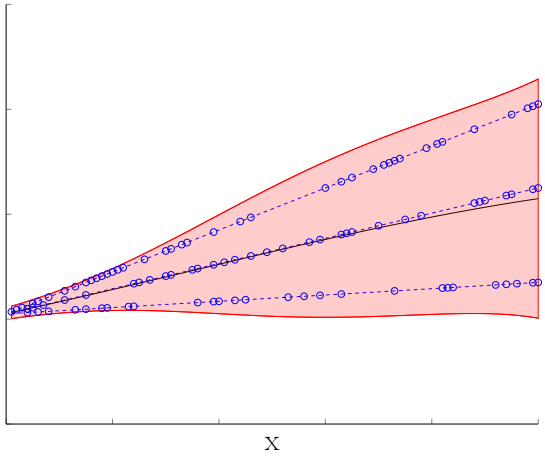
Cette partie s'intéresse aux différentes méthodes de régression applicables à un cas d'étude homoscédastique, aussi bien linéaire que non-linéaire. Il s'agit de montrer ici les avantages et limites des méthodes de régression lorsqu'utilisées pour la modélisation d'un comportement homoscédastique soit lorsque la variabilité entre les observations (données de sortie des essais)

est uniforme (Figure 4.1a).



(a) Homoscédasticité et observations indépendantes

(b) Hétéroscédasticité et observations indépendantes



(c) Hétéroscédasticité et observations dépendantes

Figure 2.1 Jeux de données présentant (a) un comportement homoscédastique, (b) un comportement hétéroscédastique and (c) un comportement hétéroscédastique avec observations dépendantes entre elles.

Les méthodes de modélisation telles que les méthodes de régression linéaire (LR), de régression multivariée par spline adaptative (MARS) et de régression symbolique (SR) s'appliquent aux cas homoscédastiques, linéaires et non linéaires (Jeon et al., 2014). Parmi ces trois méthodes, la méthode MARS se démarque, puisque celle-ci s'affranchit de l'utilisation de la connaissance à priori sur les observations en utilisant une définition formelle de spline et est aussi pertinente pour les problèmes de grande dimension. Aussi, l'avantage des méthodes LR, MARS et SR réside dans l'hypothèse de base qui considère les observations comme dé-

pendantes les unes des autres en fonction d'attributs indépendants (données d'entrée des essais). Cette dépendance des observations ne peut être occultée lorsque la base de données à modéliser est seulement composée de quelques spécimens testés.

D'autres méthodes proposent des modèles plus complexes pouvant alors capturer des comportements fortement non-linéaires. La machine à vecteurs de support (SVM) est une méthode de classification, en effet elle trie les observations suivant différentes classes de fonction (Siddique et al., 2008). Les classes de modèles ajoutent alors un degré de complexité dans la régression permettant une meilleure prédiction. La méthode de réseau de neurones (NN) aussi adaptée aux jeux de données complexes peut être associée à l'approche Bayesienne qui permet d'estimer les paramètres du modèle en prenant compte de l'incertitude sur les paramètres (Pal and Deswal, 2008; Lampinen and Vehtari, 2001; Ma et al., 2014). L'approche Bayesienne sera vue plus en détail dans le paragraphe suivant. Les méthodes SVM et NN donnent respectivement de très bons résultats quant à la prédiction de la résistance en compression de béton autoplaçant (Siddique et al. (2008)) et de la capacité d'une colonne en béton armé (Lampinen and Vehtari (2001)). Dans les deux exemples cités, la base de données s'étend à plus de 80 spécimens, cette large base de données est nécessaire à l'application des méthodes SVM et NN.

L'ensemble des méthodes citées, LR, MARS, SR, SVM, NN s'appuie sur des jeux de données riches en observations tirées de spécimens distincts. Sans cela, la modélisation est limitée par une sur-calibration des données. Cette limitation sur le jeu de données empêche l'application de ces méthodes aux jeux de données constitués de seulement quelques spécimens.

2.1.2 Approche Bayesienne

Il arrive régulièrement dans le domaine du génie civil que le comportement étudié en laboratoire soit connu et interprétable par une loi physique. La modélisation de l'essai consiste alors à inférer de façon probabiliste, les paramètres inconnus $\mathcal{P}_f$ de la loi physique. L'approche Bayesienne permet cela et tient compte des incertitudes aléatoire et épistémique (manque de connaissance) liées à cette modélisation (Der Kiureghian and Ditlevsen (2009)). Cette méthode s'est montrée performante dans la modélisation de la capacité et de la fragilité de colonnes en béton fibré ou dans la modélisation du module élastique du béton (Gardoni et al. (2002, 2007)). Comme son nom l'indique, l'approche Bayesienne s'appuie sur le théorème de Bayes (Box and Tiao (1992)). Soit une base de données récoltées suite à des essais expérimentaux, $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, incluant N observations $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ et N valeurs d'attribut $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$. À partir du théorème de Bayes appliqué au jeu de données $\mathcal{D}$, il est possible de déterminer la distribution à postériori des paramètres $\mathcal{P}_f$. Cette

distribution à postériori $p(\mathcal{P}_f|\mathcal{D})$ correspond à la connaissance des paramètres $\mathcal{P}_f$ en prenant compte des observations mesurées $\mathcal{D}$, elle est définie comme suit,

$$\overbrace{p(\mathcal{P}_f|\mathcal{D})}^{\text{À postériori}} = \frac{\overbrace{p(\mathcal{D}|\mathcal{P}_f)}^{\substack{\text{Fonction de} \\ \text{vraisemblance}}} \cdot \overbrace{p(\mathcal{P}_f)}^{\text{À priori}}}{\underbrace{p(\mathcal{D})}_{\substack{\text{Constante de} \\ \text{Normalisation}}}} \tag{2.1}$$

où la fonction de vraisemblance $p(\mathcal{D}|\mathcal{P}_f)$ représente l'information objective sur les paramètres inconnus $\mathcal{P}_f$ tirée du jeu de données $\mathcal{D}$ et $p(\mathcal{P}_f)$ est la distribution à priori de $\mathcal{P}_f$, il s'agit de la connaissance à priori des paramètres avant toute observation. La constante de normalisation est simplement un facteur de normalisation qui se décrit ainsi :

$$p(\mathcal{D}) = \int p(\mathcal{P}_f|\mathcal{D}) \cdot p(\mathcal{P}_f)\, d\mathcal{P}_f. \tag{2.2}$$

Une des difficultés réside dans la détermination de la fonction de vraisemblance, elle se définit en fonction du phénomène étudié, des informations disponibles et des observations. Quant à la définition de la distribution à priori des paramètres $p(\mathcal{P}_f)$, celle-ci par exemple, se base sur l'expérience. Si l'ingénieur ne dispose d'aucune connaissance à priori sur les paramètres physiques recherchés, il est possible d'utiliser des distributions à priori non informatives qui auront donc peu d'impact dans la formulation de Bayes (Equation 2.1). Si les paramètres inconnus sont des paramètres de location (paramètres constants suivant les valeurs d'attribut), leur distribution non informative à priori sera uniforme avec $p(\mathcal{P}_f) = 1$. Si ces paramètres sont qualifiés de paramètres d'échelle (paramètres variant avec l'attribut), alors $p(\mathcal{P}_f) = \dfrac{1}{\mathcal{P}_f}$ (Box and Tiao (1992)).

Connaître les distributions à postériori des paramètres décrivant les modèles physiques déterministes permet d'améliorer ces modèles à l'aide de termes correctifs. Ces nouveaux modèles peuvent alors être caractérisés de probabilistes, puisque s'appuient sur l'approche Bayesienne et prennent en compte toutes les incertitudes liées aux résultats de laboratoire et rendent désormais mieux compte de la réalité (Gardoni et al. (2002, 2007)).

Cette approche Bayesienne, qui ressort comme étant la méthode d'identification de paramètres la plus fiable, a été utilisée par Sloński (2010; 2011) à l'intérieur de la méthode de réseau de neurones (NN). L'approche Bayesienne permet d'inférer les distributions à postériori des paramètres à partir du jeu de données étudié $\mathcal{D}$. Ainsi, dans la méthode NN, les paramètres $\mathcal{P}_f$ ne sont plus représentés comme un simple vecteur de valeurs, mais comme des variables aléatoires. Le but de la méthode NN est ici de créer un modèle prédictif à

partir d'observations $\mathbf{y}$ seulement. Contrairement à Gardoni et al. (2002, 2007) qui partait d'un modèle physique déterministe pour l'améliorer en fonction des observations obtenues, Słoński (2010) cherche à modéliser à l'aide de la méthode NN les propriétés du béton ou la résistance en compression du BFHP à partir d'observations seulement (Sloński (2011)).

Soit le jeu de données $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, incluant N observations indépendantes $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ et N valeurs d'attribut $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$. Les observations sont définies comme une valeur exacte t à laquelle s'ajoute une erreur de mesure v, $y = t + v$. Pour un vecteur de valeurs d'attribut $\mathbf{x}_*$ non mesurées en laboratoire, la méthode NN fournit le modèle probabiliste prédictif $p(\tilde{\mathbf{t}}|\mathcal{P}_f, \mathbf{x}_*, \mathcal{D})$ conditionné sur les paramètres $\mathcal{P}_f$. Les distributions à postériori de ces paramètres $p(\mathcal{P}_f|\mathcal{D})$ sont inférées par approche Bayesienne. Alors les distributions prédictives des observations $\tilde{\mathbf{t}}$ pour $\mathbf{x}_*$ sont calculées à partir de l'intégration suivante :

$$p(\tilde{\mathbf{t}}|\mathbf{x}_*, \mathcal{D}) = \int p(\tilde{\mathbf{t}}|\mathcal{P}_f, \mathbf{x}_*, \mathcal{D}) \cdot p(\mathcal{P}_f|\mathcal{D})\, d\mathcal{P}_f \tag{2.3}$$

Or, cette intégration est dans la plupart des cas analytiquement intraitable, c'est pourquoi il est recommandé de faire appel à des méthodes d'approximation. La méthode d'échantillonnage Markov chain Monte Carlo (McMC) est souvent appliquée afin d'échantillonner les paramètres $\mathcal{P}_f$ à partir de leurs distributions à postériori $p(\mathcal{P}_f|\mathcal{D})$, et ainsi d'obtenir les échantillons $\mathcal{P}_{f,q} : \mathcal{P}_f \sim p(\mathcal{P}_f|\mathcal{D})$. La méthode Monte Carlo (MC) permet d'approximer l'intégrale présentée à l'Equation 2.3 par,

$$p(\tilde{\mathbf{t}}|\mathbf{x}_*, \mathcal{D}) \approx \sum_{q}^{Q} p(\tilde{\mathbf{t}}|\mathbf{x}_*, \mathcal{P}_{f,q}) \tag{2.4}$$

En prenant compte des incertitudes liées aux paramètres du modèle probabiliste, l'approche Bayesienne se montre adaptée aux cas de jeux de données limités en observations. Cependant, la connaissance à priori prend une grande importance dans cette approche puisqu'elle va compenser le manque de données disponibles, ainsi l'approche Bayesienne dépend fortement des distributions à priori des paramètres $p(\mathcal{P}_f)$. De plus, les méthodes d'échantillonnage telle que la méthode McMC sont très coûteuses en temps. Bien que Słoński (2010) reconnaisse les fortes aptitudes de la méthode McMC, il recommande l'utilisation de la méthode de régression par processus Gaussien (GPR : Gaussian Process Regression) associée à l'estimation par maximum de vraisemblance (MLE : Maximum Likelihood Estimation). Les méthodes GPR et MLE seront détaillées dans la sous-section suivante.

2.1.3 Régression par processus Gaussien et estimation par maximum de vraisemblance

La méthode de régression par processus Gaussien est une méthode rapide et efficace de modélisation probabiliste. Elle a permis d'estimer la résistance à la compression du BFHP (Hoang et al. (2016)) ou de modéliser la capacité d'une colonne (P.Mahesh and Deswal (2010)). P.Mahesh and Deswal (2010) précisent que pour leur cas d'étude, la méthode GPR est plus performante que la méthode SVM. Plusieurs travaux (Hoang et al. (2016), P.Mahesh and Deswal (2010), Słoński (2010)) suggèrent d'utiliser la méthode GPR pour la modélisation d'essais en laboratoire.

Soit un jeu de données $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$ avec N observations $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ supposées indépendantes les unes des autres pour N valeurs d'attribut $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$. La force de la méthode GPR réside dans sa capacité à interpoler et extrapoler, ceci en quantifiant les incertitudes liées à chaque prédiction. La méthode GPR définit les observations $\mathbf{y}$ comme étant une fonction de l'attribut $\mathbf{x}$ à laquelle s'ajoute l'erreur d'observation qui peut être l'erreur de mesure dans le cas d'observations en laboratoire (Equation 2.5). La méthode GPR suppose que les observations $\mathbf{y}$ sont décrites par une distribution Gausienne multivariée $\mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$, caractérisée par le vecteur colonne $\mathbf{M}$ et la matrice de covariance $\mathbf{\Sigma}$ (MacKay (1998); Rasmussen and Williams (2006); Ebden (2008)).

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{f(\mathbf{x})}_{\text{réalité}} + \underbrace{\mathbf{v}}_{\text{erreurs d'observations}} \qquad \text{avec} \quad \mathbf{v} : V \sim \mathcal{N}(0, \sigma_v^2) \tag{2.5}$$

Ces dernières matrices $\mathbf{M}$ et $\mathbf{\Sigma}$ décrivent la structure à priori définissant les observations, c'est pourquoi le choix de cette structure est primordial afin que celle-ci s'adapte au mieux au comportement testé. En général, il est commun de commencer la modélisation avec $\mathbf{M} = \mathbf{0}$ et la fonction de covariance exponentielle carré, telle que :

$$g(x_i, x_j) = \sigma_f^2 \exp\left[\frac{-(x_i - x_j)^2}{2\ell^2}\right] + \sigma_v^2 \delta(x_i, x_j) \tag{2.6}$$

où σ_f^2 est la variance à priori dans le modèle, σ_v^2 est la variance correspondant au bruit de mesure et ℓ est la longueur de corrélation qui décrit l'influence d'un attribut x_i sur un autre attribut x_j. La Figure 2.2 schématise cette influence, plus la longueur de corrélation ℓ est grande, plus la corrélation entre les attributs x_i et x_j sera importante.

L'estimation de ces hyper-paramètres (paramètres de la structure à priori du processus Gaussien) $\mathcal{P}_f = \{\sigma_f, \ell, \sigma_v\}$ se fait par estimation par maximum de vraisemblance (MLE) qui sera

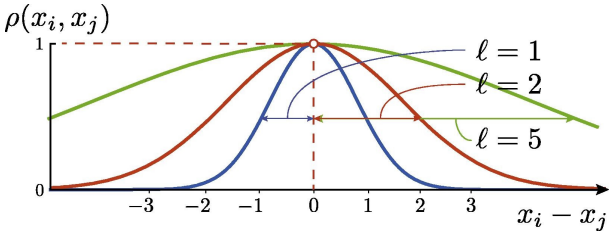


Figure 2.2 Exemple de fonctions de covariance exponentielle carré unidimensionnelles pour différentes longueurs de corrélation ℓ (tiré de Quach et al. (2017))

étudiée à la suite de la méthode GPR. Finalement la matrice de covariance $\mathbf{\Sigma}$ se construit à partir de l'équation 2.6, il est à noter que le bruit de mesure modélisé par l'écart-type σ_v apparaît seulement sur la diagonale de la matrice $\mathbf{\Sigma}$. En effet, l'erreur d'observation n'entre en jeu que pour deux mêmes attributs puisque les observations sont présumées indépendantes. Aussi, la matrice de covariance doit être définie semi-positive pour la suite des calculs,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_f^2 + \sigma_v^2 & g(x_1, x_2) & \cdots & g(x_1, x_N) \\ & \sigma_f^2 + \sigma_v^2 & \cdots & g(x_2, x_N) \\ & & \ddots & \cdots \\ \text{Sym.} & & & \sigma_f^2 + \sigma_v^2 \end{bmatrix}. \tag{2.7}$$

Une fois la structure à priori construite, la méthode GPR permet de prédire à partir du jeu de données $\mathcal{D}$ des valeurs non observées $\mathbf{f} = [f(x_{1*}), f(x_{2*}), \ldots, f(x_{P*})]^\intercal$ sous des valeurs d'attributs cibles $\mathbf{x}_* = [x_{1*}, x_{2*}, ..., x_{P*}]^\intercal$. En concaténant les covariances des valeurs observées et non observées, la distribution Gaussienne multivariée de prédiction devient :

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{M}\\ \mathbf{M}_*\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma} & \mathbf{\Sigma}_*^\intercal\\ \mathbf{\Sigma}_* & \mathbf{\Sigma}_{**}\end{bmatrix}\right), \quad \text{avec} \tag{2.8}$$

$$\mathbf{\Sigma}_* = \begin{bmatrix} g(x_{1*},x_1) & g(x_{1*},x_2) & \cdots & g(x_{1*},x_N)\\ & g(x_{2*},x_2) & \cdots & g(x_{2*},x_N)\\ & & \ddots & \cdots\\ \text{Sym.} & & & g(x_{P*},x_N)\end{bmatrix} \tag{2.9}$$

$$\mathbf{\Sigma}_{**} = \begin{bmatrix} \sigma_f^2 & g(x_{1*},x_{2*}) & \cdots & g(x_{1*},x_{P*})\\ & \sigma_f^2 & \cdots & g(x_{2*},x_{P*})\\ & & \ddots & \cdots\\ \text{Sym.} & & & \sigma_f^2\end{bmatrix} \tag{2.10}$$

Finalement, les observations prédites $\mathbf{f}$ sont décrites par une distribution Gaussienne multivariée avec :

$$\mathbb{E}[\mathbf{f}] = \mathbf{M}_* + \mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{M}), \tag{2.11}$$

$$\text{cov}(\mathbf{f}) = \mathbf{\Sigma}_{**} - \mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*{}^\intercal, \tag{2.12}$$

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbb{E}[\mathbf{f}], \text{cov}(\mathbf{f})) \tag{2.13}$$

Comme explicité plus haut, la pertinence de la prédiction par la méthode GPR repose sur le choix de la fonction de covariance et ainsi sur l'estimation des paramètres qui la décrivent. A partir de l'Equation 2.6, l'ensemble d'hyper-paramètres à estimer est $\mathcal{P}_f = \{\sigma_f, \ell, \sigma_v\}$. L'estimation de paramètres est possible par approche Bayesienne, mais il est plus courant d'associer au GPR la méthode MLE (Słoński (2010); Ebden (2008)). La méthode d'estimation MLE s'appuie, comme l'approche Bayesienne, sur la formule de Bayes présentée à l'Equation 2.1.

On peut supposer que la distribution à priori des hyper-paramètres $p(\mathcal{P}_f)$ est constante, alors la distribution à postériori $p(\mathcal{P}_f|\mathcal{D})$ atteint son maximum lorsque la fonction de vraisemblance $p(\mathcal{D}|\mathcal{P}_f)$ atteint elle-même son maximum pour $\mathcal{P}_f^*$. Sachant que la variance de la distribution

à postériori s'approche de zéro autant que le jeu de données $\mathcal{D}$ s'étend, il devient raisonnable d'estimer les hyper-paramètres à partir du maximum de vraisemblance. Cependant, la fonction de vraisemblance tend vers zéro à mesure que le jeu de données grandit à cause du nombre de termes augmentant dans la matrice de covariance. Une solution simple est de maximiser le log de la fonction de vraisemblance afin que les faibles probabilités s'ajoutent au lieu de se multiplier entre elles. L'estimation des paramètres par MLE devient :

$$\mathcal{P}_f^* = \arg\max_{\mathcal{P}_f} \{\log(p(\mathcal{P}_f|\mathcal{D}))\} \tag{2.14}$$

La méthode GPR est capable de modéliser des comportements fortement non-linéaires en fournissant les valeurs moyennes et la matrice de covariance des prédictions. Cependant, comme le montre l'Equation 2.6, les prédictions dépendent des distances entre les valeurs à prédire et les observations de $\mathcal{D}$, c'est pourquoi la méthode GPR compte aussi sur une base de données riche en observations pour fournir un modèle probabiliste prédictif judicieux. Pour des régressions simples, la méthode GPR a été implémentée dans la toolbox GPML disponible sur MATLAB (Rasmussen and Nickisch (2010)).

L'ensemble des méthodes : LR, MARS, SR, SVM, NN, GPR s'appliquent à des larges jeux de données présentant un comportement homoscédastique, c'est-à-dire présentant une variabilité uniforme entre les observations (Figure 4.1a) . Dans le domaine du génie civil, il arrive régulièrement que les résultats d'essais en laboratoire montrent un comportement hétéroscédastique (Figure 4.1b-c). En effet, pour des essais effectués sur des spécimens de BFHP par exemple, il y a une forte probabilité que les résultats varient selon les spécimens. Ceci s'explique par l'impossibilité de produire deux spécimens identiques qui peuvent varier notamment en fonction de la taille et de la répartition des granulats ou de la répartition et de l'orientation des fibres. Ces critères sont difficiles à contrôler ou à mesurer et leur variabilité se retrouve dans la variabilité inter-spécimens qui est dépendante de l'attribut lors de l'essai, soit l'hétéroscédasticité. La prochaine Section 2.2 décrira l'adaptation des méthodes précédentes à cette variabilité conditionnelle à l'attribut.

2.2 Hétéroscédasticité

2.2.1 Différentes approches de l'erreur de mesure

L'hétéroscédasticité peut être décrite comme l'erreur de mesure qui varie en fonction des valeurs d'attribut $\mathbf{x}$ et qui, additionnée à l'observation exacte, va engendrer une variabilité entre les observations dépendante de l'attribut.

La modélisation par simple régression linéaire (LR) offre une solution face aux comportements hétéroscédastiques (Bansal and Aggarwal (2007)). Cette solution s'appuie sur la définition même de l'hétéroscédasticité, étant la variabilité entre les observations évoluant en fonction de l'attribut. Il s'agit alors de définir la variance de l'observation comme fonction de l'attribut. Soit un jeu de données $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$ contenant N observations indépendantes $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ pour N valeurs d'attribut $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$. Il s'agit de la même définition de l'observation que dans la méthode GPR, mais cette fois l'observation est une fonction linéaire de l'attribut :

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{\beta\mathbf{x}}_{\text{réalité}} + \underbrace{\mathbf{v}}_{\text{erreur d'observation}} \qquad \text{avec} \quad \mathbf{v} : V \sim \mathcal{N}(0, \sigma_v^2) \tag{2.15}$$

Par hypothèse, l'observation y_i suit la distribution $y_i \sim \mathcal{N}(\beta x_i, \sigma_v^2)$ et Bansal and Aggarwal (2007) proposent une nouvelle définition du bruit de mesure qui dépendrait de l'attribut et donc changerait à chaque valeur d'attribut, soit :

$$\sigma_{v,i}^2 = \mathbb{E}[y_i]^2 = \beta^2 x_i^2, \quad i = 1, 2, \ldots, N \tag{2.16}$$

$$y_i \sim \mathcal{N}(\beta x_i, \beta^2 x_i^2), \quad i = 1, 2, \ldots, N, \quad \beta \in \mathbb{R} \tag{2.17}$$

Le coefficient β peut ensuite être estimé par approche Bayesienne. Cette méthode est donc fonctionnelle pour les comportements hétéroscédastiques.

Concernant les comportements non-linéaires plus complexes et hétéroscédastiques, la méthode NN n'utilise pas de paramètres d'erreur de mesure dépendant de l'attribut, mais va chercher à adapter la distribution de modélisation afin de prendre en compte l'hétéroscédasticité (Yeh, 2014). Yeh (2014) conclut que la distribution Log-Normale est plus adéquate que la distribution Normale pour modéliser la résistance à la compression dans le BFHP ainsi que la variabilité de l'erreur.

L'approche Bayesienne a été présentée à travers un modèle déterministe physique afin d'inférer les paramètres du modèle. Il est possible d'appliquer la même approche sur un jeu de données présentant un comportement non-linéaire hétéroscédastique en faisant du bruit de mesure une fonction de l'attribut. Aussi, cette méthode est globale, car elle s'applique aussi aux cas homoscédastiques grâce à l'ajout du paramètre d'hétéroscédasticité dans la formulation du bruit de mesure (Blau et al., 2008). Ce paramètre en prenant la valeur zéro permet de rendre le bruit de mesure constant dans les cas homoscédastiques, sinon il permet de faire varier le bruit en fonction de l'attribut. L'approche Bayesienne permet alors à la fois d'inférer

les paramètres du modèle physique étudié et les paramètres statistiques. Cette méthode est applicable lorsque le modèle physique est connu, ce qui n'est pas toujours le cas.

L'étude des méthodes de modélisation de comportements homoscédastiques a permis de souligner les forces de GPR. Pour la suite, la méthode GPR sera adaptée aux cas hétéroscédastiques.

2.2.2 Régression par processus Gaussien et hétéroscédasticité

Goldberg et al. (1997) proposent d'utiliser une approche hiérarchique de la méthode GPR telle que la variance de l'erreur de mesure est elle-même modélisée par un processus Gaussien. La nouvelle description des observations $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ indépendantes, indexées par $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$ devient donc :

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{f(\mathbf{x})}_{\text{réalité}} + \underbrace{\mathbf{v}}_{\text{erreur d'observation}} \qquad \text{avec} \quad \mathbf{v} : V \sim \mathcal{N}(0, r(\mathbf{x})) \tag{2.18}$$

En supposant une moyenne à priori nulle, les observations prédites $\mathbf{f}$ pour le vecteur cible $\mathbf{x}_*$, sont décrites par :

$$\mathbb{E}[\mathbf{f}] = \mathbf{\Sigma}_*(\mathbf{\Sigma} + \mathbf{\Sigma}_V)^{-1}\mathbf{y}, \quad \text{avec} \quad \mathbf{\Sigma}_V = \text{diag}(r(\mathbf{x})) \tag{2.19}$$

$$\text{cov}(\mathbf{f}) = \mathbf{\Sigma}_{**} + r(\mathbf{x}_*) - \mathbf{\Sigma}_*(\mathbf{\Sigma} + \mathbf{\Sigma}_V)^{-1}\mathbf{\Sigma}_*^\intercal, \tag{2.20}$$

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbb{E}[\mathbf{f}], \text{cov}(\mathbf{f})) \tag{2.21}$$

La difficulté de cette méthode réside une nouvelle fois dans l'inférence des hyper-paramètres, ici l'hyper-paramètre additionnel à inférer est la variance de l'erreur de mesure pour chaque attribut x_i. La fonction de vraisemblance n'étant plus analytiquement exprimable, la méthode MLE n'est plus applicable. La méthode d'inférence de paramètres par maximum à postériori (MAP) comparable à la méthode MLE ne peut pas être utilisée non plus (Kersting et al., 2007; Le et al., 2005) et d'autres méthodes d'inférence sont alors nécessaires. Goldberg et al. (1997) appuient leur étude sur l'échantillonnage de Gibbs qui est une méthode McMC, donc une méthode lourde. Grâce à la formule de Bayes, les échantillons $\mathbf{r} = [r(x_1), r(x_2), ..., r(x_N)]^\intercal$ sont tirés de la distribution à postériori de la variance de l'erreur de mesure, $p(\mathbf{r}|\mathbf{y})$. Face à cette méthode d'échantillonnage lente et coûteuse, plusieurs chercheurs se sont penchés sur diverses approches afin de la contourner. Aussi, Titsias and Lázaro-Gredilla (2011) suggèrent d'utiliser en remplacement de la fonction de vraisemblance, puisque non connue, sa borne inférieure

variant avec l'attribut. En comparaison avec la méthode McMC, cette dernière approche est plus rapide, seulement deux fois le temps que prend l'application de GPR standard à un cas homoscédastique. Une autre méthode d'approximation de la fonction de vraisemblance et donc de la distribution à postériori des hyper-paramètres est l'Expectation Propagation (EP) (Tolvanen et al. (2014)). Cette méthode est plus générale et sophistiquée, puisqu'elle propose d'inférer la variance du processus Gaussien σ_f (Equation 2.6) comme une fonction de l'attribut et ainsi modéliser la non-stationnarité du comportement étudié. La méthode EP est moins coûteuse en programmation, rapide et aussi performante que la méthode McMC dans le cas d'étude présenté par Tolvanen et al. (2014), aussi elle a déjà été implémentée dans la toolbox GPstuff disponible sur MATLAB (Vanhatalo et al. (2012)). L'ensemble des dernières méthodes s'appuie sur un même modèle d'observation décrit à l'Equation 2.18 qui propose une erreur de mesure dépendante de l'attribut $\mathbf{x}$. Cette hypothèse suggère que l'hétéroscédasticité proviendrait de l'erreur de mesure uniquement, mais l'hétéroscédasticité peut être expliquée par la non-maîtrise de tous les paramètres entrant en jeu dans la fabrication d'un spécimen, puisque non mesurés voire non connus. Cette approche de l'hétéroscédasticité est considérée par Wang and Neal (2012); Wang (2014) qui proposent une reformulation de la fonction de covariance.

$$g(x_i, x_j, w_i, w_j) = \sigma_f^2 \exp\left[-\frac{(x_i - x_j)^2}{2\ell^2} - \frac{(w_i - w_j)^2}{2\ell_w^2}\right] + \sigma_v^2 \delta(x_i, x_j) \tag{2.22}$$

Dans cette Equation 2.22, la covariable additionnelle $\mathbf{w} = [w_1, w_2, ..., w_N]^\intercal$ est une variable latente donc non connue, cachée. Elle représente tous ces paramètres non mesurés lors de l'essai en laboratoire et qui ont une influence sur le comportement à modéliser. Déterminer les valeurs de la covariable cachée $\mathbf{w}$ requiert une nouvelle fois une inférence par échantillonnage McMC de la distribution à postériori de $\mathbf{w}$.

Finalement, les différentes modélisations de l'hétéroscédasticité (Equations 2.18 et 2.22) présentent la même difficulté qui réside comme dans le cas homoscédastique dans la détermination des hyper-paramètres. Pour rappel, il existe diverses méthodes pour inférer les paramètres, mais leur robustesse repose sur un jeu de données $\mathcal{D}$ riche (MLE) ou demande une programmation lourde (approche Bayesienne et McMC). Pour la modélisation d'un jeu de données limité à quelques spécimens et présentant un comportement homoscédastique (Thiyagarajan and Kodagoda, 2016) ou hétéroscédastique (Figure 4.1c), la dépendance des observations entre elles ne peut pas être occultée. L'application de la méthode GPR hétéroscédastique considère les observations comme indépendantes les unes des autres et ainsi ne prend pas en compte l'incertitude due au faible nombre de spécimens ayant pour conséquence une sur-calibration du modèle. L'hypothèse d'indépendance des observations n'est pas ap-

plicable dans le cas d'un jeu de données riche en observations tirées de seulement quelques spécimens (Figure 4.1c). Les méthodes existantes supposant les observations indépendantes puisque tirées de nombreux spécimens distincts, ne sont donc pas adéquates pour le problème posé. Il s'agit alors d'élaborer une méthode de modélisation probabiliste capable d'associer hétéroscédasticité et dépendance des observations. Il est à retenir cependant que les méthodes basées sur la méthode GPR sont performantes et peu coûteuses en temps ou en programmation. L'inférence des hyper-paramètres peut s'effectuer par approche Bayesienne qui est la méthode la plus adéquate pour prendre en compte toutes les incertitudes, mais cette méthode reste très lourde. Aussi, pour des jeux de données riches en observations prélevées sur quelques spécimens, on préférera la méthode MLE qui est applicable lorsqu'un grand nombre d'observations est disponible.

2.3 BFHP : Béton Fibré à Hautes Performances

Les bétons renforcés de fibres (BRF) présentent des propriétés améliorées par rapport aux bétons ordinaires sur de multiples aspects, tels que le comportement en traction, la résistance à la fissuration, la ductilité et la perméabilité. Les BRF permettent d'améliorer le comportement en traction post-fissuration, de réduire les armatures passives dans les structures et de réduire l'ouverture des fissures pour améliorer la durabilité. Les nombreux avantages des BRF ont conduit les scientifiques et ingénieurs à s'y intéresser davantage et d'en accroître leurs capacités, c'est pourquoi depuis trente ans, les BRF sont grandement étudiés et ont évolué du béton ordinaire (BO) au béton fibré à ultra hautes performances (BFUP) en passant par le béton fibré à hautes performances (BFHP). Cette section porte d'abord sur les BHP et BFHP puisque le jeu de données étudié dans le projet regroupe des résultats d'essais de perméabilité sur des spécimens de BHP et de BFHP. Ensuite, la durabilité des bétons via la mesure de la perméabilité sera discutée.

2.3.1 Formulation du BFHP

Le béton à hautes performances (BHP) est un BO auquel ont été incorporés des ajouts minéraux qui possèdent la plupart du temps une réactivité pouzzolanique et ont un effet filler de remplissage du squelette granulaire (Charron et al. (2016)). Les ajouts courants au Québec sont la fumée de silice, la cendre volante, le laitier et le filler calcaire. Aussi, l'addition d'un superplastifiant permet de réduire le rapport Eau/Liant qui varie alors de 0.30 à 0.40, tout en gardant une bonne maniabilité du béton. Cette formulation du béton fournit une matrice de densité élevée et une résistance à la compression à 28 jours pouvant atteindre 90 MPa. Le BFHP est un BHP auquel a été ajoutée une quantité de fibres variant de 0.5% à

1.5% du volume de béton. Les fibres les plus utilisées pour obtenir un renforcement mécanique sont les fibres d'acier qui possèdent un module élastique plus élevé. Il en existe diverses sortes telles que synthétique (acrylique, nylon), naturelle (bagasse, noix de coco) ou encore en verre. Les fibres doivent avoir la propriété de se déformer et idéalement de ne pas casser dans le béton (perte de renforcement). La formulation spécifique des BRF et donc du BFHP se fait sous la méthode Baron-Lesage qui suppose que le béton le plus maniable est le plus compact. Pour une géométrie, des dimensions et une quantité de fibres fixées (ici des macrofibres pour un BFHP), il suffit de faire varier le rapport de quantités sable/granulat du mélange pour en déterminer le rapport optimal donnant la meilleure maniabilité. Cette dernière est estimée par un essai de maniabilimètre pour les BRF courants et par test d'étalement (cône d'Abrams) pour les BRF autoplaçants. Finalement, les caractéristiques générales du BHP et du BFHP sont résumées dans le Tableau 2.1.

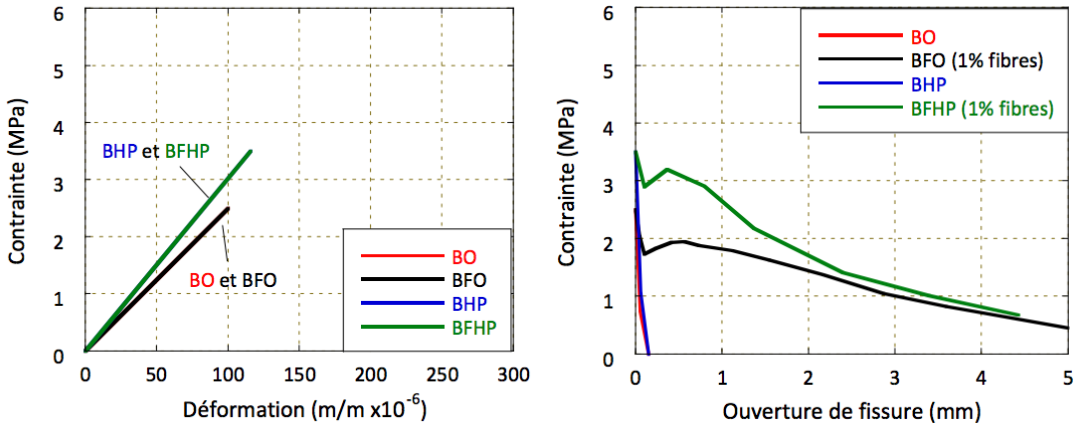
Tableau 2.1 Caractéristiques du BHP et du BFHP (adapté de Charron et al. (2016))

Composition	**BHP**	**BFHP**	**BFUP**
Rapport Eau/Liant	0.30 à 0.40	0.30 à 0.40	0.15 à 0.25
Liant	350 à 450 kg/m^3	350 à 450 kg/m^3	800 à 1000 kg/m^3
Ajouts minéraux	oui	oui	oui
Sable	oui	oui	oui
Pierre	oui	oui	non
Fibres	non	40 à 120 kg/m^3	160 à 480 kg/m^3
Propriétés	**BHP**	**BFHP**	**BFUP**
Compression - f'_c	50 à 90 MPa	50 à 90 MPa	120 à 180 MPa
Traction - f'_t	3.0 à 4.5 MPa	3.0 à 4.5 MPa	7 à 15 MPa
Module - E_c	25 à 35 GPa	25 à 35 GPa	30 à 45 GPa

2.3.2 Comportement en traction

L'un des avantages du BFHP est son comportement ductile en traction. L'ajout de macrofibres n'a que peu d'impact sur le comportement en compression, en revanche il influence celui en traction. Comme le Tableau 2.1 l'indique, la résistance en traction f'_t du BFHP est identique à celle du BHP, cependant la rupture après le pic est plus ductile comme le montre la Figure 2.3. Avant le pic de résistance soit au comportement pré-pic, le béton ordinaire (BO), le béton fibré ordinaire (BFO), le BHP et le BFHP présentent un comportement similaire qui est très court et avec peu de déformations. C'est une fois la macrofissure créée, soit au pic de résistance, que les macrofibres entrent en jeu et que les comportements des bétons diffèrent. Le comportement des bétons sans fibre, BO et BHP est très fragile, on observe une

pente abrupte à la suite du pic de résistance. En revanche, les bétons renforcés de fibres de 1%, BFO et BFHP présentent un comportement ductile en traction.

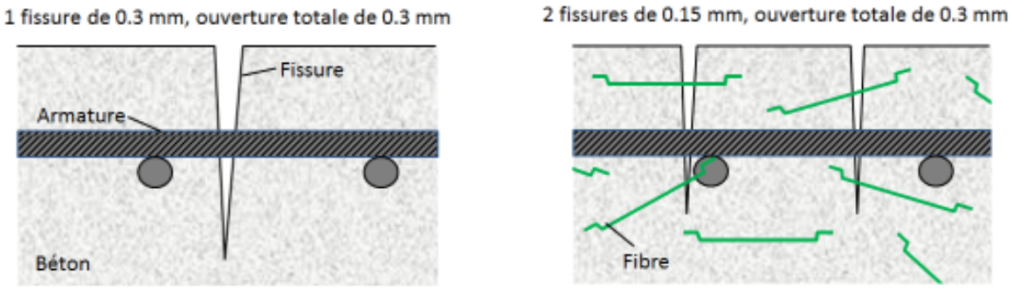


(a) Comportement pré-pic (b) Comportement post-pic

Figure 2.3 Comportement en traction uniaxiale des bétons (tiré de Charron et al. (2016))

La Figure 2.4 schématise l'action des fibres dans le béton. Les macrofibres n'influent pas sur la résistance maximale en traction, mais sur le comportement post-pic des bétons. La ductilité observée à la Figure 2.3 provient de l'action de fermeture des fibres sur les fissures. Les fibres agissent comme des ponts : ancrées de part et d'autre des fissures, elles vont permettre aux contraintes de passer à travers des fissures et ainsi procurer une résistance résiduelle à la traction (ductilité) tout en limitant la propagation des fissures. Sous une même charge, des sections équivalentes de bétons BO et BHP présentent typiquement une macrofissure d'ouverture de 0.3 mm, alors que les bétons BFO et BFHP présenteront plutôt deux macrofissures de 0.15 mm chacune (Figure 2.4). Les macrofissures dans les BRF sont donc plus fines et plus nombreuses.

L'orientation des fibres est le facteur prépondérant quant à l'efficacité de l'action des fibres plutôt que la quantité de fibres ajoutée au béton. Il est très difficile de contrôler l'orientation des fibres lors de la mise en place de BRF, cependant il est possible de favoriser l'orientation des fibres perpendiculairement aux fissures pour accroître leur action de fermeture (Martinie and Roussel (2011); Plagué et al. (2017)). Dans le cas des BFHP, la macrofissuration est localisée et le rôle des macrofibres est de coudre les macrofissures et ainsi de retarder la rupture. Pour cela, le pourcentage de macrofibres doit être faible et les fibres doivent être longues à gros diamètre permettant l'augmentation de la capacité portante et de la ductilité en flexion et à l'effort tranchant. L'orientation des fibres dépend de nombreux paramètres tels que la direction de l'écoulement (Plagué et al. (2017)), l'effet de parois (rapport entre



(a) Fissuration avec un BO ou un BHP (b) Fissuration avec un BFO ou un BFHP

Figure 2.4 Patron de fissuration dans différents bétons pour une ouverture de fissure totale de 0.3 mm (tiré de Charron et al. (2016))

les dimensions de la structure et celles de la fibre) (Rossi (1998)), la technique de mise en place (on préférera la vibration externe si requise pour éviter de générer des zones sans fibre dans le béton) et la présence des armatures et des chaises de coffrage qui peuvent dévier les fibres. Difficile à contrôler, l'orientation des fibres joue un rôle d'autant plus important dans la perméabilité des BRF, puisqu'une bonne orientation implique des fissures plus fines et donc une perméabilité plus faible (Plagué et al. (2017)).

2.3.3 Durabilité du béton et perméabilité à l'eau

La durabilité des bétons dépend de la pénétration à l'eau et des agents agressifs dans la porosité et dans les fissures. La pénétration de l'eau se fait principalement par 3 mécanismes : la diffusion et l'adsorption par la porosité entre les fissures et par la perméabilité dans les fissures. Si les fissures demeurent ouvertes, c'est ce dernier mécanisme qui devient prépondérant pour la durabilité du béton environnant une fissure. La durabilité du béton armé en condition de service et fissuré est donc significativement influencée par la perméabilité à l'eau. C'est dans ce contexte que les travaux de Maxime Hubert ont été réalisés (Hubert et al., 2015) pour l'étude de la perméabilité à l'eau de différentes gammes de béton (BHP, BFHP et BFUP) en condition de service. La présente étude étant basée sur les travaux de Hubert et al. pour l'application d'une nouvelle méthode probabiliste, la revue de cette section portera essentiellement sur la perméabilité des bétons.

La perméabilité à l'eau se définit comme la pénétration de l'eau, soumise à un gradient de pression, à travers le béton. Elle peut être très néfaste si importante puisqu'elle laisse pénétrer l'eau qui transporte des agents agressifs tels que les chlorures engendrant la dégradation de la structure. En fait, à l'état fissuré et plus précisément au voisinage d'une fissure, le mode de transport prédominant des agents agressifs est la perméabilité (Charron et al. (2016)). La

perméabilité à l'eau dépend directement de l'ouverture des fissures (Banthia and Bhargava (2007)), celle-ci est proportionnelle au cube de l'ouverture de fissure (Hubert et al. (2015)), donc $k \propto w^3$.

L'étude du comportement en traction des BFHP a démontré précédemment que l'ajout de macrofibres dans le béton permet de ralentir la propagation des macrofissures. En effet, les macrofibres se comportant comme des ponts pour la diffusion des contraintes permettent la création de fissures plus nombreuses mais plus fines. Ceci, en opposition aux BO qui présentent sous un même cas de charge des macrofissures moins nombreuses mais plus larges. La Figure 2.5 présente les patrons de fissurations pour les cas de tirants en BO et en BRF et montre que l'addition de fibres diminue l'ouverture des fissures qui sont alors plus diffuses.

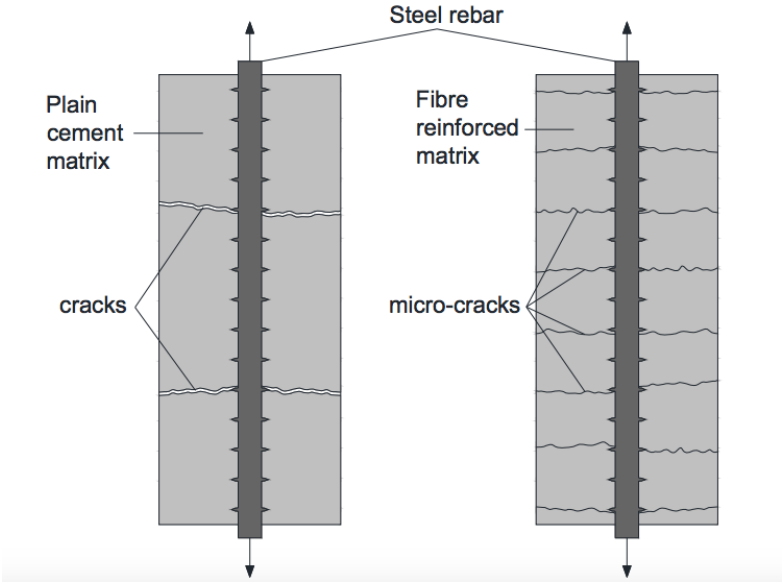


Figure 2.5 Patrons de fissuration des tirants de BO et de BRF (tiré de Berrocal et al. (2013))

L'addition de fibres permet de réduire la perméabilité à l'eau de façon significative. Rapoport et al. (2002) concluent de leurs essais de perméabilité que celle-ci est réduite avec un ajout de fibres de 0.5% du volume par rapport à un BO, et est encore plus réduite avec un ajout de 1%. Ceci concorde avec les résultats de Desmettre and Charron (2012) qui indiquent que la perméabilité décroît de 60-70% dans un tirant en BFHP par rapport à un tirant en BO, tous deux soumis à un même cas de charge. Ces résultats positifs confirment le potentiel des BFR dans l'amélioration de la durabilité des structures en béton armé.

Ainsi, le présent projet portera sur la modélisation probabiliste de résultats d'essais de perméabilité. La sous-section suivante décrira l'essai de perméabilité développé à l'Ecole Poly-

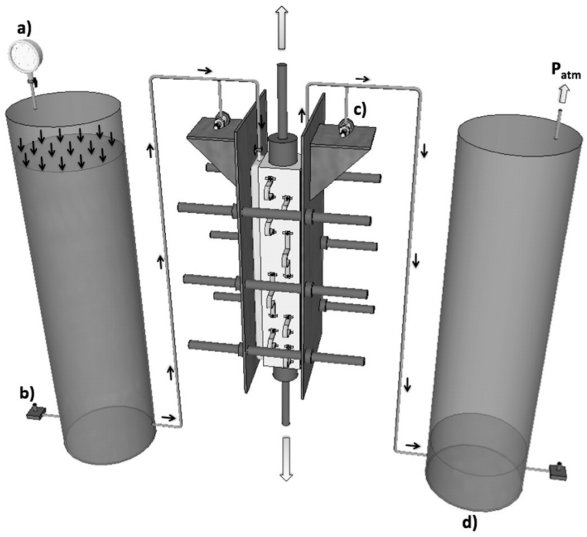
technique de Montréal, les spécimens testés et les résultats obtenus.

2.3.4 Essai de perméabilité sur des tirants en BFHP

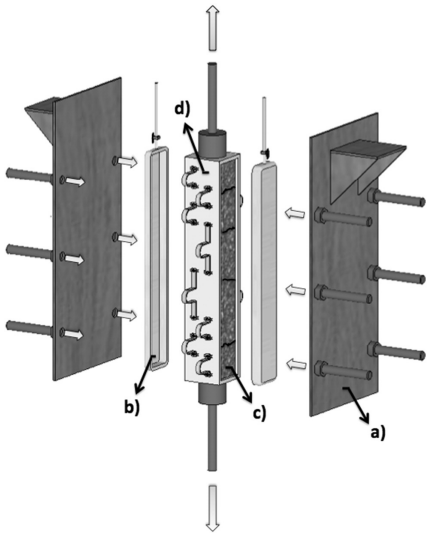
Le dispositif d'essai de perméabilité est schématisé à la Figure 2.6a. Il a été développé par Desmettre and Charron (2012), réutilisé par Hubert et al. (2015) pour caractériser l'influence de la quantité de fibres sur la perméabilité et réutilisé par Plagué et al. (2017) afin d'étudier l'impact de l'orientation des fibres sur la perméabilité. Dans cet essai, le tirant de béton est soumis à un chargement en traction uniaxial quasi-statique induit par un déplacement axial imposé à un taux de 0.05 mm/min grâce à une presse hydraulique de capacité de 2.5 MN. Une vue éclatée de la cellule de perméabilité (Figure 2.6c) met en évidence les deux surfaces de béton exposées à la pénétration de l'eau. Les autres faces sont recouvertes d'une membrane étanche permettant un écoulement unidirectionnel de l'eau. A l'état initial de l'essai, le tirant de béton pourvu d'une barre d'armature est saturé en eau et un système de serrage permet d'éviter les fuites. Pendant l'essai, le spécimen est sollicité en traction et la pression appliquée au cylindre d'entrée associée à la pression atmosphérique dans le cylindre de sortie, vide initialement, permet de créer un gradient de pression et ainsi déplacer l'eau à travers le tirant. Des capteurs de hauteur différentielle installés à la base des cylindres mesurent le volume d'eau traversant le tirant et donc les débits d'eau entrant et sortant qui mènent à la mesure de la perméabilité. Des capteurs PI sont aussi installés le long des faces longitudinales du tirant (Figure 2.6a), ceux-ci mesurent l'ouverture des fissures et l'allongement total du tirant apparaissant au fur et à mesure de l'augmentation de la force appliquée. Ce dispositif permet donc de mesurer l'évolution de la perméabilité et de l'ouverture des fissures dans un spécimen de béton soumis à un effort de traction croissant. De plus, il est possible d'évaluer la contrainte moyenne dans l'armature du tirant à tout moment.

Ce projet est basé sur les résultats d'essai de Hubert et al. (2015). Plus précisément, l'objectif est de modéliser la perméabilité des BHP et des BFHP pour les deux taux de fibres 0.75% et 1.5% en fonction des contraintes appliquées au long de l'essai de perméabilité présenté ci-dessus. Trois conditions de spécimens différentes seront considérées, la variable étant la quantité de fibres. Les trois cas sont identifiés dans le Tableau 2.2. Chaque spécimen mesure 610 mm de long et a une section carrée de 90×90 mm^2. Il contient une barre d'armature 10M avec un enrobage de 40 mm. Pour chaque condition de volume de fibres, 0%, 0.75% et 1.5%, trois spécimens répliqués ont été testés. Le Tableau 2.3 décrit la composition des 3 bétons étudiés. Une augmentation du dosage en superplastifiant et en viscosant a été requise pour le BFHP - 1.5% afin de garder une bonne maniabilité.

Ainsi, neuf spécimens ont été testés sous l'essai de perméabilité, soient trois spécimens répli-



(a) Dispositif de chargement et de perméabilité. a) Système de serrage, b) Réservoir d'eau en aluminium, c) Manomètre et cylindre d'entrée, d) Capteur de hauteur différentielle, e) Capteur de pression, f) Cylindre de sortie



(b) Cellule de perméabilité. a) Système de serrage, b) Réservoir d'eau en aluminium, c) Béton exposé, d) Membrane étanche et capteurs PI

Figure 2.6 Instrumentation de l'essai de perméabilité (tiré de Plagué et al. (2017))

Tableau 2.2 Programme expérimental (adapté de Hubert et al. (2015))

identification	**Béton**	**Taux de fibres (%)**	**Taux d'armature ρ (%)**
BHP - 0% - 10 M	BHP - 50 MPa	0	1.23
BFHP - 0.75% - 10 M	BFHP - 50 MPa	0.75	1.23
BFHP - 1.5% - 10 M	BFHP - 50 MPa	1.5	1.23

Tableau 2.3 Composition des bétons (adapté de Hubert et al. (2015))

Matériau	**BHP**	**BFHP-0.75%**	**BFHP-1.5%**
Ciment (kg/m^3)	500	500	500
Fumée de silice (kg/m^3)	50	50	50
Eau (kg/m^3)	237	237	237
Superplastifiant (l/m^3)	9.20	9.20	45.9
Viscosant (l/m^3)	0.70	0.70	3.55
Sable (kg/m^3)	814	814	823
Pierre (kg/m^3)	678	658	593
Dosage en fibre (kg/m^3)	0	58.9	117
Rapport Eau/Liant	0.43	0.43	0.43

qués par taux de fibres 0%, 0.75% et 1.5%. Les résultats sont représentés dans la Figure 2.7, les courbes représentent les valeurs moyennes de perméabilité en fonction de la contrainte

appliquée, auxquelles sont associées les valeurs minimales et maximales de perméabilité mesurées pour les trois spécimens. Le graphique montre les résultats pour les BFHP et pour un béton fibré à ultra-hautes performances (BFUP ou UHPFRC : ultra high performance fiber reinforced concrete) qui ne sera pas étudié dans ce projet, puisque l'objectif est de comparer l'efficacité de l'addition de fibres sur la perméabilité dans les BFHP. Le coefficient de perméabilité reste stable et inférieur à 1×10^{-10} m/s jusqu'à atteindre la résistance en traction du béton où les premières fissures apparaissent. À charge équivalente et à l'état fissuré, la perméabilité diminue avec l'augmentation du taux de fibres. A l'ultime, les tirants présentent trois, six et dix macrofissures pour les pourcentages de fibres 0%, 0.75% et 1.5% respectivement. Ces résultats confirment l'effet "pont" des fibres pour le transfert des contraintes dans le tirant, ce qui améliore le contrôle de la fissuration. Finalement, augmenter le pourcentage de fibres dans les spécimens de BFHP de 0% à 0.75% et de 0.75% à 1.5% permet de réduire la perméabilité de 31% et de 92% respectivement.

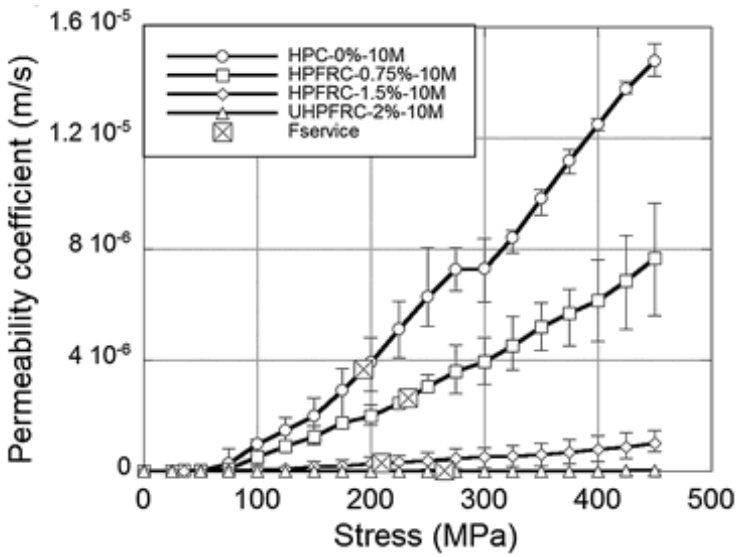


Figure 2.7 Perméabilité en fonction de la contrainte dans l'armature pour quatre dosages de fibres (tiré de Hubert et al. (2015))

L'interprétation des résultats des essais de perméabilité sur le BFHP a permis de mettre en évidence l'avantage de l'addition de fibres sur la réduction de la perméabilité à l'eau mais aucun modèle de comportement de la perméabilité n'a été fourni, limitant alors la portée de ces dernières conclusions. En effet la modélisation et rendue difficile face au faible nombre de spécimens testés, soit trois par taux de fibres, et à la variabilité des observations dépendante des valeurs de contrainte (hétéroscédasticité), représentée par les barres d'erreur à la Figure

2.7. La modélisation probabiliste de ces essais permettrait de prédire le comportement d'un quatrième spécimen non testé à partir des données brutes des trois spécimens testés. De plus, la modélisation de chaque taux de fibres permettrait de les comparer et d'estimer la probabilité d'obtenir une perméabilité plus faible avec un certain pourcentage de fibres par rapport à un autre.

2.4 Conclusion

Cette revue de la documentation a permis de mettre en évidence le manque de méthodes de modélisation de données tirées de quelques spécimens au comportement hétéroscédastique. Les méthodes existantes sont fiables lorsque les observations peuvent être considérées indépendantes les une des autres, que leur affiliation au spécimen n'a pas d'impact. Or cette hypothèse n'est valable que pour des jeux de données formés d'observations tirées de spécimens testés distincts. La modélisation d'un jeu de données contenant peu de spécimens au comportement hétéroscédastique doit être capable de prendre en compte la dépendance des observations aux spécimens dont elles sont tirées afin de modéliser l'incertitude liée à leur faible nombre. Il s'agit donc d'élaborer une méthode de modélisation probabiliste qui tient compte à la fois de la dépendance des observations et de l'hétéroscédasticité.

La deuxième partie de la revue centrée sur les BFHP a permis d'introduire les essais de perméabilité en laboratoire choisis pour l'application de la méthode. La durabilité des structures en BFHP, à l'état fissuré, dépend de la perméabilité à l'eau. Les essais de perméabilité ont montré l'avantage de l'incorporation de fibres puisque la perméabilité est réduite significativement du BHP au BFHP-1.5% pour tous les niveaux de contraintes de traction. Cependant, aucun modèle prédictif de perméabilité en fonction de la contrainte de traction et de la teneur en fibres n'a été tiré des données résultantes des essais et seule la réduction de perméabilité a été quantifiée entre les essais. La modélisation probabiliste vise donc à confirmer les tendances observées de façon plus générale en prenant compte de l'incertitude liée au faible nombre de spécimens testés. L'application de la méthode probabiliste a pour objectif de quantifier le bénéfice de l'ajout de fibres dans le BFHP en comparant les modèles prédictifs de perméabilité pour les différents taux de fibres testés.

CHAPITRE 3 MÉTHODOLOGIE

Ce chapitre présentera une vue d'ensemble de la méthodologie suivie dans ce projet de recherche. Elle permettra avant tout de faire le lien entre les résultats d'essais en laboratoire et le modèle probabiliste numérique développé au Chapitre 4.

3.1 Essai expérimental à l'étude

Le comportement physique étudié et à modéliser est celui de la perméabilité à l'eau dans des tirants en BHP et en BFHP soumis à un chargement en traction. Les résultats d'essais disponibles pour la modélisation ont été tirés des essais de perméabilité décrits à la Section 2.3.4 et conduits par Hubert et al. (2015).

3.1.1 Jeu de données

Comme précisé à la Section 2.3.4, neuf spécimens de BHP et BFHP ont été testés sous l'essai de perméabilité (Tableau 2.2) soient trois spécimens de BHP, trois spécimens de BFHP à 0.75% de fibres et trois spécimens à 1.5% de fibres. La Figure 3.1 regroupe les résultats d'essais, soit pour chaque spécimen testé, le coefficient de perméabilité $\mathbf{k}$ exprimé comme une vitesse en m/s en fonction de la contrainte moyenne $\boldsymbol{\sigma}$ appliquée dans la barre d'armature au sein du spécimen de béton, elle est exprimée en MPa. La Figure 3.2 rend compte des mesures de chaque cas, soit chaque taux de fibres testé pour trois spécimens répliqués. L'intervalle de contraintes choisi pour l'étude soit, [150-450] MPa, représente les conditions en service et à l'ultime des infrastructures en béton armé. Pour chaque taux de fibres testé, les observations présentent un comportement hétéroscédastique en effet la variabilité entre les observations varie suivant la contrainte. Deux tendances principales se détachent de cette représentation graphique des résultats. De manière générale, la perméabilité augmente avec la contrainte appliquée, ce qui est cohérent puisque l'augmentation de la contrainte engendre l'apparition de macrofissures et l'ouverture de celles-ci. Mais ce qui est le plus intéressant dans ces essais est l'impact des fibres sur la perméabilité. En effet, plus grand est le taux de fibres incorporé dans le BFHP, plus la réduction de la perméabilité est importante. Il semble donc que l'incorporation de fibres ait un impact positif de réduction sur la perméabilité et permet donc la prolongation de la durabilité du BFHP. C'est ce comportement et ce bénéfice qui seront quantifiés à travers la modélisation probabiliste. Aussi, chaque jeu de données montre certaines particularités qu'il est nécessaire de modéliser afin de représenter fidèlement le com-

portement physique observé. À 275 MPa, les données du BHP présentent un décrochage dans la perméabilité. Ce phénomène s'explique par l'apparition d'une macrofissure qui va venir contrôler la fissuration. L'ouverture soudain de la fissure requiert un certain temps d'adaptation du dispositif de chargement de l'essai qui impose un déplacement constant, l'énergie libérée par l'ouverture de la fissure est repris par le dispositif. Suite à cette macrofissure, il est possible d'observer que les courbes se rejoignent à les contraintes plus élevées. En effet, le béton est devenu fortement perméable suite à cette macrofissure. Pour le BFHP-0.75%, les données se regroupent à 230-240 MPa, en effet lors des essais les spécimens ont montré chacun une ouverture de fissures totale sur quatre fissures similaire engendrant une faible variabilité à ce niveau de contraintes. La méthode de modélisation doit donc être capable de rendre compte de ces diverses particularités observées.

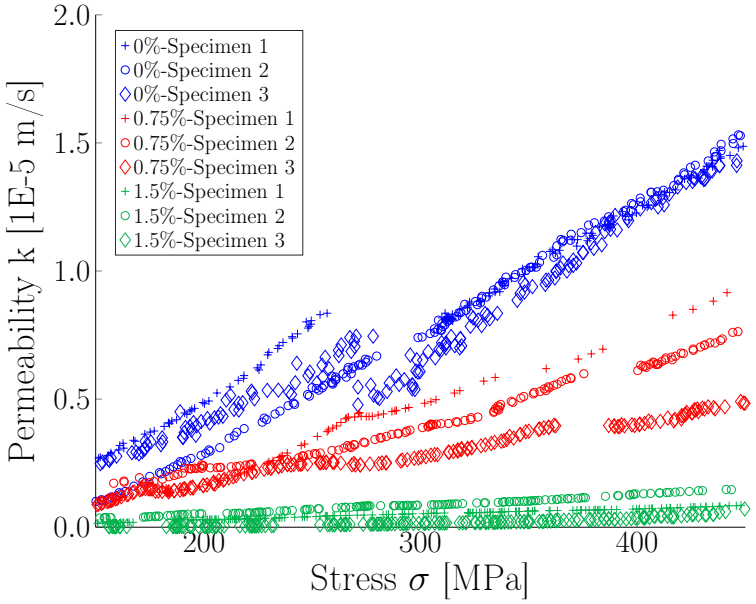
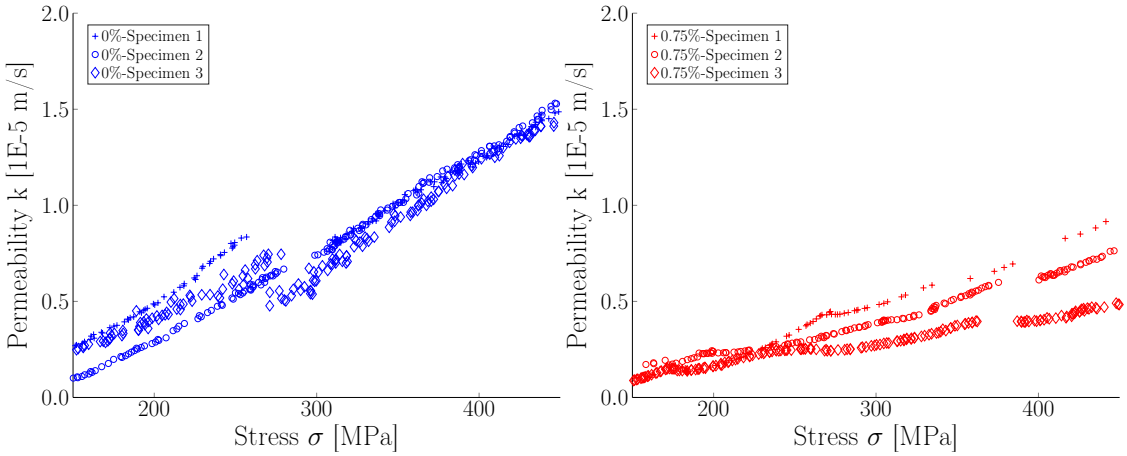


Figure 3.1 Représentation graphique du jeu de données résultant des essais de perméabilité sur neuf spécimens de BHP et BFHP

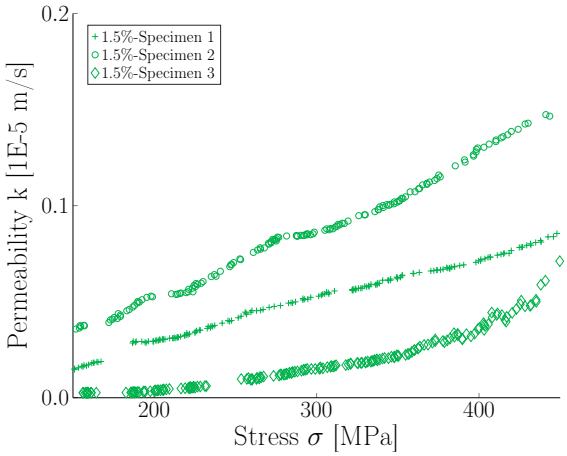
3.1.2 Objectifs de la modélisation probabiliste

La modélisation probabiliste des résultats d'essais présentés à la Figure 3.1 a tout d'abord pour objectif de fournir, pour chaque taux de fibres testé, un modèle prédictif de la perméabilité en fonction de la contrainte. Ce modèle prédictif peut être interprété comme la prédiction du comportement d'un quatrième spécimen non testé, ce comportement lui-même modélisé par une moyenne et un intervalle de confiance représentatif de l'incertitude liée à cette prédiction. Le défi à relever dans cette modélisation est d'établir une méthodologie capable de



(a) Taux de fibres 0%

(b) Taux de fibres 0.75%



(c) Taux de fibres 1.5%

Figure 3.2 Représentations graphiques par taux de fibres des jeux de données résultant des essais de perméabilité sur trois spécimens de BHP ou BFHP

modéliser à la fois l'hétéroscédasticité des données ou encore la variabilité inter-spécimens dépendante de la contrainte et l'incertitude liée au faible nombre de spécimens, au nombre de trois, constituant le jeu de données. La superposition et comparaison des trois modèles prédictifs, soit un modèle par taux de fibres, 0%,0.75%,1.5%, permettra de quantifier le bénéfice de l'incorporation de fibres dans le béton.

3.2 Modélisation : de l'expérimental au numérique

Dans cette section, le lien entre l'expérimental et le numérique est fait en explicitant l'utilisation des données expérimentales au sein du modèle probabiliste numérique.

3.2.1 Interprétation probabiliste des données

Le jeu de données en sortie des essais de perméabilité, pour un taux de fibres testé, est constitué des données mesurées sur trois spécimens répliqués. Le jeu de données $\mathcal{D}$ est constitué de N points de mesure de perméabilité pris sur les trois spécimens répliqués d'un même taux de fibres. Chaque point est une observation de la perméabilité en fonction de l'attribut contrainte appliquée. Or, toute observation n'est pas une mesure exacte et comprend une erreur de mesure. L'observation y_i de la perméabilité est interprétée comme la valeur exacte la perméabilité k_i, elle-même fonction de la contrainte σ_i, à laquelle vient s'ajouter une erreur de mesure v. Ainsi, le jeu de données $\mathcal{D}$ comprend le vecteur de contraintes en entrée $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_N]$ et le vecteur d'observations de perméabilité en sortie $\mathbf{y} = [y_1, \ldots, y_N]$ et s'écrit alors $\mathcal{D} = \{(\sigma_i, y_i), i = 1, \ldots, N\}$ avec,

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{\mathbf{k}}_{\text{réalité}} + \underbrace{\mathbf{v}}_{\text{erreur d'observation}} \tag{3.1}$$

$$= f(\boldsymbol{\sigma}) + \mathbf{v}. \tag{3.2}$$

3.2.2 Deux approches probabilistes

Au cours de cette recherche deux approches différentes ont été étudiées. La première méthode par processus Gaussien sera détaillée à travers un article scientifique présenté au Chapitre 4. Cette première méthode par processus Gaussien se révèlera être performante pour répondre aux objectifs de ce projet, mais ne présentera que peu d'ouvertures pour une extension aux cas multi-attributs. Aussi, une seconde approche sera présentée dans le Chapitre 5, une approche

plus lourde en temps et programmation mais offrant plus de perspectives. Cependant, cette seconde méthode par approche Bayesienne est toujours en développement et n'a pu fournir de résultats pertinents quant au jeu de données tiré des essais de perméabilité.

3.3 Méthode par processus Gaussien

3.3.1 Description

La méthodologie à développer afin de modéliser des essais en laboratoire effectués sur peu de spécimens répliqués doit tenir compte de la variabilité inter-spécimens (hétéroscédasticité) et de la dépendance des observations aux spécimens, cette dernière pouvant être interprétée comme l'incertitude liée au faible nombre de spécimens. Pour répondre à ces critères, la première méthode présentée au Chapitre 4 peut être décomposée en trois étapes. La première étape consiste à prédire le comportement de la perméabilité pour chacun des trois spécimens pour toute contrainte au sein d'un même processus Gaussien. Connaissant alors trois points d'observation pour chaque contrainte, la deuxième étape permet de modéliser à la fois l'hétéroscédasticité et l'incertitude sur le nombre de spécimens à l'aide d'une distribution à priori conjuguée définie à partir de ces trois points d'observation. Finalement, la prédiction de la perméabilité d'un quatrième spécimen pour une contrainte donnée, se fait à partir de la distribution prédictive déduite de la distribution à postériori des trois points précédents.

3.3.2 Résultats

La modélisation probabiliste des résultats d'essais de perméabilité fournit un modèle prédictif pour chaque taux de fibres testé permettant ainsi de prédire dans chaque cas le comportement d'un quatrième spécimen. La comparaison deux à deux de ces trois modèles permet d'estimer la probabilité d'obtenir une perméabilité plus faible avec un taux de fibres par rapport à un autre. Ceci quantifie alors l'apport de l'incorporation de fibres dans le béton et finalement permettra de souligner le réel potentiel de durabilité du BFHP par rapport au BHP. Les résultats détaillés sont présentés au Chapitre 4 sous forme d'article.

3.4 Méthode par approche Bayesienne

La méthode par processus Gaussien s'est révélée performante et a fourni des résultats concluants. Cependant, sa capacité à élargir sa prédiction à des taux de fibres non testés est moindre. Ainsi, une seconde méthode par approche Bayesienne a été développée offrant une autre perspective.

3.4.1 Description

Cette seconde méthode propose une approche originale de l'hétéroscédasticité. En effet, les spécimens testés sont représentés par des coordonnées spatiales se répartissent dans un espace multi-dimensionnel. L'hétéroscédasticité est alors représentée par les distances séparant les spécimens : plus cette distance virtuelle est grande, plus importante est la variabilité. Les coordonnées virtuelles des spécimens permettant de calculer les distances sont alors des nouveaux hyper-paramètres à inférer à travers une formulation de Bayes hiérarchique. Cette approche Bayesienne offre la possibilité de prendre en compte toutes les incertitudes liées aux paramètres, mais requiert une implémentation plus lourde, puisqu'elle nécessite de faire appel à la méthode d'échantillonnage Markov chain Monte Carlo (McMC).

3.4.2 Résultats

Le jeu de données étudié au sein de ce projet ne présente que trois spécimens par taux de fibres et ne permet donc pas de présenter de résultats concluants par l'application de cette seconde méthode. Cependant, la représentation spatiale et virtuelle des spécimens offre la possibilité d'ajouter le taux de fibres comme un second attribut dans la modélisation et ainsi prédire le comportement d'un taux de fibres non testé. Les limitations et les perspectives de la méthode par approche Bayesienne seront présentées dans le Chapitre 5.

CHAPITRE 4 ARTICLE 1 : PROBABILISTIC MODELING OF HETEROSCEDASTIC LABORATORY EXPERIMENTS USING GAUSSIAN PROCESS REGRESSION

Lucie Tabor, James-A. Goulet, Jean-Philippe Charron, Clélia Desmettre

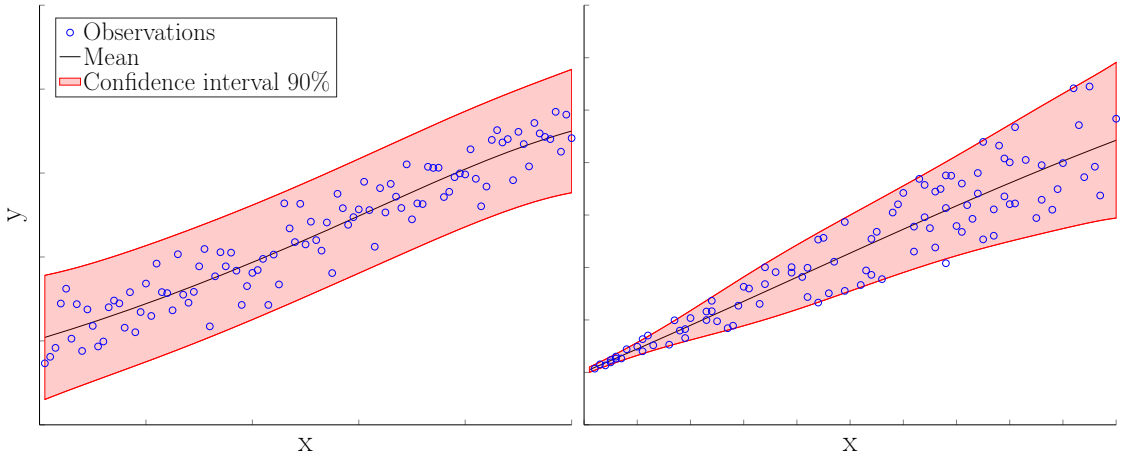
Article submitted to *Journal of engineering mechanics*

4.1 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 in 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.

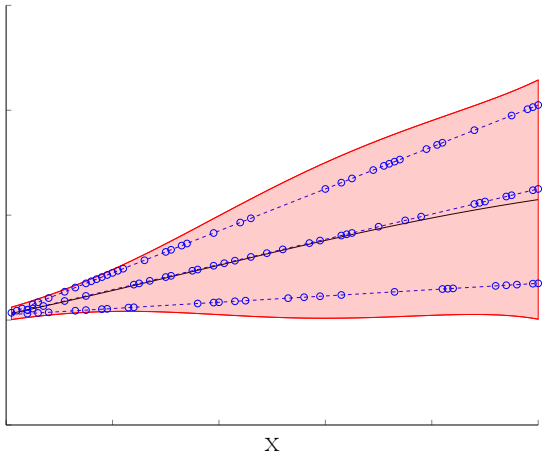
4.2 Introduction

Modelling the variability in the results of laboratory experiments is difficult when only few specimens are available for the study. 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 4.1a & 4.1b respectively show an example of a homoscedastic and of a heteroscedastic behaviour. For Figure 4.1a, the observations variability is independent of the covariate x, which is not the case for Figure 4.1b. 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 several observations for different covariates, are obtained from the same specimen as displayed in Figure 4.1c.



(a) Homoscedasticity and independent observations

(b) Heteroscedasticity and independent observations



(c) Heteroscedasticity and dependent observations

Figure 4.1 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), however they are not best suited to model a heteroscedastic behaviour. Sloński (2011) 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. A similar method has been extended to heteroscedastic behaviours 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 modelisation challenges is Gaussian Process regression (GPR) (Rasmussen and Williams, 2006; MacKay, 1998). Słoński (2010) 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. GPR can handle highly nonlinear sets of data. Simple regression problems such as the one presented in Figure 4.1a can readily be processed with the GPR method as implemented 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 (Wang and Neal, 2012; Wang, 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 as demonstrated by Kersting et al. (2007), Le et al. (2005) and Titsias and Lázaro-Gredilla (2011).

The shaded region in Figures 4.1a-c describe 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. 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. Existing methods are currently not able to model the response from observations obtained from only few specimens 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. 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.

4.3 Gaussian Process Regression

4.3.1 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 attribute $\mathbf{x} = [x_1, x_2, ..., x_N]^\intercal$, Gaussian Process Regression can predict an unobserved value, given a new attribute value x_*. The vector $\mathbf{x}$ can be extended to a matrix of several attributes. 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 attributes $\mathbf{x}$ and are described by a multivariate Gaussian distribution $\mathbf{y} : Y \sim \mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$ which includes observations errors $\mathbf{v}$ (Equation 4.1), characterized by a mean column vector $\mathbf{M}$ and a covariance matrix $\mathbf{\Sigma}$.

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{f(\mathbf{x})}_{\text{reality}} + \underbrace{\mathbf{v}}_{\text{observations errors}}, \qquad \text{with} \quad \mathbf{v} : V \sim \mathcal{N}(0, \sigma_v^2) \tag{4.1}$$

These matrices describing the prior structure 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,

$$g(x_k, x_l) = \sigma_f^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell^2}\right] + \sigma_v^2 \delta(x_k, x_l) \tag{4.2}$$

where σ_f^2 is the process noise variance and ℓ is the correlation length which defines the influence of one attribute x_k on another attribute x_l. The longer is the correlation length, the higher will be the correlation between an attribute 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.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_f^2 + \sigma_v^2 & g(x_1, x_2) & \cdots & g(x_1, x_N) \\ & \sigma_f^2 + \sigma_v^2 & \cdots & g(x_2, x_N) \\ & & \ddots & \cdots \\ \text{Sym.} & & & \sigma_f^2 + \sigma_v^2 \end{bmatrix} \tag{4.3}$$

GPR can estimate unobserved values $\mathbf{f} = [f(x_{1*}), f(x_{2*}), \ldots, f(x_{P*})]^\intercal$ given the target attribute 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 so the multivariate Gaussian distribution becomes,

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{M}\\ \mathbf{M}_*\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma} & \mathbf{\Sigma}_*{}^\intercal\\ \mathbf{\Sigma}_* & \mathbf{\Sigma}_{**}\end{bmatrix}\right), \quad \text{with} \tag{4.4}$$

$$\mathbf{\Sigma}_* = \begin{bmatrix} g(x_{1*},x_1) & g(x_{1*},x_2) & \cdots & g(x_{1*},x_N)\\ & g(x_{2*},x_2) & \cdots & g(x_{2*},x_N)\\ & & \ddots & \cdots\\ \text{Sym.} & & & g(x_{P*},x_N)\end{bmatrix} \tag{4.5}$$

$$\mathbf{\Sigma}_{**} = \begin{bmatrix} \sigma_f^2 & g(x_{1*},x_{2*}) & \cdots & g(x_{1*},x_{P*})\\ & \sigma_f^2 & \cdots & g(x_{2*},x_{P*})\\ & & \ddots & \cdots\\ \text{Sym.} & & & \sigma_f^2\end{bmatrix} \tag{4.6}$$

The estimated values $\mathbf{f}$ are described by a multivariate Gaussian distribution,

$$\mathbb{E}[\mathbf{f}] = \mathbf{M}_* + \mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{M}), \tag{4.7}$$

$$\text{cov}(\mathbf{f}) = \mathbf{\Sigma}_{**} - \mathbf{\Sigma}_*\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_*{}^\intercal, \tag{4.8}$$

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbb{E}[\mathbf{f}], \text{cov}(\mathbf{f})) \tag{4.9}$$

4.3.2 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 $\mathcal{P}_f = \{\sigma_f, \ell, \sigma_v\}$. According to Bayes theorem, the posterior probability for hyper-parameters values is given by :

$$p(\mathcal{P}_f|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{P}_f)\cdot p(\mathcal{P}_f)}{p(\mathcal{D})} \tag{4.10}$$

$$\propto p(\mathcal{D}|\mathcal{P}_f)\cdot p(\mathcal{P}_f) \tag{4.11}$$

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 maximum likelihood estimate (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 $\mathcal{P}_f^* = \arg\max_{\mathcal{P}_f}\{\log(p(\mathcal{P}_f|\mathcal{D}))\}$.

4.4 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 context, the aim is the characterization of the posterior predictive probability density function (pdf) when posterior hyper-parameters uncertainty is marginalized

$$\tilde{T} \sim p(t|\mathcal{D}_i) = \int p(t; \mathcal{P}_t)\cdot p(\mathcal{P}_t|\mathcal{D}_i)d\mathcal{P}_t. \tag{4.12}$$

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

$$p(\mathcal{P}_t|\mathcal{D}_i) = \frac{p(\mathcal{D}_i|\mathcal{P}_t)\cdot p(\mathcal{P}_t)}{p(\mathcal{D}_i)} \tag{4.13}$$

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 4.2, 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.

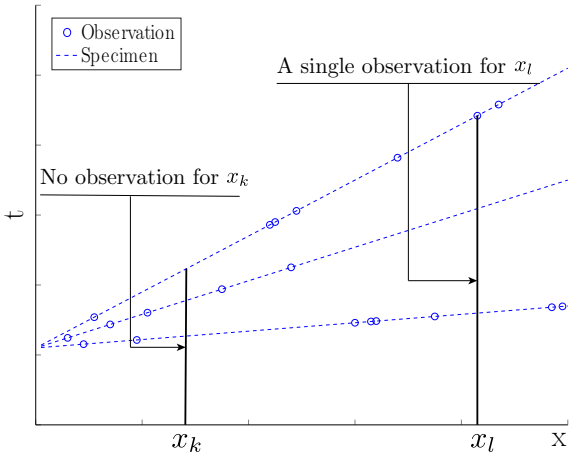


Figure 4.2 Example of observations obtained on replicated specimens for a set of covariates values. This Figure illustrates the challenge that for most covariate x either only one or no observation is available.

4.4.1 Combining GPR and conjugate priors

The first aspect of the method proposed consists in employing the GPR method to build a joint model 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 attribute x_i, it will be possible to employ conjugate priors to characterize inter-specimen variability.

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

$$g(x_k, x_l, s_k, s_l) = \left(\sigma_f^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell^2}\right] + \sigma_v^2 \delta(x_k, x_l)\right) \cdot \delta(s_k, s_l). \tag{4.14}$$

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 4.3.2, 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 4.3 schematizes the predictions of the GPR. Because of the covariance function in Equation 4.14, 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 $\mathbf{\Sigma} = \text{cov}(\mathbf{f})$.

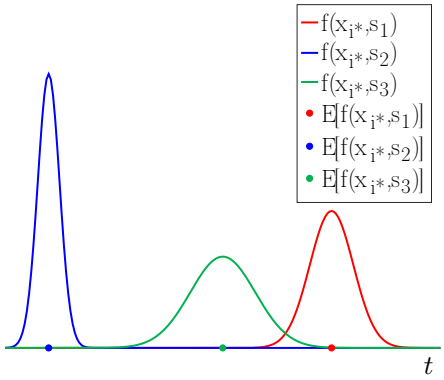


Figure 4.3 Example of predicted marginal distributions for three specimens for the attribute value x_{i*}.

4.4.2 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 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 $f(x_{i*}, \mathbf{s})$.

As explained in Section 4.4, the predictive inter-specimens variability $\tilde{T} \sim p(t; \mathcal{D}_i)$ can be modeled with analytic formulations provided by conjugate priors. However, these formulations can only be applied to a dataset of deterministic values. In this project, 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 $\tilde{T}$. For now, the model 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 and passed to the conjugate prior so that $\mathcal{D}_i^q = \{(x_{i*}, \mathbf{f}_q)\}$.

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,

$$p(\mathcal{P}_t|\mathcal{D}_i^q) = \frac{p(\mathcal{D}_i^q|\mathcal{P}_t)\cdot p(\mathcal{P}_t)}{p(\mathcal{D}_i^q)} \tag{4.15}$$

$$\propto \underbrace{p(\mathcal{D}_i^q|\mathcal{P}_t)}_{\substack{\text{Likelihood}\\ \mathcal{N}(\mu_i,\sigma_i^2)}} \cdot \underbrace{p(\mathcal{P}_t)}_{\substack{\text{Prior}\\ \mathcal{NIG}(m_0,V_0,a_0,b_0)}} . \tag{4.16}$$

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 [1], their initial values $\mathcal{P}_c=\{m_0,V_0,a_0,b_0\}$ is chosen to be equal to zero. The hyper-parameters posterior follows :

$$p(\mu_i,\sigma_i^2|\mathcal{D}_i^q) = \mathcal{NIG}(m_{N_s},V_{N_s},a_{N_s},b_{N_s}) \tag{4.17}$$

$$V_{N_s}^{-1} = V_0^{-1}+N_s = N_s, \tag{4.18}$$

$$\frac{m_{N_s}}{V_{N_s}} = V_0^{-1}m_0 + N_s\overline{\mathbf{f}_q} = N_s\overline{\mathbf{f}_q}, \tag{4.19}$$

$$a_{N_s} = a_0+\frac{N_s}{2}=\frac{N_s}{2}, \tag{4.20}$$

$$b_{N_s} = b_0+\frac{1}{2}\left[m_0^2V_0^{-1}+\sum {f_q}^2 - m_{N_s}^2V_{N_s}^{-1}\right] \tag{4.21}$$

$$= \frac{1}{2}\left[\sum {f_q}^2 - m_{N_s}^2V_{N_s}^{-1}\right]. \tag{4.22}$$

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,

1. Parameters of the hyper-parametres

$$\tilde{T}_q \sim p(t_i|\mathcal{D}_i^q) = \iint p(t_i|\mu_i,\sigma_i^2)\cdot p(\mu_i,\sigma_i^2|\mathcal{D}_i^q)\,d\mu_i\,d\sigma_i^2 \tag{4.23}$$

$$= t_{2a_{N_s}}\left(m_{N_s},\frac{b_{N_s}(1+V_{N_s})}{a_{N_s}}\right) \tag{4.24}$$

$$= t_{2a_{N_s}}(\mu_{\tilde{T}},\sigma^2_{\tilde{T}_\Sigma}). \tag{4.25}$$

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

$$\sigma^2_{\tilde{T}} = \frac{\nu}{\nu-2}\sigma^2_{\tilde{T}_\Sigma}, \quad \text{with } \nu = 2a_{N_s} \text{ posterior degrees of freedoms.} \tag{4.26}$$

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

4.4.3 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*},

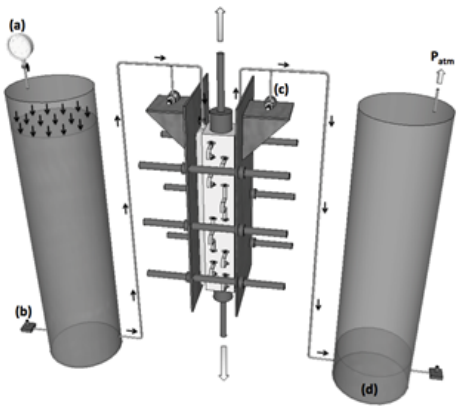
$$\mathbb{E}[\tilde{T}] \approx \frac{1}{Q}\sum_q \tilde{t}_q \tag{4.27}$$

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 $N+1$ 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.

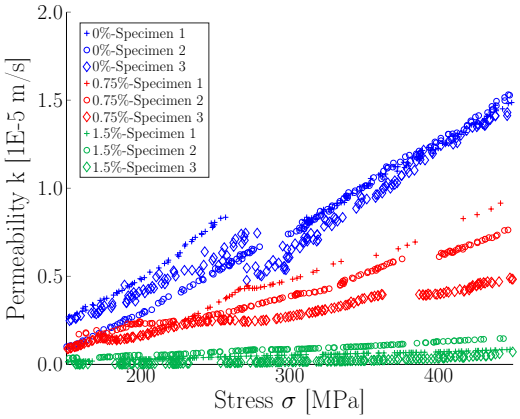
4.5 Case-Study : Permeability of High Performance Fiber-Reinforced Concrete

4.5.1 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 4.4a. 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 4.4b pre-



(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 4.4 Permeability measurement setup and representation of the test results dataset.

sents 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 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.

4.5.2 Probabilistic models

Hypotheses

In this case-study, heteroscedasticity is observable in permeability measurements for three replicated specimens which are function of one attribute, the stress measured in rebars. Indeed, the tensile loading on three replicated specimens under the same attribute (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. In order to maximise the quality of predictions, two prior models are each having a different covariance function are considered. 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,

$$\mu(x_i) = a \cdot x_i + b \qquad \text{with } a, b \in \mathbb{R} \tag{4.28}$$

and covariance function,

$$\text{Model 1 :} \qquad g_1(x_k, x_l, s_k, s_l) = \left(\sigma_{f_1}^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell_1^2}\right] + \sigma_v^2 \delta(x_k, x_l)\right) \cdot \delta(s_k, s_l)$$

$$\text{Model 2 :} \qquad g_2(x_k, x_l, s_k, s_l) = \left(\sum_{r=1}^{3} \sigma_{f_r}^2 \exp\left[\frac{-(x_k - x_l)^2}{2\ell_r^2}\right] + \sigma_v^2 \delta(x_k, x_l)\right) \cdot \delta(s_k, s_l)$$

The main structure of the covariance function is the square exponential (SE). The first one is the generic SE function ; the second adds two other covariance functions, each having a different correlation length. It allows to consider the impact of one attribute 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 GPML. To ensure a strictly positive water permeability values, the method is applied in the log-space which means using the log-permeability, the dataset becomes $\mathcal{D} = \{(x_i, \log(y_i)), i = 1, \cdots, N\}$. Figure 4.5 presents the data set plot in log-space. Note that the scale of the vertical axis has been modified by the log-transformation.

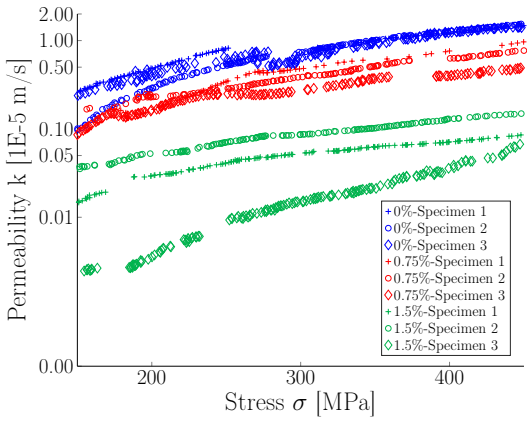


Figure 4.5 Representation of permeability observations function of the attribute stress in log-space.

Hyper-parameters calibration

Once GPR prior structure is defined, hyper-parameters are identified by Maximum Likelihood Estimation (MLE). As two models are studied, two sets of hyper-parameters, per fiber ratio, have to be estimated, $\mathcal{P}_{f1} = \{a, b, \sigma_{f_1}, \ell_1, \sigma_v\}$ for the first model, $\mathcal{P}_{f2} = \{a, b, \sigma_{f_1}, \ell_1, \sigma_{f_2}, \ell_2, \sigma_{f_3}, \ell_3, \sigma_v\}$ for the second model.

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. Tables 4.1 and 4.2 gather the results of the highest Log-Likelihood (LL) and the hyper-parameters associated, for the three fiber ratios 0%, 0.75%, 1.5%, for each model.

The second model is more accurate for all fiber ratios as its Log-likelihoods are higher in the three cases. Model 2 is thus selected for further calculation. The following section gathers the results based on a covariance function of three correlation lengths with the parameters values presented in Table 4.2.

Table 4.1 MLE Results - Model 1

Fiber Ratio	**Hyper-parameters**					**LL**
	a	b	σ_{f_1}	ℓ_1	σ_v	$\cdot 10^3$
0%	0.0043	−13.15	0.092	49.01	0.048	1.56
0.75%	0.0062	−14.54	0.103	103.92	0.055	1.20
1.5%	0.0084	−17.53	0.156	113.69	0.066	1.42

Table 4.2 MLE Results - Model 2

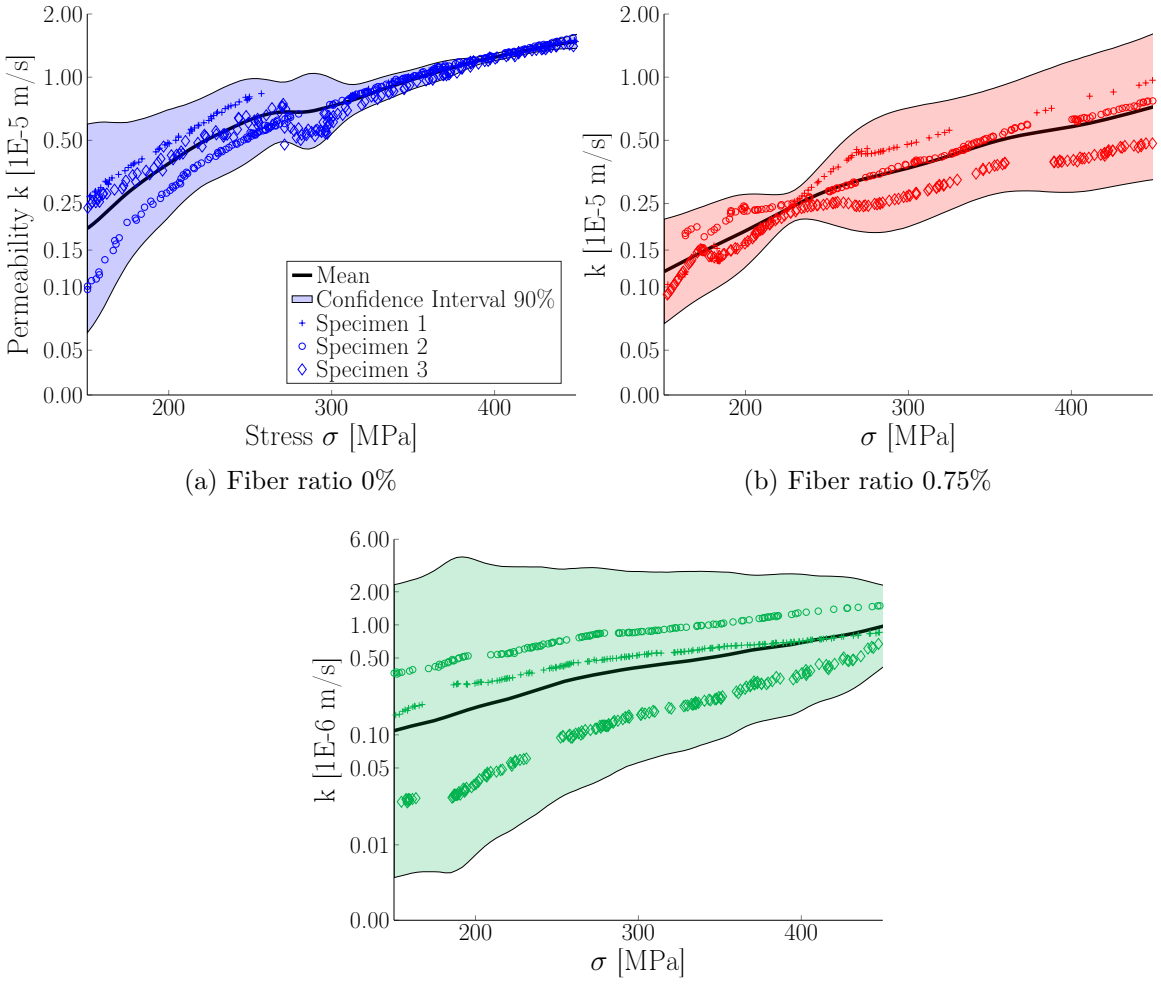
Fiber Ratio	**Hyper-parameters**									**LL**
	a	b	σ_{f_1}	ℓ_1	σ_{f_2}	ℓ_2	σ_{f_3}	ℓ_3	σ_v	$\cdot 10^3$
0%	0.0049	−13.33	0.213	288.42	0.094	76.22	0.012	9.46	0.032	1.70
0.75%	0.0060	−14.53	0.124	222.45	0.017	31.84	0.0024	10.25	0.031	1.30
1.5%	0.0080	−17.50	0.337	271.76	0.041	53.15	0.013	8.01	0.029	2.10

4.5.3 Results

Prediction of a new specimen

Figures 4.6 and 4.7 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 4.6 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 4.6c 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 4.7 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.

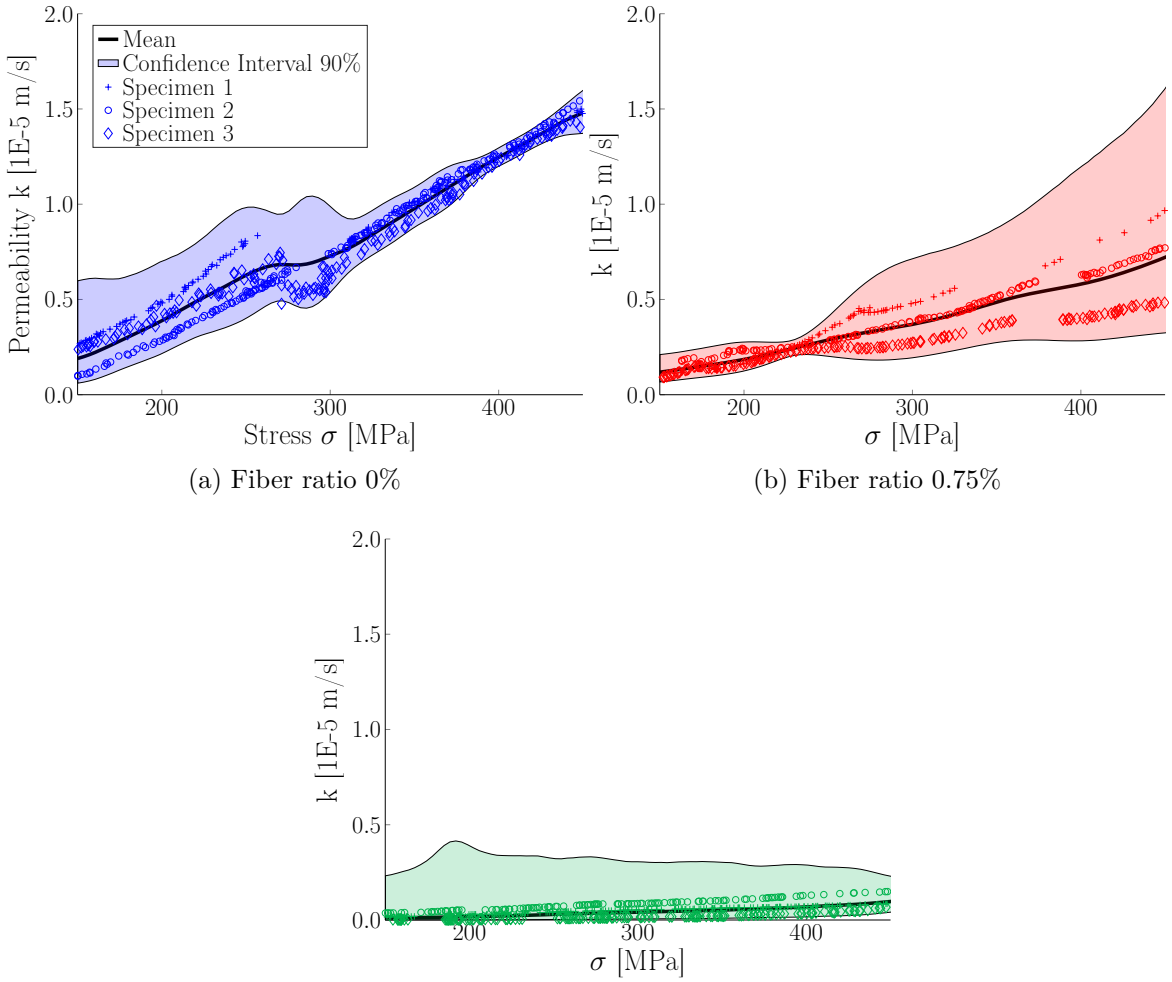


(a) Fiber ratio 0%

(b) Fiber ratio 0.75%

(c) Fiber ratio 1.5%

Figure 4.6 Prediction of water permeability for a fourth untested specimen in the log-space.



(a) Fiber ratio 0%

(b) Fiber ratio 0.75%

(c) Fiber ratio 1.5%

Figure 4.7 Prediction of water permeability for a fourth and untested specimen in the original space.

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 4.8 juxtapose the three predictive models in the log and original spaces. Figure 4.8 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 4.5.2 for the GPR model. The results of samples comparisons are summarized in Figure 4.8c. In this case, the variable $\tilde{T}_q$ follows a Multivariate t-distribution :

$$\tilde{T}_q \sim T(\mathbf{M}_{\tilde{T}}, \mathbf{\Sigma}_{\tilde{T}}, \nu) \quad \text{with,} \tag{4.29}$$

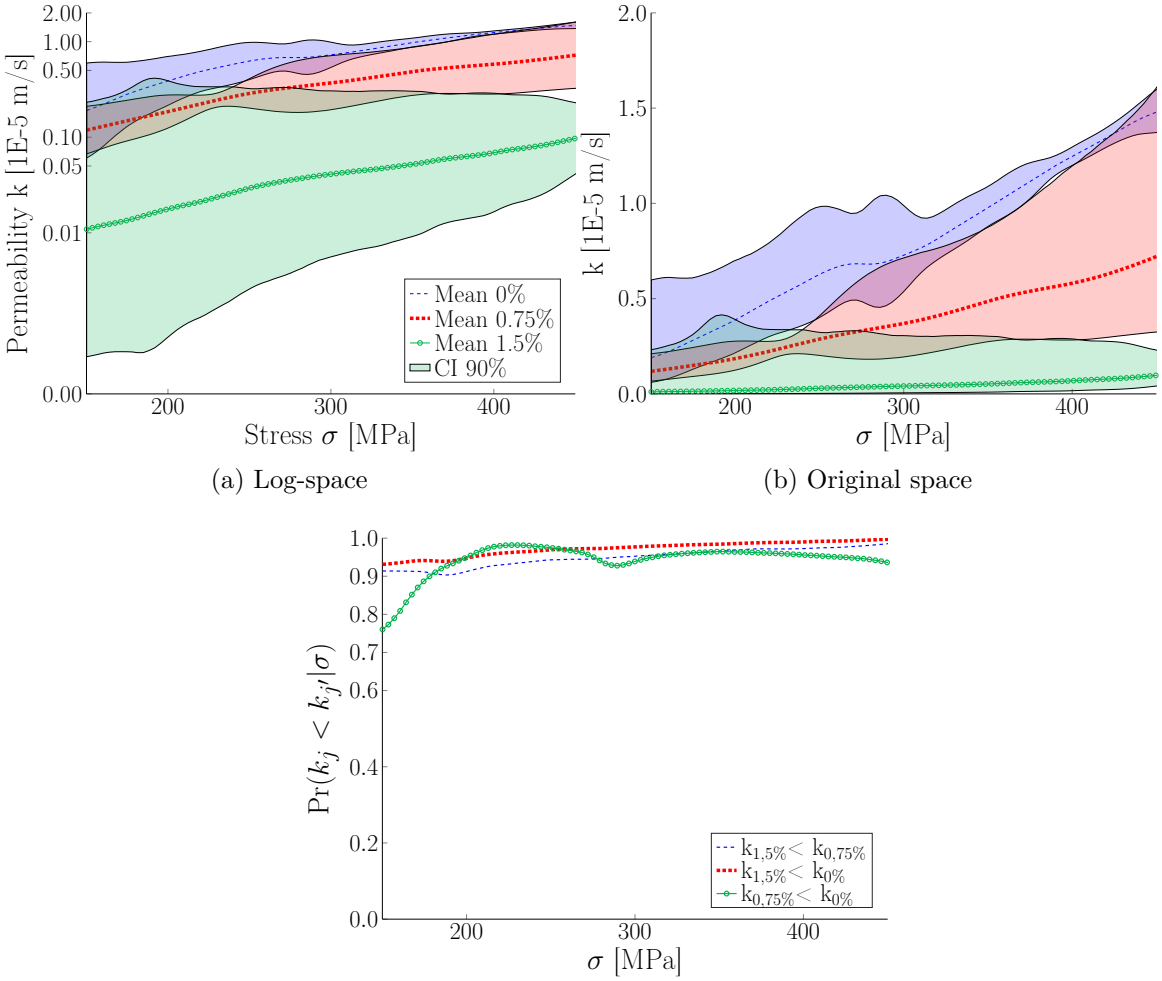
$$\mathbf{M}_{\tilde{T}} = \begin{bmatrix} \mu_{\tilde{T}_1} \\ \mu_{\tilde{T}_2} \\ \vdots \\ \mu_{\tilde{T}_N} \end{bmatrix}, \tag{4.30}$$

$$\mathbf{\Sigma}_{\tilde{T}} = \begin{bmatrix} \sigma^2_{\tilde{T}_{\Sigma_1}} & \rho_{1,2}\sigma_{\tilde{T}_{\Sigma_1}}\sigma_{\tilde{T}_{\Sigma_2}} & \cdots & \rho_{1,N}\sigma_{\tilde{T}_{\Sigma_1}}\sigma_{\tilde{T}_{\Sigma_N}} \\ \rho_{2,1}\sigma_{\tilde{T}_{\Sigma_2}}\sigma_{\tilde{T}_{\Sigma_1}} & \sigma^2_{\tilde{T}_{\Sigma_2}} & \cdots & \rho_{2,N}\sigma_{\tilde{T}_{\Sigma_2}}\sigma_{\tilde{T}_{\Sigma_N}} \\ \vdots & \vdots & \ddots & \cdots \\ \rho_{N,1}\sigma_{\tilde{T}_{\Sigma_N}}\sigma_{\tilde{T}_{\Sigma_1}} & \rho_{N,2}\sigma_{\tilde{T}_{\Sigma_N}}\sigma_{\tilde{T}_{\Sigma_2}} & \cdots & \sigma^2_{\tilde{T}_{\Sigma_N}} \end{bmatrix}, \tag{4.31}$$

$$\nu = 2a_{N_s} = \frac{N_s}{2} \tag{4.32}$$

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 $R = \text{chol}(\Sigma_{\tilde{T}})$, $(R^\intercal R = \Sigma_{\tilde{T}})$.

Figure 4.9 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



(a) Log-space

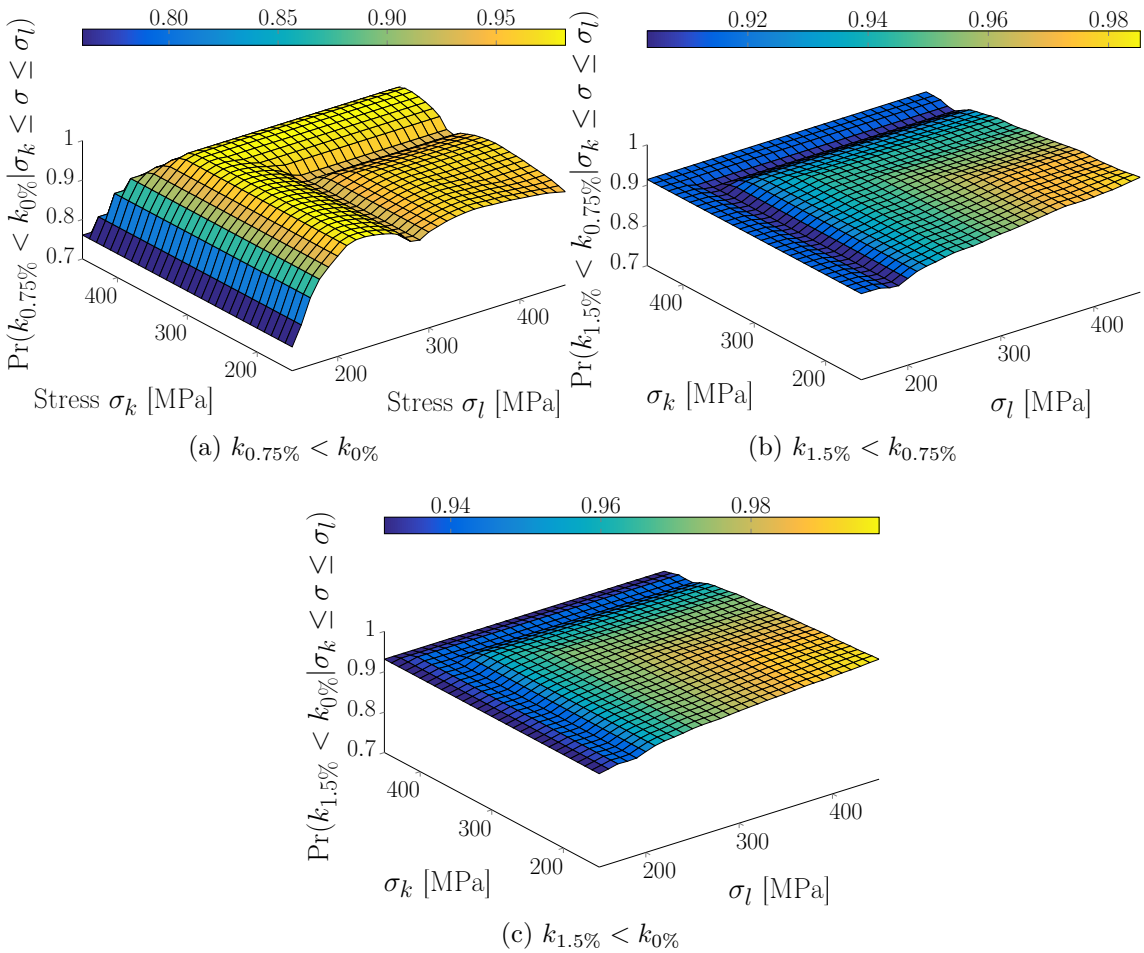
(b) Original space

(c) Conditional probabilities

Figure 4.8 (a),(b) Predictive models and (c) comparison of permeability in high-performance fiber reinforced concrete for the tested fiber reinforcement ratios.

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 4.8c.

The probabilities $\Pr(k_j < k_{j'} | 150\,\text{MPa} \leq \sigma \leq 450\,\text{MPa})$ are presented in Table 4.3 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.



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

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

Figure 4.9 Global probabilities of $\Pr(k_j < k_{j'} | \sigma_k \leq \sigma \leq \sigma_l)$ for permeability across each pair of fiber ratios

4.6 Discussion

Probabilistic analyses demonstrate that an introduction of 0.75% or 1.5% of fibers 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 biaised hyper-parameters. A careful

Tableau 4.3 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

choice of initial parameters values is therefore essential. 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.

4.7 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

A. K. Bansal et P. Aggarwal, "Bayes prediction for a heteroscedastic regression superpopulation model using balanced loss function", *Communications in Statistics—Theory and Methods*, vol. 36, no. 8, pp. 1565–1575, 2007.

G. Blau, M. Lasinski, S. Orcun, S.-H. Hsu, J. Caruthers, N. Delgass, et V. Venkatasubramanian, "High fidelity mathematical model building with experimental data : A bayesian approach", *Computers & Chemical Engineering*, vol. 32, no. 4, pp. 971–989, 2008.

P. Gardoni, A. Der Kiureghian, et K. Mosalam, "Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations", *Journal of Engineering Mechanics*, vol. 128, no. 10, pp. 1024–1038, 2002.

P. Gardoni, K. Nemati, et T. Noguchi, "Bayesian statistical framework to construct probabilistic models for the elastic modulus of concrete", *Journal of Materials In Civil Engineering*, vol. 19, no. 10, pp. 898–905, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, et D. B. Rubin, *Bayesian data analysis*. Taylor & Francis, 2014, vol. 2.

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

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

J.-S. Jeon, A. Shafieezadeh, et R. DesRoches, "Statistical models for shear strength of rc beam-column joints using machine-learning techniques", *Earthquake Engineering & Structural Dynamics*, vol. 43, no. 14, pp. 2075–2095, 2014.

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

J. Lampinen et A. Vehtari, "Bayesian approach for neural networks—review and case studies", *Neural networks*, vol. 14, no. 3, pp. 257–274, 2001.

Q. V. Le, A. J. Smola, et S. Canu, “Heteroscedastic gaussian process regression”, dans *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 489–496.

Y. Ma, L. Wang, J. Zhang, Y. Xiang, et Y. Liu, “Bridge remaining strength prediction integrated with bayesian network and in situ load testing”, *Journal of Bridge Engineering*, vol. 19, no. 10, p. 04014037, 2014.

D. MacKay, “Introduction to gaussian processes”, *NATO ASI Series F Computer and Systems Sciences*, vol. 168, pp. 133–166, 1998.

K. P. Murphy, “Conjugate bayesian analysis of the gaussian distribution”, p. 16, 2007.

M. Pal et S. Deswal, “Modeling pile capacity using support vector machines and generalized regression neural network”, *Journal of geotechnical and geoenvironmental engineering*, vol. 134, no. 7, pp. 1021–1024, 2008.

C. Rasmussen et H. Nickisch, “Gaussian processes for machine learning (gpml) toolbox”, *The Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

C. Rasmussen et C. Williams, “Gaussian processes for machine learning”, *the MIT Press*, vol. 2, no. 3, p. 4, 2006.

R. Siddique, P. Aggarwal, Y. Aggarwal, et S. Gupta, “Modeling properties of self-compacting concrete : support vector machines approach”, *Computers and Concrete*, vol. 5, no. 5, pp. 123–129, 2008.

M. Słoński, “A comparison of model selection methods for compressive strength prediction of high-performance concrete using neural networks”, *Computers & structures*, vol. 88, no. 21, pp. 1248–1253, 2010.

M. Sloński, “Bayesian neural networks and gaussian processes in identification of concrete properties”, *Computer Assisted Mechanics and Engineering Sciences*, vol. 18, no. 4, pp. 291–302, 2011.

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

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

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

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, et A. Vehtari, "Bayesian modeling with gaussian processes using the gpstuff toolbox", *arXiv preprint arXiv :1206.5754*, 2012.

C. Wang, "Gaussian process regression with heteroscedastic residuals and fast mcmc methods", Thèse de doctorat, University of Toronto, 2014.

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

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

J. Zhong, P. Gardoni, D. Rosowsky, et T. Haukaas, "Probabilistic seismic demand models and fragility estimates for reinforced concrete bridges with two-column bents", *Journal of engineering mechanics*, vol. 134, no. 6, pp. 495–504, 2008.

CHAPITRE 5 DISCUSSION GÉNÉRALE ET APPROCHE COMPLÉMENTAIRE

Ce chapitre est une analyse critique de la méthode probabiliste, par processus Gaussien développée au chapitre précédent et des résultats en découlant. Une approche alternative sera aussi présentée pour répondre à la problématique de ce projet de recherche offrant une nouvelle perspective d'application, mais cette méthode est encore en développement. Finalement, une comparaison des deux méthodes sera proposée afin d'identifier les champs d'application privilégiés de chacune.

5.1 Discussion générale : méthode par processus Gaussien

Cette section portera sur les avantages et limites de la méthode par processus Gaussien présentée dans l'article au Chapitre 4. Elle permettra de souligner les apports de cette méthode dans les domaines de génie, mais aussi les certaines limites de son application.

5.1.1 Résultats de l'application de la méthode par processus Gaussien aux essais de perméabilité

Suite au développement théorique de la méthode probabiliste associant GPR et distributions à priori conjuguées au Chapitre 4, celle-ci fût appliquée au cas pratique d'essais de perméabilité à l'eau sur des spécimens de BHP et de BFHP. Les objectifs de cette modélisation étaient de confirmer une tendance de comportement de la perméabilité à l'eau en fonction de la contrainte appliquée dans la barre d'armature, mais aussi de quantifier le bénéfice d'incorporation de fibres au BHP. Ces objectifs ont été atteints, en effet, la Figure 4.6 présente l'évolution de la perméabilité dans les trois conditions d'essais, c'est-à-dire pour les trois taux de fibres testés, il apparaît que la prédiction d'un quatrième spécimen confirme l'augmentation de la perméabilité à l'eau en fonction de l'augmentation de la contrainte dans l'armature du tirant, mais aussi sa diminution importante avec l'augmentation du volume de fibres dans le béton. Les trois modèles probabilistes produisent chacun la prédiction d'un quatrième spécimen associé à un intervalle de confiance. Une tendance de comportement se dégage dans chaque cas, puisque cet intervalle de confiance est relativement modéré, permettant alors de comparer les variations de perméabilité à l'eau entre les différents taux de fibres. Avec de trop importants intervalles de confiance, les comparaisons des modèles ne fourniraient pas de résultats si tranchés, puisqu'il y aurait la même probabilité d'obtenir une réduction de

perméabilité à l'eau avec des taux de fibres différents, ceci à cause du chevauchement des intervalles de confiance des trois modèles. Dans le cas présent, ces chevauchements existent pour certaines contraintes (Figure 4.8), mais les intervalles de confiance se distinguent et augmentent alors les probabilités d'obtenir une perméabilité plus faible avec un taux de fibres plus élevé. Les résultats du Tableau 4.3 confirment alors ce qui précède, les probabilités d'obtenir une perméabilité plus faible avec l'ajout de fibres varient de 76% à 93%, consolidant alors les tendances relevées au cours des essais. Finalement, la modélisation probabiliste associée à la méthode par processus Gaussien a permis de mettre en évidence et de confirmer la forte réduction de la perméabilité pour les taux de fibres 0.75% et 1.5% des BFHP en comparaison au BHP (0%). Et ces trois modèles ont permis de dégager quantitativement le bénéfice d'incorporation de fibres dans le béton confirmant ainsi le potentiel du BFHP à présenter une excellente durabilité par rapport au BHP. Cette cohérence des résultats entre l'expérimental et le numérique renforce la pertinence de la méthode par processus Gaussien.

5.1.2 Une méthode performante pour un grand champ d'applications

La méthode de modélisation probabiliste développée est une méthode efficace car elle permet d'obtenir en quelques heures le modèle prédictif des données des essais en laboratoire. Sa force réside dans la formulation analytique de la distribution à postériori prédictive des observations et du choix de la méthode MLE pour inférer les hyper-paramètres. De plus, elle peut être utilisée par les ingénieurs dans divers domaines de recherche, puisque la méthode est adaptable à tout type d'essai. Finalement, les scientifiques sont capables grâce à cette méthode d'interpréter les résultats d'expériences en laboratoire. En effet, cette modélisation permet de confirmer une tendance générale du comportement étudié, et ainsi de comparer quantitativement différentes variations de l'essai. La méthode par processus Gaussien répond aux objectifs visés dans ce projet. À partir de quelques spécimens testés présentant un comportement hétéroscédastique, il est possible de modéliser le comportement de l'essai en fournissant la tendance moyenne et l'intervalle de confiance associé, qui lui-même prend en compte l'incertitude liée au faible nombre de spécimens. En résumé, cette méthode rapide et performante est accessible par les ingénieurs de tous les domaines grâce à l'adaptabilité de son application à tout essai et grâce à sa formulation analytique claire.

5.1.3 Nombre de spécimens testés

L'application de la méthode par processus Gaussien a été effectuée sur un jeu de données constitué de trois spécimens donnant alors de très bons résultats quant à la probabilité d'obtenir une perméabilité plus faible avec un taux de fibres plus important qu'un autre. Il est

pertinent de s'interroger sur l'impact sur les résultats du modèle qu'aurait l'ajout d'un quatrième spécimen dans le jeu de données, c'est-à-dire observer comment l'intervalle de confiance réduirait dans le modèle prédictif et de ce fait augmenterait les probabilités de comparaison des trois taux de fibres. Pour cela, un quatrième spécimen peut être simulé numériquement afin d'étendre le jeu de données et suite à cela la méthode par processus Gaussien peut être appliquée à ce nouveau jeu de données. Suivant les résultats, l'ingénieur peut décider de lancer d'autres essais si l'incertitude sur le modèle se voit réduite significativement. Dans le cas contraire, si la diminution de l'incertitude est négligeable alors le nombre de spécimens testés est suffisant pour établir un modèle performant. Cette idée peut être étendue à tout essai, grâce à la rapidité de l'application de la méthode, le scientifique peut à chaque spécimen testé choisir d'en produire un autre pour rendre le modèle statistique optimal ou d'arrêter les tests si celui-ci l'est déjà. Afin d'éviter aux ingénieurs de devoir réimplanter numériquement la méthode à chaque essai, il serait utile de créer une application accessible à tous, dans laquelle la méthode serait déjà programmée. Cette application avec une configuration standard prendrait les données d'essais en entrée pour en fournir le modèle prédictif associé.

5.1.4 Limites de la méthode par processus Gaussien

Le chapitre précédent a présenté une méthode peu coûteuse en programmation, rapide et performante pour la modélisation probabiliste d'essais en laboratoire avec peu de spécimens. Cependant, cette méthode par processus Gaussien présente quelques limitations. En effet, il ne serait pas possible d'ajouter la covariable taux de fibres à la covariable contrainte et de prédire le comportement de la perméabilité à l'eau dans le BFHP pour un taux de fibres non testé, puisque seulement trois taux de fibres sont connus. Cette approche requiert plus d'observations avec des taux de fibres différents, l'incertitude entre deux taux de fibres connus serait bien trop élevée dans la prédiction d'un quatrième taux de fibres. A titre de comparaison, le jeu de données d'observations dispose de la perméabilité pour des centaines de valeurs de contraintes différentes, alors qu'il n'y a que trois valeurs de pourcentage de fibres. Il serait alors avantageux d'élaborer une méthode capable de se détacher du nombre de valeurs d'une des covariables.

De plus, dans la méthode par processus Gaussien, l'inférence des hyper-paramètres $\mathcal{P}_f$ s'appuie sur la méthode MLE qui fournit comme valeurs d'hyper-paramètres le mode de leurs distributions négligeant l'incertitude liée aux hyper-paramètres. Cette incertitude est négligeable lorsque le jeu de données est très riche, c'est ce qui a permis de l'utiliser. Il y avait certes peu de spécimens, mais de nombreux points d'observations. La méthode MLE n'est pas toujours performante, puisqu'il peut arriver de tomber sur un maximum local de la

fonction de vraisemblance et non sur le maximum global, rendant les valeurs de sortie des hyper-paramètres non optimales. Trouver le maximum global peut prendre du temps, cela demande de faire varier les valeurs initiales des hyper-paramètres et de chercher le maximum de la fonction de vraisemblance par tâtonnement. La revue de documentation (Chapitre 2) a démontré que l'approche Bayesienne pour l'inférence des paramètres reste, bien que lourde, la méthode la plus fiable, puisqu'elle prend en compte toutes les incertitudes. C'est pourquoi il pourrait être intéressant de l'appliquer à la problématique étudiée afin de contourner les limitations liées à la méthode MLE.

5.2 Méthode par approche Bayesienne

Cette section développera une approche alternative pour répondre à la problématique du projet, une méthode utilisant aussi GPR mais offrant différentes perspectives. Une nouvelle représentation de l'hétéroscédasticité sera introduite et la méthode de modélisation proposée sera basée sur l'approche Bayesienne pour inférer les hyper-paramètres.

5.2.1 Représentation spatiale de l'hétéroscédasticité

La sous-section 2.2.2 a décrit une approche originale de l'hétéroscédasticité, qui consistait à ajouter des covariables cachées dans la fonction de covariance présentée à l'Equation 2.22. Ces variables latentes w_i représentent alors les paramètres variant d'un spécimen à un autre, mais non mesurés ni identifiés (Wang and Neal (2012); Wang (2014)).

Identifier les paramètres physiques variant d'un spécimen à l'autre est presque impossible, ceux-ci sont très nombreux et peut-être non connus ainsi les covariables latentes à ajouter à la fonction de covariance pourraient être très nombreuses. Une nouvelle interprétation spatiale de ces covariables cachées permet de déjouer cette dernière limitation. Au lieu d'être décrites comme des paramètres physiques, les variables latentes peuvent être considérées comme des coordonnées spatiales. Chaque spécimen testé en laboratoire est identifié par des coordonnées spatiales, si N_s spécimens sont testés alors ils peuvent être chacun représentés par un point dans un espace à $N_s - 1$ dimensions. La distance séparant deux spécimens représente alors les différences entre les spécimens qui peuvent être dues à la variation d'un ou plusieurs paramètres. Ainsi les spécimens sont tous éloignés les uns des autres à des distances modélisant la variabilité inter-spécimens. A titre d'exemple, le cas simple représenté dans la Figure 5.1, d'un espace 2-D comprenant donc trois spécimens testés. Chaque spécimen est modélisé par un couple de coordonnées (w_{iX}, w_{iY}) et est séparé d'un autre spécimen d'une distance d_{ij}.

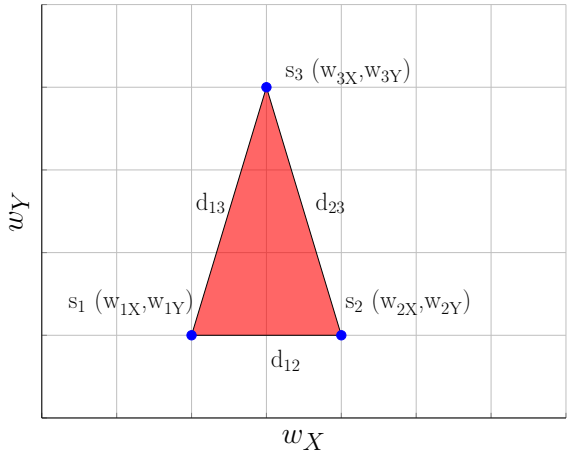


Figure 5.1 Représentation spatiale des covariables latentes pour trois spécimens dans un espace 2-D

Soit un deuxième exemple prenant en données d'entrée les résultats d'essais sur quatre spécimens (Figure 5.2). Ces quatre spécimens sont alors représentés par un triplet de coordonnées (w_{iX}, w_{iY}, w_{iZ}) dans un espace à trois dimensions.

La formulation de la méthode par approche Bayesienne est présentée ici dans le cas où le jeu de données est limité à trois spécimens, ce qui représente le cas d'étude sur la perméabilité à l'eau des BHP et BFHP et qui facilitera aussi la compréhension (Figure 5.1). La méthode est facilement adaptable à N_s spécimens dans un espace à $N_s - 1$ dimensions. La modélisation du comportement étudié à partir des trois spécimens requiert donc d'inférer les trois couples de coordonnées soient six hyper-paramètres $\{w_{1X}, w_{2Y}, w_{2X}, w_{2Y}, w_{3X}, w_{3Y}\}$. Afin de diminuer le nombre d'hyper-paramètres à inférer on supposera $w_{1X} = w_{1Y} = w_{2Y} = 0$, ce qui est raisonnable puisque fixer ces coordonnées n'influe pas sur les distances d_{ij} séparant les spécimens.

D'après les résultats de la méthode MLE présentée au Chapitre 3, la fonction de covariance la plus pertinente à utiliser dans le cas d'étude est décrite par trois longueurs de corrélation et devient avec l'addition des covariables latentes,

$$g(x_i, x_j, d_{ij}) = \sum_{s=1}^{3} \sigma_{f_s}^2 \exp\left[\frac{-(x_i - x_j)^2}{2\ell_s^2} - d_{ij}^2\right] + \sigma_v^2 \delta(x_i, x_j) \tag{5.1}$$

$$g(x_i, x_j, w_i, w_j) = \sum_{s=1}^{3} \sigma_{f_s}^2 \exp\left[\frac{-(x_i - x_j)^2}{2\ell_s^2} - (w_i - w_j)^2\right] + \sigma_v^2 \delta(x_i, x_j) \tag{5.2}$$

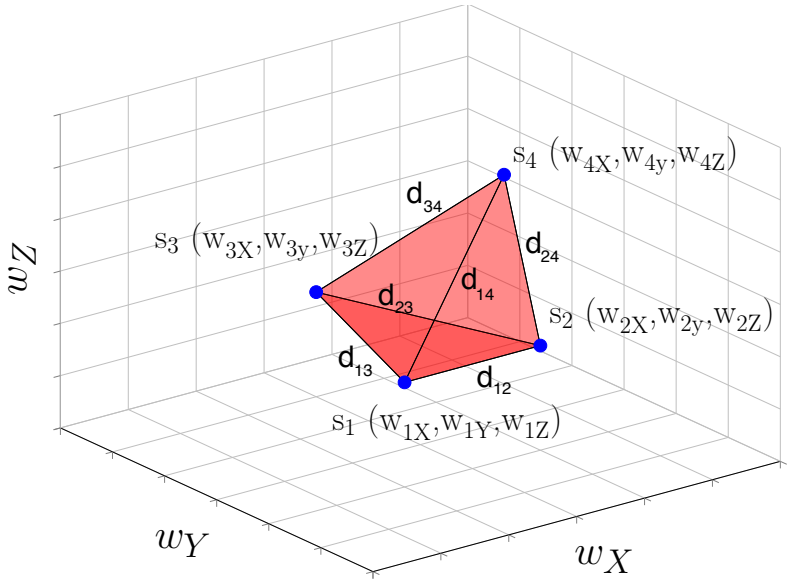


Figure 5.2 Représentation spatiale des covariables latentes pour quatre spécimens dans un espace 3-D

La formulation de la fonction de covariance avec les coordonnées virtuelles w_i est préférée par rapport à celle avec les distances virtuelles d_{ij}. En effet, dans la prédiction d'un quatrième spécimen non testé, les coordonnées initiales inférées des trois spécimens testés sont utilisées afin de déterminer par échantillonnage les coordonnées virtuelles du quatrième et nouveau spécimen non testé, tel représenté dans la Figure 5.3. Le quatrième spécimen est inféré à partir des trois spécimens testés, il se trouve ainsi dans le plan formé de ces trois derniers spécimens. Choisir d'inférer les distances virtuelles dans la fonction de covariance implique de déterminer les coordonnées à partir des distances pour estimer l'emplacement du nouveau spécimen. Déterminer les coordonnées à partir des distances est tout à fait possible, en effet la représentation spatiale des spécimens est un simplexe qui est une généralisation du triangle à une dimension quelconque. Un simplexe est l'enveloppe convexe d'un ensemble de N_s points utilisé pour former un repère affine dans un espace affine de dimension $N_s - 1$. À partir de l'exemple présenté à la Figure 5.1, il faudrait calculer trois coordonnées $\{w_{2X}, w_{3X}, w_{3Y}\}$ connaissant trois distances $\{d_{12}, d_{13}, d_{23}\}$, le système d'équations a été décrit par Erlandson (2016). Pour des dimensions d'espace supérieures à 3D il est fort possible que les solutions d'équations, soient les coordonnées des spécimens, se trouvent dans l'espace complexe $\mathbb{C}$ rendant alors la matrice de covariance inadmissible. Pour éviter des résultats de coordonnées dans l'espace complexe, la formulation de la fonction de covariance (Equation 5.2) retenue est celle décrite avec les coordonnées virtuelles w_i contraintes à l'espace réel $\mathbb{R}^2$. Inférer

directement les distances d_{ij} requiert d'ajouter un critère de vérification de la définition semi-positive de la matrice de covariance. Pour chaque paire de spécimens, à la sortie de la

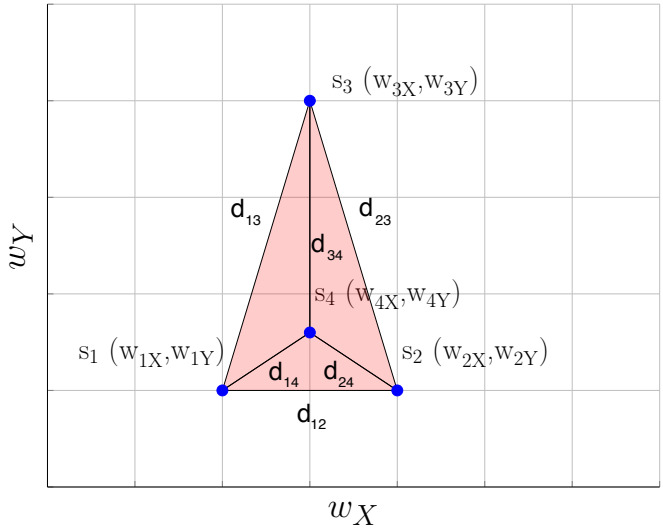


Figure 5.3 Nouveau et quatrième spécimen échantillonnée à partir des trois spécimens testés

méthode d'inférence des hyper-paramètres un contrôle est effectué sur la distance virtuelle d_{ij}. En effet, comme précisé plus haut la matrice de covariance n'est pas admissible si les coordonnées virtuelles w_i et w_j sont obtenues dans l'espace complexe $\mathbb{C}$. Si tel est le cas, la méthode est d'ajouter une variation Δ_{ij} à la distance inférée jusqu'à obtenir des coordonnées virtuelles dans l'espace réel $\mathbb{R}$.

Le simplexe formé est de 3 dimensions, correspondant au nombre de spécimens testés. Chaque spécimen est représenté par des coordonnées virtuelles réelles dans l'espace 2-D. Les distances virtuelles représentent donc les différences entre deux spécimens, la variabilité inter-spécimens.

Cette méthode de vérification de la semi-positivité de la matrice a été appliquée dans le cadre de la modélisation de sols contaminés (Quach et al., 2017). Cette dernière application s'est appuyée sur la méthode MLE pour inférer les hyper-paramètres de la structure à priori du processus Gaussien. Or, comme souligné dans la discussion générale sur la méthode par processus Gaussien (Section 5.1), la méthode MLE n'est performante que si le jeu de données est riche, puisqu'elle occulte les incertitudes liées aux hyper-paramètres. Mais elle peut se heurter à un minimum local de la fonction de vraisemblance et donc ne pas fournir les paramètres optimaux. C'est pourquoi il est préférable d'utiliser une approche Bayesienne pour inférer les hyper-paramètres et surmonter les limites de la méthode MLE.

5.2.2 Echantillonnage des hyper-paramètres et hyper-hyper-paramètres

L'étape principale de la méthode est l'inférence des hyper-paramètres. Le modèle prédictif reste basé sur la méthode GPR, étant donné le jeu de données $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, et sachant que les observations suivent la loi normale multivariée $\mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$, leur définition reste la suivante,

$$\underbrace{\mathbf{y}}_{\text{observations}} = \underbrace{f(\mathbf{x})}_{\text{réalité}} + \underbrace{\mathbf{v}}_{\text{erreur d'observation}} \qquad \text{avec} \quad \mathbf{v} : V \sim \mathcal{N}(0, \sigma_v^2) \tag{5.3}$$

En supposant une moyenne à priori affine, l'hypothèse basée sur le modèle défini au Chapitre 3, telle que $\mathbf{M} = a \cdot \mathbf{x} + b$ et à partir de la fonction de covariance développée à l'Equation 5.2, les hyper-paramètres de la structure à priori de GPR à estimer sont donc $\mathcal{P}_{gp} = \{a, b, \ell_1, \sigma_{f1}, \ell_2, \sigma_{f2}, \ell_3, \sigma_{f3}\, \sigma_v\}$ et $\mathcal{P}_w = \{w_{2X}, w_{3X}, w_{3Y}\}$. Afin de modéliser l'hétéroscédasticité, soit la variabilité inter-spécimen dépendante de l'attribut, l'écart-type sera décrit par la fonction affine, $\sigma_{f1} = c \cdot x_i + d$. Et, les paramètres $\mathcal{P}_{gp}$ à inférer deviennent $\mathcal{P}_{gp} = \{a, b, \ell_1, c, d, \ell_2, \sigma_{f2}, \ell_3, \sigma_{f3}\, \sigma_v\}$.

Les coordonnées virtuelles $\mathcal{P}_w$ dépendent elles-mêmes d'hyper-hyper-paramètres, en assumant qu'elles suivent une distribution normale multivariée,

$$\begin{bmatrix} w_{iX} \\ w_{iY} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_{XY}) \quad \text{avec} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \quad \text{et} \quad \mathbf{\Sigma}_{XY} = \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \tag{5.4}$$

Ainsi, afin d'inférer les hyper-paramètres $\mathcal{P}_w = \{w_{2X}, w_{3X}, w_{3Y}\}$, il faut inférer les hyper-hyper-paramètres $\mathcal{P}_f = \{\mu_X, \mu_Y, \rho_{XY}, \sigma_X, \sigma_Y\}$. Cette structure hiérarchique des paramètres est interprétable par une formulation empirique du théorème de Bayes. En effet, l'inférence des hyper-paramètres $\mathcal{P}_w$ et hyper-hyper-paramètres $\mathcal{P}_f$ s'appuie sur la formule de Bayes afin de prendre en compte les incertitudes liées à chaque paramètre. L'ensemble des paramètres de la structure à priori de GPR sera échantillonné à partir de la formule de Bayes empirique. Ces échantillons permettront alors de prédire un nouveau spécimen, les détails de l'étape de prédiction seront décrits dans la Section 5.2.3. Il faut d'abord décrire la formulation Bayesienne empirique pour l'échantillonnage. Les échantillons des hyper et hyper-hyper-paramètres $\mathcal{P}_q$ sont tirés de leur distribution à postériori.

$$\mathcal{P}_q \equiv \mathcal{P}_q|\mathcal{D} = \{\mathcal{P}_{gp,q}, \mathcal{P}_{w,q}, \mathcal{P}_{f,q}\}|\mathcal{D} \tag{5.5}$$

$$\sim p(\mathcal{P}_{gp}, \mathcal{P}_w, \mathcal{P}_f|\mathcal{D}) \tag{5.6}$$

$$= \frac{p(\mathbf{y}|\mathcal{P}_{gp}, \mathcal{P}_w) \cdot p(\mathcal{P}_{gp}) \cdot p(\mathcal{P}_w|\mathcal{P}_f) \cdot p(\mathcal{P}_f)}{p(\mathbf{y})} \tag{5.7}$$

$$\propto p(\mathbf{y}|\mathcal{P}_{gp}, \mathcal{P}_w) \cdot p(\mathcal{P}_{gp}) \cdot p(\mathcal{P}_w|\mathcal{P}_f) \cdot p(\mathcal{P}_f). \tag{5.8}$$

La fonction de vraisemblance marginale $p(\mathbf{y}|\mathcal{P}_{gp}, \mathcal{P}_w)$ est directement estimée grâce à sa formulation analytique. Soit l'ensemble des paramètres excepté l'erreur de mesure σ_v, $\mathcal{P} = \{a, b, \ell_1, c, d, \ell_2, \sigma_{f2}, \ell_3, \sigma_{f3}, w_{2X}, w_{3X}, w_{3Y}, \mu_X, \mu_Y, \rho_{XY}, \sigma_X, \sigma_Y\}$ et la fonction de covariance, de même sans σ_v,

$$g_{xx}(x_i, x_j, w_i, w_j) = \sum_{k=1}^{3} \sigma_{f_k}^2 \exp\left[\frac{-(x_i - x_j)^2}{2\ell_k^2} - (w_i - w_j)^2\right]. \tag{5.9}$$

Alors, d'après la définition des observations par GPR, les valeurs exactes (observations sans bruit de mesure) $\mathbf{f}$ suivent la loi Normale $\mathbf{f}|\mathcal{P} \sim \mathcal{N}(\mathbf{M}, \mathbf{\Sigma}_{xx})$ et les observations (incluant le bruit de mesure) suivent la loi Normale $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_v^2 I)$.La formulation analytique de la fonction de vraisemblance est alors :

$$p(\mathbf{y}|\mathcal{P}_{gp}, \mathcal{P}_w) = \int p(\mathbf{y}|\mathbf{f}, \mathcal{P}, \sigma_v) \cdot p(\mathbf{f}|\mathcal{P}) \tag{5.10}$$

$$\log(p(\mathbf{y}|\mathcal{P}_{gp}, \mathcal{P}_w)) = -\frac{1}{2}\mathbf{y}^\intercal \mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\log(|\mathbf{\Sigma}|) - \frac{N}{2}\log(2\pi). \tag{5.11}$$

Il a été supposé que les coordonnées virtuelles $\mathcal{P}_w = \{w_{2X}, w_{3X}, w_{3Y}\}$ sont décrites par la distribution normale multivariée $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_{XY})$. Il en résulte que la fonction de vraisemblance des hyper-paramètres $\mathcal{P}_w$ suit cette même loi.

$$p(\mathcal{P}_w|\mathcal{P}_f) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_{XY}) \tag{5.12}$$

Les distributions à priori des hyper et hyper-hyper-paramètres sont plus délicates à estimer. En effet, il n'y a aucune connaissance à priori de ces paramètres puisque le comportement

étudié au laboratoire n'est modélisé par aucune loi physique. De plus, en considérant le faible nombre de spécimens constituant le jeu de données $\mathcal{D}$, les distributions à priori sont essentielles dans la modélisation. En effet, le modèle va tendre vers la connaissance à priori puisqu'il sera difficile d'apprendre les paramètres avec peu de données. La revue de la documentation a mis en évidence cette limitation de l'approche Bayesienne, la distribution à postériori des hyper-paramètres dépend fortement de leur distribution à priori.

Dans le cas où il n'existe aucune connaissance à priori sur les paramètres, l'utilisation de distributions à priori non-informatives est recommandée (Box and Tiao (1992); Syversveen (1998)). En supposant les paramètres indépendants les uns des autres, les distributions à priori conjointes des hyper et hyper-hyper paramètres sont le produit des distributions à priori marginales. Pour des paramètres de location, la distribution à priori non-informative est uniforme, les coordonnées ou les paramètres liés à la moyenne entrent dans cette catégorie. Et pour des paramètres d'échelle, dont les paramètres écart-type et longueurs de corrélation, la distribution à priori non-informative est la fonction inverse du paramètre. On suppose que la distribution à priori de l'hyper-hyper paramètre ρ_{XY} est uniforme sur l'intervalle $[-1, 1]$, sachant que le coefficient de corrélation ne peut prendre de valeurs qu'entre -1 et 1.

$$p(\mathcal{P}_{gp}) \cdot p(\mathcal{P}_f) \propto \frac{1}{\ell_1} \cdot \frac{1}{\ell_2} \cdot \frac{1}{\ell_3} \cdot \frac{1}{\sigma_1} \cdot \frac{1}{\sigma_2} \cdot \frac{1}{\sigma_3} \cdot \frac{1}{\sigma_v} \cdot \frac{1}{\sigma_X} \cdot \frac{1}{\sigma_Y} \tag{5.13}$$

Ces distributions à priori non-informatives permettent ainsi de se détacher de la dépendance de la connaissance a priori dans l'approche Bayesienne. Grâce à la formulation empirique de l'Equation 5.8, il est possible d'échantillonner les hyper-paramètres et hyper-hyper-paramètres $\mathcal{P}_q$ de leur distribution à postériori $p(\mathcal{P}_{gp}, \mathcal{P}_w, \mathcal{P}_f|\mathcal{D})$. La méthode d'échantillonnage McMC associée à la méthode de Newton-Raphson peut être appliquée (Wang and Neal (2012); Wang (2014)).

5.2.3 Prédiction

La modélisation probabiliste du comportement étudié en laboratoire sur trois spécimens testés est équivalente à prédire le comportement d'un quatrième spécimen non testé. Ainsi, l'objectif du modèle est de prédire pour les valeurs de covariables, contrainte $\mathbf{x}_*$ et coordonnées virtuelles du quatrième spécimen $\mathbf{w}_*$, le comportement du quatrième spécimen. Les coordonnées d'un quatrième spécimen sont échantillonnées Q fois à partir de la distribution

suivante (Figure 5.3),

$$\mathbf{w}_{*,q}|\mathcal{D} \sim p(\mathbf{w}_*|\mathcal{D}) \tag{5.14}$$

$$= \int p(\mathcal{P}_w|\mathcal{P}_f) \cdot p(\mathcal{P}_f) d\mathcal{P}_f \tag{5.15}$$

À partir du jeu de données initial $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, il s'en suit que la prédiction par GPR des valeurs non observées $\mathbf{f}_q$ sous les covariables $\mathbf{x}_*$ et $\mathbf{w}_{*,q}$ est estimée à partir des Q échantillons (Equation 5.8) des hyper et hyper-hyper-paramètres $\mathcal{P}_q$ (Wang, 2014).

$$\mathbf{f}_q|\mathbf{x}_*, \mathcal{D} \sim p(\mathbf{f}|\mathbf{x}_*, \mathcal{D}) \tag{5.16}$$

$$= \iiint p(\mathbf{f}|\mathcal{P}_{gp}, \mathcal{P}_w, \mathcal{D}) \cdot p(\mathcal{P}_{gp}) \cdot p(\mathcal{P}_w|\mathcal{P}_f) \cdot p(\mathcal{P}_f)\, d\mathcal{P}_f\, d\mathcal{P}_w\, d\mathcal{P}_{gp} \tag{5.17}$$

$$\mathbb{E}[\mathbf{f}|\mathbf{x}_*, \mathcal{D}] \approx \frac{1}{Q} \sum_{q}^{Q} \mathbb{E}[\mathbf{f}_q|\mathcal{P}_{gp,q}, \mathcal{P}_{w,q}, \mathbf{w}_*, \mathbf{x}_*, \mathcal{D}] \tag{5.18}$$

$$\mathrm{Cov}[\mathbf{f}|\mathbf{x}_*, \mathcal{D}] \approx \mathrm{Cov}[\mathbb{E}[\mathbf{f}_q|\mathcal{P}_{gp,q}, \mathcal{P}_{w,q}, \mathbf{w}_*, \mathbf{x}_*, \mathcal{D}]] + \frac{1}{Q} \sum_{q}^{Q} \mathrm{Cov}[\mathbf{f}_q|\mathcal{P}_{gp,q}, \mathcal{P}_{w,q}, \mathbf{w}_*, \mathbf{x}_*, \mathcal{D}] \tag{5.19}$$

5.2.4 Structure d'ensemble de la méthode par approche Bayesienne

La Figure 5.4 schématise le réseau Bayesien présenté au cours de cette section, appliqué à l'essai de perméabilité à l'eau. Le réseau Bayesien se construit à partir de noeuds et de flèches qui marquent les liens de causalité entre les éléments. Pour trois valeurs de la covariable x, soit dans le cas d'étude trois valeurs de contraintes $\{x_i, x_j, x_k\}$, les essais de perméabilité fournissent respectivement trois observations $\{y_i^1, y_j^2, y_k^3\}$, avec chaque observation tirée d'un spécimen distinct. Le jeu de données $\mathcal{D}$ disponible est représenté par les noeuds colorés. Les observations sont la somme de la valeur exacte de perméabilité $\{k_i^1, k_j^2, k_k^3\}$ (double cercle) dépendante de la covariable contrainte x et du bruit de mesure v. D'après l'hypothèse de GPR, les observations suivent une loi normale multivariée $\mathcal{N}(\mathbf{M}, \mathbf{\Sigma})$. La structure à priori de GPR décrite par le vecteur moyenne $\mathbf{M}$ et la matrice de covariance $\mathbf{\Sigma}$ est constituée d'hyper-paramètres et d'hyper-hyper-paramètres (Equation 5.2) à inférer à partir de leur distribution conjointe à postériori, elle-même construite par une formulation empirique de Bayes (Equation 5.8). Les cadres en pointillé englobant les différents paramètres soulignent la hiérarchie proposée dans la formulation empirique. Ainsi, en partant de bas en haut, les hyper-hyper-paramètres $\mathcal{P}_f$ permettent de déterminer les hyper-paramètres soient les coordonnées

virtuelles $\mathcal{P}_w$. L'ensemble des hyper-paramètres constituant la structure à priori du processus Gaussien, associé au jeu de données $\mathcal{D}$ permet par la méthode GPR de prédire la perméabilité k_*^4 pour une contrainte cible x_* et un quatrième spécimen non testé.

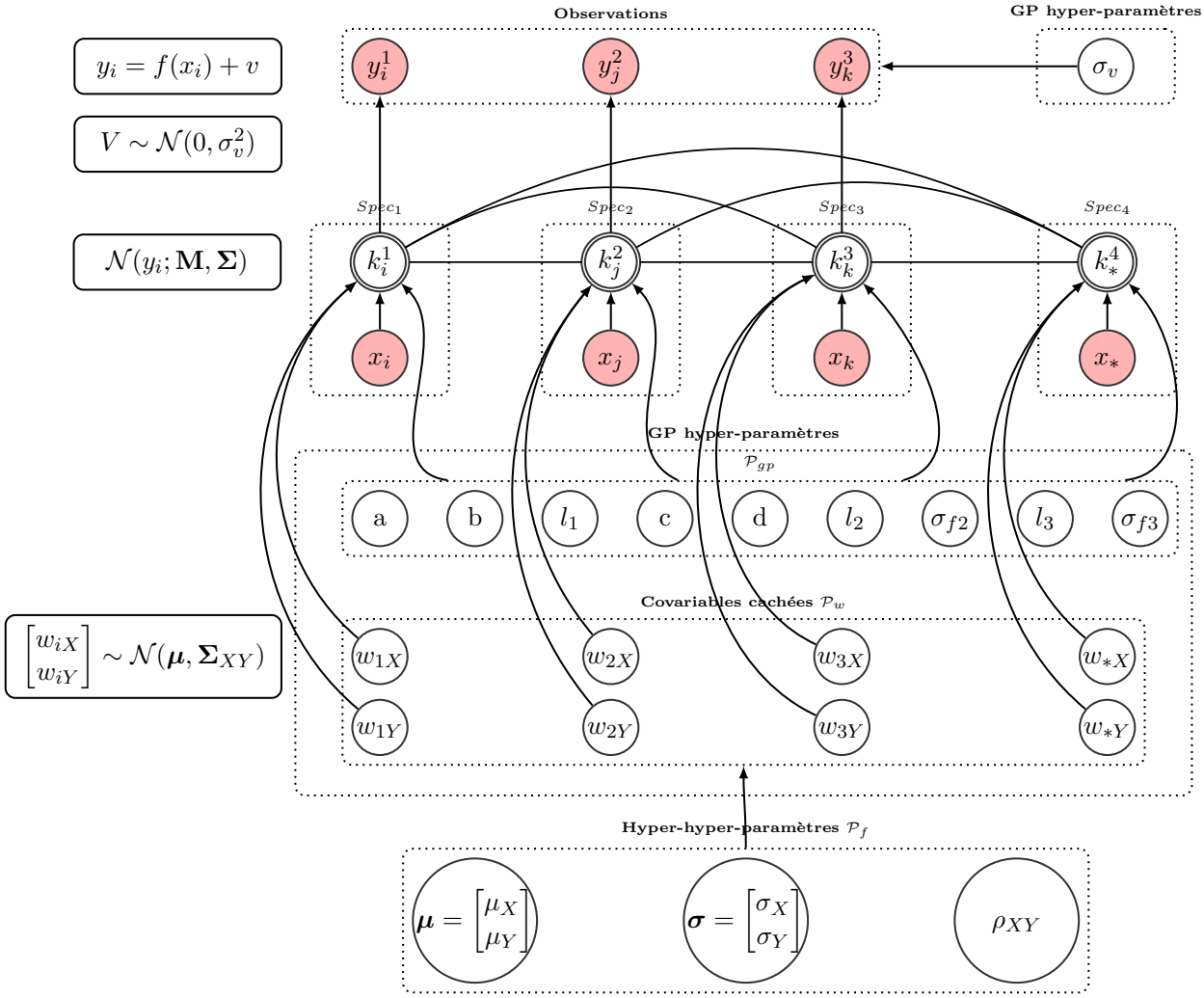


Figure 5.4 s

5.2.5 Conclusion

L'application de cette méthode au jeu de données tiré des essais de perméabilité sur le BFHP n'a malheureusement pas fourni de résultats cohérents. Dû au jeu de données pauvre en spécimens, la distribution à postériori des paramètres est fortement dépendante de leur connaissance à priori. L'hypothèse des distributions à priori non-informatives est une bonne solution, puisqu'elle n'apporte pas d'information sur les paramètres. Cependant, les valeurs

initiales des hyper et hyper-hyper-paramètres prises pour débuter la méthode d'échantillonnage joue tout autant un rôle important, et comme la distribution à priori, la distribution à postériori en est très dépendante. Un jeu de données plus riche permettrait de surmonter cette limite de dépendance et de se détacher de la connaissance à priori des hyper et hyper-hyper-paramètres.

5.3 Discussion générale : méthode par approche Bayesienne

Cette section permettra de mettre en lumière les avantages et perspectives de la méthode par approche Bayesienne présentée à la Section 5.2 précédente, tout en présentant ses quelques limites. Enfin, les deux méthodes développées au sein de ce projet seront comparées afin d'expliciter leurs cas d'utilisation.

5.3.1 Une approche interprétable physiquement

L'approche spatiale de l'hétéroscédasticité offre une visualisation concrète de la variabilité inter-spécimens et offre une compréhension physique de cette variabilité. En effet, dans le cas d'étude, chaque spécimen est représenté par un point dans un espace 2-D et chaque distance séparant deux spécimens modélise la variation de paramètres physiques difficiles à contrôler, telles que la disposition et la taille des granulats ou l'orientation des fibres. Ces distances virtuelles quantifient la variabilité entre deux spécimens. Ainsi, cette approche spatiale peut s'adapter à tout problème physique puisque les paramètres peuvent être interprétés physiquement. La caractérisation spatiale des spécimens permet de réduire le nombre d'étapes dans la méthode. En effet, au sein de la formulation de la fonction de covariance (Equation 2.22), la variabilité entre les observations selon la covariable $\mathbf{x}$ par spécimen et la variabilité inter-spécimens sont modélisées. La dépendance des observations est prise en compte dans les variables latentes, chaque observation est associée aux coordonnées fictives des spécimens testés.

5.3.2 Vers de plus grandes dimensions

L'approche spatiale de l'hétéroscédasticité offre des perspectives de développement. En gardant l'exemple du cas d'étude de la perméabilité à l'eau dans les BHP et BFHP. Pour un taux de fibres donné les trois spécimens testés forment un plan (Figure 5.1) dans l'espace (w_X,w_Y) et les distances séparant les spécimens représentent la variabilité des paramètres physiques non identifiés ou mesurés. Il est alors possible d'ajouter la covariable taux de fibres % en tant que troisième dimension et ainsi former un espace 3-D (w_X,w_Y,%) contenant les trois plans

formés par les trois spécimens spécifiques au taux de fibre testé. La Figure 5.5 illustre la disposition des neuf spécimens formant trois plans distincts dans l'espace à trois dimensions. Les neufs spécimens étant alors dans un même espace sont séparés un à un par des distances virtuelles. Ces distances pourraient être inférées et utilisées afin de prédire le comportement d'un spécimen pour un taux de fibres non testé en laboratoire. Ce début d'approche reste à développer, puisque tel que démontré l'inférence des hyper-paramètres soient des distances ou coordonnées virtuelles est délicate.

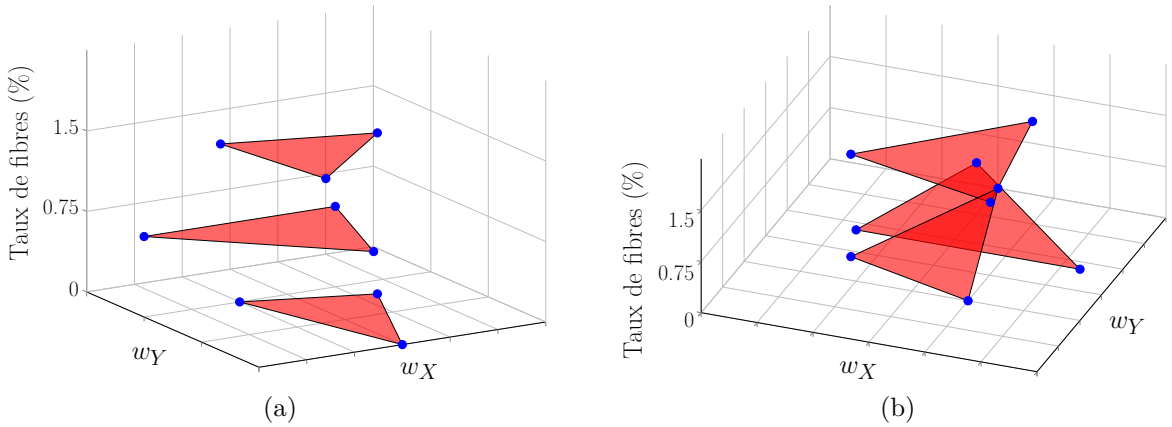


Figure 5.5 Distribution spatiale des neufs spécimens testés en fonction de la covariable taux de fibres

5.3.3 Limite de la méthode par approche Bayesienne

L'approche Bayesienne pour inférer les paramètres est une méthode intéressante, puisqu'elle tient compte des incertitudes sur les paramètres. Mais cette approche requiert d'échantillonner les hyper-paramètres car il n'existe pas de formulation simple de la distribution prédictive des observations. Les méthodes d'échantillonnage sont lourdes et coûteuses en temps. De plus, l'importance du jeu de données étudié a un fort impact sur les résultats de cette approche, tel que dans le cas d'étude. Ce jeu de données ne comprend que trois spécimens, étant donné que chaque taux de fibre est étudié individuellement. Le nombre d'hyper-paramètres et d'hyper-hyper-paramètres est très grand en comparaison au faible nombre de spécimens. Alors la dépendance de la distribution à postériori des paramètres à leur distribution à priori est très forte et vient biaiser le modèle prédictif. Sans un jeu de données riche, l'approche Bayesienne ne peut se détacher de la connaissance à priori. Le jeu de données limité à trois spécimens n'a pas pu fournir de résultats cohérents. Il serait intéressant de tester quelques autres spécimens dans les mêmes conditions d'essai pour augmenter le jeu de données et ainsi estimer le nombre de spécimens nécessaires à la réalisation performante de cette approche Bayesienne.

5.3.4 Comparaison des deux approches

La première méthode basée sur GPR et les distributions à priori conjuguées est une méthode rapide et performante qui a mené à de très bons résultats. A partir d'un jeu de données de quelques spécimens présentant un comportement hétéroscédastique, cette méthode est capable de mettre en évidence une tendance et de fournir un modèle prédictif du comportement. Elle est plutôt rapide et adaptable à tout essai. L'inférence des hyper-paramètres par la méthode MLE est fiable puisque de nombreuses observations par spécimen sont disponibles, elle est un bon compromis face à l'approche Bayesienne qui est plus exacte mais plus lourde. Cependant, cette première méthode par processus Gaussien n'offre pas de perspective de développement, ajouter une seconde covariable telle que le taux de fibres en données d'entrée nécessiterait un nombre de valeurs de taux de fibres aussi grand que le nombre de valeurs de contraintes pour obtenir des résultats comparables. La seconde méthode par approche Bayesienne inclut une représentation spatiale des spécimens et permettrait de prédire le comportement d'un spécimen pour un taux de fibres non testé et offre donc plus de perspectives d'application. Mais cette approche requiert d'inférer un grand nombre de paramètres et seul un jeu de données initial plus large le permet, que ce soit avec la méthode MLE ou l'approche Bayesienne. Finalement, la première méthode par processus Gaussien est fortement recommandée, facile d'accès aux ingénieurs de tous les domaines, elle permet de modéliser efficacement des essais en laboratoire menés sur peu de spécimens à comportement hétéroscédastique. Pour des modèles plus poussés, avec plusieurs covariables, la deuxième approche propose des pistes de recherche intéressantes.

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS

Ce chapitre conclut ce mémoire. Il revient sur la problématique et les différents objectifs posés en début de ce projet et sur les différentes approches proposées afin de répondre à ces derniers. De cette synthèse seront tirées diverses conclusions qui seront complétées par les limites de la solution proposée. Enfin, quelques recommandations et possibilités d'amélioration pour de futures recherches seront présentées.

6.1 Synthèse des travaux

Les essais expérimentaux requièrent de grandes ressources financières, c'est pourquoi seuls quelques spécimens répliqués viennent à être testés lors des programmes expérimentaux. Il est difficile et délicat de tirer un modèle de comportement à partir de seulement quelques spécimens, de plus il est commun que les spécimens testés présentent un comportement hétéroscédastique ou qu'il existe une variabilité inter-spécimens. En effet, il est particulièrement difficile de reproduire des spécimens à l'identique et encore plus difficile de contrôler chaque paramètre physique intervenant dans la fabrication de ceux-ci, ainsi les comportements observés sur des spécimens répliqués, suite aux essais en laboratoire, peuvent varier. La modélisation probabiliste est une méthode s'appuyant sur les données expérimentales permettant de rendre compte et confirmer les comportements observés par la prédiction d'un spécimen non testé. La problématique de ce projet était de développer une méthode probabiliste permettant de modéliser des essais en laboratoire à comportement hétéroscédastique avec seulement quelques spécimens testés. Les principales difficultés étant de modéliser l'incertitude liée au faible nombre de spécimens formant le jeu de données, ou encore la dépendance des observations et en même temps l'hétéroscédasticité. Aussi, cette recherche a eu pour objectif d'appliquer cette méthode à un cas concret, plus précisément aux essais de perméabilité à l'eau sur les BHP et les BFHP et ainsi de confirmer les tendances observées expérimentalement et de quantifier le bénéfice d'incorporation de fibres dans le béton. Pour répondre à la problématique, deux approches ont été proposées au cours de ce mémoire. La première approche consiste en la prédiction de chaque spécimen au sein d'une même régression par processus Gaussien (GPR) puis de s'appuyer sur les formulations analytiques des distributions à priori conjuguées afin de décrire la variabilité inter-spécimens. La seconde méthode est une formulation plus générale par approche Bayesienne empirique, dans ce cas l'hétéroscédasticité est représentée et décrite spatialement. Cette seconde approche est toujours en cours de développement et n'a pu fournir de résultats quant à son application au cas d'étude

proposé au sein de ce projet. En revanche, l'application de la méthode par processus Gaussien a fourni de très bons résultats prometteurs. Il en a résulté que l'ajout de fibres dans le béton conduit à diminuer la perméabilité et les probabilités sont très grandes de réduire la perméabilité dans le béton en augmentant le volume de fibres dans sa formulation.

6.2 Conclusions

Cette section rend compte des conclusions tirées des travaux et résultats de cette recherche, présentant à la fois les forces des méthodes proposées et leurs limites.

- La méthode par processus Gaussien est performante, rapide et peu coûteuse dans la modélisation d'essais en laboratoire effectués sur peu de spécimens à comportement hétéroscédastique ;
- La méthode par processus Gaussien est applicable à tout type d'essai et adaptable dans tous les domaines de génie ;
- L'inférence des hyper-paramètres par estimation par maximum de vraisemblance (MLE) présente quelques limites, car elle peut s'arrêter à un maximum local de la fonction de vraisemblance ;
- L'application de la méthode par processus Gaussien a permis de modéliser la perméabilité à l'eau dans les BHP et BFHP en fonction de la contrainte appliquée dans la barre d'armature du tirant ;
- La méthode par processus Gaussien a permis de quantifier le bénéfice d'incorporation de fibres dans le béton. En comparaison aux BHP, les probabilités de réduire la perméabilité à l'eau chez les BFHP à 0.75% et 1.5% de fibres sont respectivement de 0.91 et 0.93 ;
- Il est statistiquement démontré que l'ajout de fibres réduit la perméabilité dans le béton et qu'il s'agit d'une méthode efficace pour prolonger la durabilité de structures en béton ;
- La méthode par processus Gaussien ne permet pas de prédire le comportement de la perméabilité pour un taux de fibres non testé ;
- La méthode par approche Bayesienne est plus précise dans l'inférence des hyper-paramètres grâce à sa formulation hiérarchique ;
- La méthode par approche Bayesienne offre la possibilité d'ajouter la covariable taux de fibres dans la modélisation de la perméabilité à l'eau grâce à sa représentation spatiale de l'hétéroscédasticité ;

- L'approche Bayesienne et la nécessité d'utiliser une méthode d'échantillonnage rend la méthode plus lourde et plus coûteuse en temps et programmation ;
- Le manque de spécimens testés limite l'inférence des hyper-paramètres, l'approche Bayesienne ne peut se détacher de la connaissance à priori ;

6.3 Recommandations

- Dans le cadre de ce projet, la méthode par processus Gaussien est une méthode de modélisation efficace comparativement à la méthode par approche Bayesienne qui requiert plus de développement. La méthode par processus Gaussien peut s'appliquer à tout essai et dans tout domaine, elle permet de modéliser les observations en fonction d'une ou plusieurs covariables dont les valeurs testées sont nombreuses. Il appartient à l'ingénieur d'adapter la fonction de covariance au comportement étudié afin de modéliser au mieux la variabilité. Il est donc recommandé de travailler sur la base de la fonction de covariance exponentielle carré et de la modifier au besoin. La comparaison de différentes fonctions de covariance se fait par l'estimation du maximum de la fonction de vraisemblance. En parallèle, les hyper-paramètres des fonctions covariance et de moyenne à inférer par la méthode MLE ont un gros impact sur le modèle probabiliste prédictif. Afin d'obtenir les hyper-paramètres optimaux et donc le maximum global de la fonction de vraisemblance, il est conseillé de tester plusieurs jeux de valeurs initiales d'hyper-paramètres afin de confirmer la convergence vers un même maximum global.
- La méthode par approche Bayesienne n'a pas pu permettre d'aboutir à des résultats cohérents à cause du grand nombre d'hyper-paramètres à inférer et du jeu de données limité. Ainsi pour des jeux de données plus riches et des modèles plus complexes, la méthode par approche Bayesienne peut devenir plus adéquate puisque plus exacte et offrant plus de possibilités. Dans cette approche Bayesienne, le choix de la distribution à priori reste essentiel, seul un jeu de données étendu permet de se détacher de cette dépendance. Ainsi, cette méthode par approche Bayesienne est recommandée pour des jeux de données à multi-attributs riches en observations ou dont la connaissance à priori du comportement étudié est étendue.
- Dans le cadre des essais de perméabilité il a été montré que pour obtenir une perméabilité plus faible et donc une durabilité prolongée, il est préférable d'incorporer des fibres au béton. En effet, le taux de fibres de 1.5% donnent les meilleurs résultats parmi ceux étudiés. Des études similaires pourraient être conduites à plus grande échelle pour quantifier l'effet bénéfique des fibres sur la durabilité des infrastructures en béton armé.

6.4 Améliorations futures

- La méthode par processus Gaussien a permis de modéliser la perméabilité à partir de trois spécimens et d'atteindre les objectifs fixés. Aussi, la méthode peut être optimisée en connaissant le nombre de spécimens nécessaires à une modélisation optimale. Dans le cas d'application de ce projet, la simulation d'un quatrième spécimen aurait indiqué si tester un autre spécimen aurait pu améliorer le modèle de façon significative.
- Il serait avantageux d'implémenter un code simple d'utilisation ou une application qui prendrait les données d'essais en laboratoire et en ressortirait les modèles. La simplicité et l'efficacité de son utilisation permettrait aux scientifiques d'obtenir un modèle probabiliste facilement. Et ceci combiné à l'idée d'identifier le nombre de spécimens nécessaires à la modélisation permettrait à chaque spécimen testé de produire un modèle prédictif et de définir si il est requis de tester un autre spécimen.
- Sachant que la méthode par approche Bayesienne reste à développer et montre un grand potentiel, elle est sujette à de nombreuses améliorations. Il s'agit d'établir une distribution à priori adéquate dans la formulation de Bayes, d'optimiser la méthode d'échantillonnage, de décrire une méthode simple pour définir les hyper-paramètres et hyper-hyper-paramètres initiaux.
- La méthode par approche Bayesienne peut être étendue en nombre de dimensions, c'est-à-dire ajouter des covariables indépendamment de leur nombre de valeurs. Cette extension pourrait être très intéressante, en effet elle permettrait de prédire le comportement de la perméabilité pour un taux de fibres non testé, par exemple 0.5% ou 1% de fibres.

RÉFÉRENCES

A. K. Bansal et P. Aggarwal, "Bayes prediction for a heteroscedastic regression superpopulation model using balanced loss function", *Communications in Statistics—Theory and Methods*, vol. 36, no. 8, pp. 1565–1575, 2007.

N. Banthia et A. Bhargava, "Permeability of stressed concrete and role of fiber reinforcement", *ACI materials journal*, vol. 104, no. 1, pp. 70–76, 2007.

C. G. Berrocal, K. Lundgren, et I. Löfgren, "Influence of steel fibres on corrosion of reinforcement in concrete in chloride environments : A review", dans *7th International Conference Fibre Concrete 2013 Proceedings*, 2013.

G. Blau, M. Lasinski, S. Orcun, S.-H. Hsu, J. Caruthers, N. Delgass, et V. Venkatasubramanian, "High fidelity mathematical model building with experimental data : A bayesian approach", *Computers & Chemical Engineering*, vol. 32, no. 4, pp. 971–989, 2008.

G. Box et G. Tiao, *Bayesian inference in statistical analysis*. New-York : Wiley, 1992.

J.-P. Charron, C. Desmettre, et C. Androuet, "Utilisation de béton renforcé de fibres (brf) pour les glissières en béton pour chantier (gbc)", 2016.

A. Der Kiureghian et O. Ditlevsen, "Aleatory or epistemic ? does it matter ?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.

C. Desmettre et J.-P. Charron, "Water permeability of reinforced concrete with and without fiber subjected to static and constant tensile loading", *Cement and Concrete Research*, vol. 42, no. 7, pp. 945–952, 2012.

M. Ebden, "Gaussian processes for regression : A quick introduction", 2008.

E. Erlandson. (2016, March) Computing simplex vertex locations from pairwise object distances. En ligne : `http://erikerlandson.github.io/blog/2016/03/26/computing-simplex-vertex-locations-from-pairwise-vertex-distances/`

P. Gardoni, A. Der Kiureghian, et K. Mosalam, "Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations", *Journal of Engineering Mechanics*, vol. 128, no. 10, pp. 1024–1038, 2002.

P. Gardoni, K. Nemati, et T. Noguchi, “Bayesian statistical framework to construct probabilistic models for the elastic modulus of concrete”, *Journal of Materials In Civil Engineering*, vol. 19, no. 10, pp. 898–905, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, et D. B. Rubin, *Bayesian data analysis*. Taylor & Francis, 2014, vol. 2.

P. W. Goldberg, C. K. Williams, et C. M. Bishop, “Regression with input-dependent noise : A gaussian process treatment”, *Advances in neural information processing systems*, vol. 10, pp. 493–499, 1997.

N.-D. Hoang, A.-D. Pham, Q.-L. Nguyen, et Q.-N. Pham, “Estimating compressive strength of high performance concrete with gaussian process regression model”, *Advances in Civil Engineering*, vol. 2016, 2016.

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

J.-S. Jeon, A. Shafieezadeh, et R. DesRoches, “Statistical models for shear strength of rc beam-column joints using machine-learning techniques”, *Earthquake Engineering & Structural Dynamics*, vol. 43, no. 14, pp. 2075–2095, 2014.

K. Kersting, C. Plagemann, P. Pfaff, et W. Burgard, “Most likely heteroscedastic gaussian process regression”, dans *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 393–400.

J. Lampinen et A. Vehtari, “Bayesian approach for neural networks—review and case studies”, *Neural networks*, vol. 14, no. 3, pp. 257–274, 2001.

Q. V. Le, A. J. Smola, et S. Canu, “Heteroscedastic gaussian process regression”, dans *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 489–496.

Y. Ma, L. Wang, J. Zhang, Y. Xiang, et Y. Liu, “Bridge remaining strength prediction integrated with bayesian network and in situ load testing”, *Journal of Bridge Engineering*, vol. 19, no. 10, p. 04014037, 2014.

D. MacKay, “Introduction to gaussian processes”, *NATO ASI Series F Computer and Systems Sciences*, vol. 168, pp. 133–166, 1998.

L. Martinie et N. Roussel, “Simple tools for fiber orientation prediction in industrial practice”, *Cement and Concrete research*, vol. 41, no. 10, pp. 993–1000, 2011.

K. P. Murphy, “Conjugate bayesian analysis of the gaussian distribution”, p. 16, 2007.

K. Murphy, *Machine learning : a probabilistic perspective*. The MIT Press, 2012.

M. Pal et S. Deswal, “Modeling pile capacity using support vector machines and generalized regression neural network”, *Journal of geotechnical and geoenvironmental engineering*, vol. 134, no. 7, pp. 1021–1024, 2008.

T. Plagué, C. Desmettre, et J.-P. Charron, “Influence of fiber type and fiber orientation on cracking and permeability of reinforced concrete under tensile loading”, *Cement and Concrete Research*, vol. 94, pp. 59–70, 2017.

P.Mahesh et S. Deswal, “Modelling pile capacity using gaussian process regression”, *Computers and Geotechnics*, vol. 37, no. 7, pp. 942–947, 2010.

A. N.-O. Quach, L. Tabor, D. Dumont, B. Courcelles, et J.-A. Goulet, “A machine learning approach for characterizing soil contamination in the presence of physical site discontinuities and aggregated samples”, *Advanced Engineering Informatics*, vol. 33, pp. 60–67, 2017.

J. Rapoport, C.-M. Aldea, S. P. Shah, B. Ankenman, et A. Karr, “Permeability of cracked steel fiber-reinforced concrete”, *Journal of materials in civil engineering*, vol. 14, no. 4, pp. 355–358, 2002.

C. Rasmussen et H. Nickisch, “Gaussian processes for machine learning (gpml) toolbox”, *The Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

C. Rasmussen et C. Williams, “Gaussian processes for machine learning”, *the MIT Press*, vol. 2, no. 3, p. 4, 2006.

P. Rossi, “Bétons de fibres métalliques (bfm)”, *Techniques de l’ingénieur. Construction*, vol. 2, no. C2214, pp. C2214–1, 1998.

R. Siddique, P. Aggarwal, Y. Aggarwal, et S. Gupta, “Modeling properties of self-compacting concrete : support vector machines approach”, *Computers and Concrete*, vol. 5, no. 5, pp. 123–129, 2008.

M. Słoński, “A comparison of model selection methods for compressive strength prediction of high-performance concrete using neural networks”, *Computers & structures*, vol. 88, no. 21,

pp. 1248–1253, 2010.

M. Sloński, “Bayesian neural networks and gaussian processes in identification of concrete properties”, *Computer Assisted Mechanics and Engineering Sciences*, vol. 18, no. 4, pp. 291–302, 2011.

A. R. Syversveen, “Noninformative bayesian priors. interpretation and problems with construction and applications”, *Preprint Statistics*, vol. 3, 1998.

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

M. K. Titsias et M. Lázaro-Gredilla, “Variational heteroscedastic gaussian process regression”, dans *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 841–848.

V. Tolvanen, P. Jylänki, et A. Vehtari, “Expectation propagation for nonstationary heteroscedastic gaussian process regression”, dans *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2014, pp. 1–6.

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, et A. Vehtari, “Bayesian modeling with gaussian processes using the gpstuff toolbox”, *arXiv preprint arXiv :1206.5754*, 2012.

C. Wang, “Gaussian process regression with heteroscedastic residuals and fast mcmc methods”, Thèse de doctorat, University of Toronto, 2014.

C. Wang et R. M. Neal, “Gaussian process regression with heteroscedastic or non-gaussian residuals”, *arXiv preprint arXiv :1212.6246*, 2012.

I. C. Yeh, “Estimating distribution of concrete strength using quantile regression neural networks”, *Applied Mechanics and Materials*, vol. 584, p. 1017, 2014.

J. Zhong, P. Gardoni, D. Rosowsky, et T. Haukaas, “Probabilistic seismic demand models and fragility estimates for reinforced concrete bridges with two-column bents”, *Journal of engineering mechanics*, vol. 134, no. 6, pp. 495–504, 2008.