Surrogate modeling of stochastic simulators

Soumaya Azzi

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAT009

Thèse de doctorat

Surrogate modeling of stochastic simulators

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 4 Juin 2020, par

SOUMAYA AZZI

Composition du Jury :

Jean-Marc Bourinet Institut Pascal, SIGMA Clermont	Rapporteur
Philippe De Doncker Université Libre de Bruxelles	Rapporteur
Laurent Decreusefond Télécom Paris	Examinateur
Alain Sibille Télécom Paris	Examinateur
Martine Liénard Université de Lille	Examinatrice, Présidente
Joe Wiart Télécom Paris	Directeur de thèse
Bruno Sudret ETH Zürich	CoDirecteur de thèse

Acknowledgements

First and foremost, I would like to address my special thanks to both my supervisors, Joe Wiart and Bruno Sudret, for the support and encouragement over these three years. Your advice has been utterly priceless and genuinely helped me become the researcher I am today.

My warmest thanks go to everyone who accepted to review this work and go through every single page of it, namely my supervisors and the jury members.

I would like to thank my colleagues at the chair C2M for the enlightening conversations, for their help and the support, and the many meals and the laughs we shared together. You made the three years of my PhD delightful and the hard days bearable. I will never forget you, Taghrid, Yuanyuan, Amirezza, Zicheng, Mauricio, Bader, Maarouf, Xi, and Shanshan.

My gratitude also goes to people I met in Télécom Paris. Thank you Zachary for proofreading my paper. Thank you Bethany for proofreading my PhD thesis and for always being such a sweetheart. I also want to address my gratitude to Yvonne, Chantal and Hamidou; you make our department the best. I would like to thank my referent F.Suchanek and professor O.Rioul for the fruitful discussions we had. Finally, I am eternally grateful to my fellow PhD students: Elaine, Maciel, Sahar, Etienne and Natasha. It's been wondering to live this experience with you and to soak in your glow and warmth.

As this PhD is not only the achievement of the last three years, I would like to thank all my teachers and mentors, especially those who inspired me to love math and research. You are the ones I looked up to and admired while growing up.

Through various conferences I had the chance to speak at and attend, I met lovely people that became life-long friends. Through the span of my life I had the chance to meet precious friends, to whom I address my gratitude, *les amis, c'est la vie*. That was a blast.

And finally, I would like to thank my family, for always willing to go the extra mile for me, in particular my dear siblings, for the lovely letters you wrote to me saying how much you miss me, and my loving parents, to whom I dedicate this work.

Abstract

This thesis is a contribution to the surrogate modeling and the sensitivity analysis on stochastic simulators. Stochastic simulators are a particular type of computational models, they inherently contain some sources of randomness and are generally computationally prohibitive. To overcome this limitation, this manuscript proposes a method to build a surrogate model for stochastic simulators based on Karhunen-Loève expansion.

This thesis also aims to perform sensitivity analysis on such computational models. This analysis consists on quantifying the influence of the input variables onto the output of the model. In this thesis, the stochastic simulator is represented by a stochastic process, and the sensitivity analysis is then performed on the differential entropy of this process.

The proposed methods are applied to a stochastic simulator assessing the population's exposure to radio frequency waves in a city. Randomness is an intrinsic characteristic of the stochastic city generator. Meaning that, for a set of city parameters (e.g. street width, building height and anisotropy) does not define a unique city. The context of the electromagnetic dosimetry case study is presented, and a surrogate model is built. The sensitivity analysis is then performed using the proposed method.

Résumé

Cette thèse propose des outils statistiques pour étudier l'impact qu'a la morphologie d'une ville sur l'exposition des populations induite par un champ électromagnétique provenant d'une station de base. Pour cela l'exposition a été évaluée numériquement en propageant (via des techniques de lancer de rayons) les champs émis dans une antenne dans des villes aléatoires. Ces villes aléatoires ont les mêmes caractéristiques macroscopiques (e.g. hauteur moyenne des immeubles, largeur moyenne des rues et anisotropie) mais sont distinctes les unes des autres. Pour les mêmes caractéristiques de nombreuses villes aléatoires ont été générées et l'exposition induite a été calculée pour chacune. Par conséquent, chaque combinaison de variables correspond à plusieurs valeurs d'exposition. L'exposition est décrite par une distribution statistique non nécessairement gaussienne. Ce comportement stochastique est présent en plusieurs problèmes industriels et souvent les nombreuses simulations menées ont un cout de calcul important.

Les travaux de cette thèse étudient la modélisation de substitution des fonctions aléatoires. Le simulateur stochastique est considéré comme un processus stochastique. On propose une approche non paramétrique basée sur la décomposition de Karhunen-Loève du processus stochastique. La fonction de substitution a l'avantage d'être très peu coûteuse à exécuter et à fournir des prédictions précises.

En effet, l'objective de la thèse consiste à évaluer la sensibilité de l'exposition aux caractéristiques morphologiques d'une ville. On propose une approche d'analyse de sensibilité tenant compte de l'aspect stochastique du modèle. L'entropie différentielle du processus stochastique est évaluée et la sensibilité est estimée en calculant les indices de Sobol de l'entropie. La variance de l'entropie est exprimée en fonction de la variabilité de chacune des variables d'entrée.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Context of the thesis

The wireless technology brought people in a much closer world in the last three decades. Communication through mobile phones for example is quite simple, therefore attracting more and more users, wherever they may be. The number of phone users in France in 2019 is estimated to more than 51 millions [2]. This number keeps increasing, especially with the emergence of more connected devices and smart environments.

In parallel with the widespread use of wireless systems, an increased risk perception related to radio-frequency electromagnetic fields (RF-EMF) has been observed [63], and the assessment of the human exposure to RF-EMF has aroused social attention. To respond to such concerns, large efforts have been carried out to establish methods to verify compliance with exposure limits. The human EMF exposure is quantified in terms of Specific Absorption Rate (SAR) expressed in W/kg and representing the RF power absorbed per unit of mass of biological tissues.

As a matter of fact, the RF-EMF sources are the combination of uplink and downlink radiations coming from, respectively, personal wireless devices (e.g. smartphones or tablets) and cellular base stations or access points. In this respect, advanced computational propagation tools were used in many studies [40, 94, 39] to characterize the signal attenuation between a transmitter and a receiver. Such tools can provide accurate path loss results, however they are strongly dependent on detailed building and terrain data.

Stochastic geometry has proven its ability to describe the complex structures of a city [25] via a limited number of parameters, such as building density, street width, number of intersections, etc. Based on statistical distributions of the city features, i.e., building height, street width, anisotropy[1], the stochastic geometry simulator developed in [25] was used to generate various random 3D cities. Figure 1.1 illustrates various city samples generated with the same morphological features.

The objective is to explore the link between the exposure and the city parameters. To this aim a 3-D ray launching technique [93] based on propagation mechanisms such as reflections and diffractions, commonly used to propagate EMF in urban areas, is implemented in the virtual city generated using stochastic geometry. This so-called *ray tracing* technique depends on the digital geographical map extracted from the real environment, allowing for an accurate estimation of the path loss between the base

[1]The anisotropy defines the street system (street angle). This parameter goes continuously from 0 to 1 (1 for a Manhattan-like city).

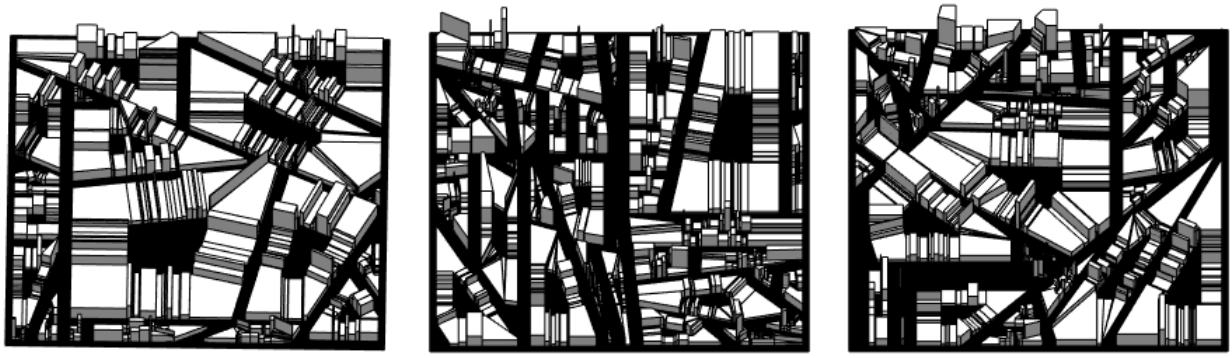


Figure 1.1: Examples of 3D stochastic city models with identical values of morphological features (street width = 13 m, building height=16 m and anisotropy =0.6).

station antenna and the wireless device. The emitted and received power can then be estimated and used to evaluate the EMF exposure. A limit of such a technique is the very high computational cost due to the use of complex deterministic propagation models.

When evaluating the exposure, the focus is on the path loss exponent (PLE) that represents the attenuation of the energy between the transmitter and the receiver. It depends on the transmitter characteristics and the propagation environment and can be evaluated following the ray tracing step.

The main goal of this thesis is to evaluate the link existing between the city parameters and the PLE. With the presented context, two obstacles have to be dealt with:

- The relationship between a set of city parameters and a virtual city is not deterministic (e.g. Figure 1.1). The model inherently contains some sources of randomness, mainly because generating the stochastic city, for example the streets architecture makes use of random processes. Consequently, having fixed a set of city parameters: the exposure is not unique, the value is different for each realization of the city. The model is thus referred to as a *stochastic simulator*, and the assessment of PLE over a city can be seen as a random function of the morphological features of this city.

- The stochastic model is computationally prohibitive. Once a virtual city is generated, an antenna is located in the city and millions of rays are launched. The signal attenuation map can thus be obtained by assessing the received power in the 'measurement' plane (1.5 m above the ground to represent the human

exposure). This computation takes more than one hour[2]. The computational burden escalates when multiple runs are needed to evaluate the possible values of exposure in a fixed city.

To overcome the second limitation, a mathematical function called *metamodel* or *surrogate model* is built. It mimics the behaviour of the simulator and runs in a reasonable cost. This surrogate model shall be adapted here to the characteristics of the original model, namely the stochastic nature of the city generator used to evaluate the path loss exponent.

This PhD thesis thesis addresses the problematic as follow:

- A non-parametric method to build a stochastic metamodel for the city generator is built. The original model is considered as a random process, and the method developed builds a surrogate random process emulating at best the original model. This step enables the prediction of the path loss exponent for different cities. It also gets rid of the computational burden limiting the use of the original stochastic model.

- The impact of the city characteristics onto the path loss exponent is evaluated using sensitivity analysis. A method is proposed to spot the most impactful variables among the three considered (street width, building height and anisotropy).

To this aim, a computer experiment on the stochastic city generator was planned. Numerous calls to the simulator were performed and the domain of definition was reasonably explored. For each point in the domain, repetitions were made such that the randomness is also reasonably explored. The design of the experiments has to be planned wisely to cope with the huge global computational costs (several months).

1.2 Surrogate modeling

Building a surrogate model for a deterministic model is quite documented in the literature. The most popular are Gaussian process modeling (a.k.a Kriging) [79], generalized polynomial chaos expansion GPCE [32, 97] and low rank tensor approximations [18, 48, 19]. Metamodeling of stochastic functions is a less mature

[2]by means of a computer type Intel Xeon E5-2620V3 2.4GHz 6Core 15Mo and NVIDIA TESLA K80

field. Assuming that the model output is a Gaussian field trajectory, recent studies [14, 56, 4, 15, 44] build two independent or joint deterministic metamodels to fit the mean and the covariance of the assumed Gaussian process. Also based on the joint metamodeling approach, [44] simultaneously surrogates the mean and the dispersion using two interlinked generalized additive models. Alternatively, the study carried out in [61] focused on projecting the output density on a basis of chosen probability density functions. With this approach, the coefficients are computed by solving constraint optimization problems for the purpose of building a local metamodel. This method is not ideal for assessing certain quantities of interest (e.g., quantiles). The goal is to overcome these limitations, and propose a non-parametric method, based on the Karhunen-Loève expansion, to build a surrogate model of random functions.

1.3 Sensitivity analysis

The path loss exponent to some extent, depends on the features governing the city structure, such as the organization of buildings into blocks, the street intersections and the street network anisotropy. Its variability will be explored by performing sensitivity analysis, which measures how the uncertainty in the output of a model is related to the input variables. In our case, we will investigate how the variability of the PLE is related to the variability of the city parameters.

The most commonly used approach for sensitivity analysis is the variance-based approach where the variance of the output is expanded as a sum of contributions of each input variable, or their combinations [82]. Regression-based measures (like Pearson correlation coefficient) are also used for models with linear behavior. Alternative global SA methods are available such as the Morris method [60] as well as the moment-independent indicators [13]. Concerning stochastic models, the literature is once more less mature. Sensitivity analysis was applied on the mean and the dispersion of the random output [56]. In this case the sensitivity analysis results does not take into account the influence of higher moments of the random variable output. In this thesis, the stochastic simulator is represented as a stochastic process and the sensitivity analysis is performed on the differential entropy of the stochastic process. The performance of the method is also evaluated.

1.4 Objectives and outline of the thesis

Methods developed in this thesis arise from a practical need to build a surrogate model to the heavy stochastic city generator, and also to characterize the impact the city morphological variables have on the exposure. Both issues were addressed by introducing tools from different disciplines namely electromagnetic, dosimetry, stochastic geometry, statistical learning and information theory, to name a few.

Chapter 2 introduces the general idea of statistical learning. The main methods to build metamodels are presented as well as possible post processing steps such as the error evaluation, the cost function and the model validation. Sensitivity analysis is presented next. The main methods used in deterministic contexts are briefly viewed. This chapter summarizes the tools used or mentioned in the rest of the thesis.

Chapter 3 is dedicated to the surrogate modeling of this particular type of computational models called stochastic simulators, which inherently contain some source of randomness. In this particular case the output of the simulator in a given point is a probability density function. The stochastic simulator is represented as a stochastic process and the surrogate model is build using the Karhunen-Loève expansion. In a first approach, the stochastic process covariance is surrogated using polynomial chaos expansion, meanwhile in a second approach the eigenvectors are interpolated. The performance of the method is illustrated on a toy example. Means to measure the accuracy of the surrogate are also provided.

In Chapter 4, the interest is to quantify the sensitivity of the random output to the model input variables. This is achieved by reducing the output random variable to its differential entropy. Thus instead of considering the sensitivity of the stochastic model, the sensitivity of the differential entropy of the stochastic model is considered. In practice, following the sampling of the stochastic model on a predefined design of experiments, differential entropy is evaluated on each DoE point. The next step consists of building a surrogate model of the differential entropy of the stochastic process to then apply standard methods of sensitivity analysis (SA), in this case, via evaluating Sobol' indices [82].

Chapter 5 describes in details the case study at hand i.e. the human exposure in cities, the evaluation of the exposure using ray-tracing as well as the experiences planned and realized to sample the stochastic simulator. Samples from the stochastic city generator are drawn and the exposure is evaluated. Based on the data collected, a metamodel of the path loss exponent is built and the sensitivity analysis is performed

following methods introduced in chapters 3 and 4. The results are described and interpreted.

At the end, a conclusion sums up the main contributions of the thesis and enumerates the main prospects to state from this work.

Chapter 2

Uncertainty quantification on deterministic models

Contents

2.1 Statistical learning

Parallel to the advances in fields ranging from biology to finance to electrodynamics to astrophysics, vast and complex data sets have emerged. Statistical learning refers to tools and methods for modeling, prediction and classification techniques. For example, let us consider the relation between the wages and age groups of males from somewhere in the world via a data set; we might foresee that the wage increases with age but then decreases again after approximately age 60. The mathematical function describing this relation is unknown, and the statistical learning in this case consists of predicting properties of the unknown function. Consider now a second example from the field of mechanical engineering. The governing equations consist of a set of partial differential equations (PDEs) whose solutions are numerically approximated by finite element methods or finite-difference time-domain methods and solved by computer codes. These codes have reached a high level of sophistication allowing a high accuracy for the PDE solutions at the expense of the computational cost. The unitary computational time typically ranges from minutes to hours or even days for complex systems and high-fidelity models. However, running a costly code thousands to millions of times is not feasible even with high performance computational infrastructures. In this case statistical learning consists of substituting the computational model solving the PDE with a mathematical function that mimics the behavior of the original model, at much cheaper cost. Therefore, by gathering knowledge from experience (data sets, computer codes, etc.), the aim is to allow computers to learn, predict, and infer in the following ways:

- Build a surrogate model (also known as learners, metamodels, interpolators or response surfaces) that accurately mimics the knowledge at hand.
- Use the surrogate model to predict the output for new values of the input parameters. In the event of heavy computational codes, the surrogate model is supposed to have short execution time.
- Perform a sensitivity analysis and identify the most contributing inputs (or set of inputs) that explain, at best, the variability of the output.
- Estimate other aspects of statistical learning including quantities of interest such as confidence intervals, credible intervals, quantiles, failure probabilities, etc.

Suppose we observe a quantitative response (target) t and p different input variables (features), $x_1, x_2, \ldots, x_p$. We assume that there is a relationship between t and $x_1, x_2, \ldots, x_p$, which can be written in the particularly general form

$$t = y(\boldsymbol{x}) + \epsilon, \tag{2.1}$$

where y is an unknown function representing the systematic information that $x_1, x_2, \ldots, x_p$ provides about t, and ϵ is a zero-mean random error term which is independent from the input variables. Eq. (2.1) captures a property of several real data sets, namely that they possess an underlying regularity, which we wish to learn, but that individual observations are corrupted by random noise. Usually, the only available information about the relationship between t and $x_1, x_2, \ldots, x_p$ are n observations. For $i = 1, \ldots, n$, we observe

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{1,1} & \ldots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n,p}, \quad \boldsymbol{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \in \mathbb{R}^n. \tag{2.2}$$

The context presented here (fitting a model that relates the response to the inputs where the aim is to accurately predict the response for future samples) is called supervised learning. For each i^{th} measurement $\{\boldsymbol{x}_i,\ i = 1, \ldots, n\}$ there is an associated response measurement t_i. It is worth mentioning that some statistical learning problems can be unsupervised, meaning that no response t_i is associated to the measurements to supervise the statistical analysis, for example, clustering problems.

The setting handled in this chapter, that each i^{th} measurement $\{\boldsymbol{x}_i,\ i = 1, \ldots, n\}$ is associated to a response measurement t_i, is called deterministic. The model yields a unique output t_i for each set of inputs $\boldsymbol{x}$; In the second chapter we introduce a particular type of model called *stochastic simulators* which, due to additional sources of randomness, run with the same input vector and provide different outputs.

Experimental Design When possible, and within the budget allocated, a sample is drawn and represents the only available information about the model. That sample set is called a design of experiments set (DoE) $DoE = \{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_n, t_n)\}$. In order to make the most of the budget, the DoE requires careful planning [68]. Once the DoE is set, the calls to the simulators can be launched. Among the strategies implemented for computer experiments, examples include Monte Carlo Sampling [17], Latin hypercube sampling (LHS) [59], Sobol' sequences [84], Halton sequences [34], factorial design [30], etc.

Cost functions What is at stake is proposing an estimate $y(\boldsymbol{x})$ of the value of $\boldsymbol{t}$ for each input $\boldsymbol{x}$. A cost (or loss) is attributed to each estimate candidate $y(\boldsymbol{x})$, $J(\boldsymbol{t}, y(\boldsymbol{x}))$. The average cost is given by

$$\mathbb{E}[J] = \int\int J(\boldsymbol{t}, y(\boldsymbol{x}))p(\boldsymbol{x}, \boldsymbol{t})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{t}. \tag{2.3}$$

A common choice of cost function is the squared error loss function $J(\boldsymbol{t}, y(\boldsymbol{x})) = (y(\boldsymbol{x}) - \boldsymbol{t})^2$; the average cost is then given by

$$\mathbb{E}[J] = \int\int (y(\boldsymbol{x}) - \boldsymbol{t})^2 p(\boldsymbol{x}, \boldsymbol{t})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{t}. \tag{2.4}$$

Over-fitting The over-fitting occurs when a surrogate model fits too closely or exactly with a particular set of data. In other words, the surrogate model learns the details and the noise of the training data, therefore, it fails to reliably predict the output for new inputs. In this respect, we rely on measures of goodness-of-fit to detect over-fitting and under-fitting phenomenons.

Model validation Most machine learning algorithms have hyperparameters that shall be estimated. The data set available enables to set those hyperparameters to an adequate value adapted to the data. To avoid the over-fitting problem, a validation set is needed. Specifically, the data is partitioned into k subsets of equal size. At each step a single subsample is retained as the validation set for testing the model (test set), and the remaining data are used to build the surrogate model (training set). This approach is called k-fold cross-validation procedure. The k-fold validation is repeated for several partitions of the data. The error is evaluated using the cost function of choice. When $k = n$, it is called the leave-one-out cross-validation, where only one observation is used to test the goodness of fit at each trial. This technique is mainly used when the dataset is too small (typically when a large data collection is not affordable).

2.2 Regression methods for deterministic models

The goal of regression is to predict the value of a set of target variables t given the value of a p dimensional vector $\boldsymbol{x}$ of input variables based on $n \geq 1$ observations. For $i = 1, \ldots, n$, we observe $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ and $t_i \in \mathbb{R}$ the output.

2.2.1 Linear methods

Linear regression is a very simple approach for supervised learning. It has been around for a long time and is the topic of countless textbooks. It may seem somewhat dull to use linear regression compared to other modern statistical learning approaches, but after all, many of the fancy statistical learning approaches can be seen as generalizations or extensions of linear regression.

Consider the same framework of n observed input variables $\boldsymbol{x}$ and target values $\boldsymbol{t}$ as in Eq. (2.2). To include the bias in the scalar product, a slight change of notation is made:

$$\mathbb{X} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n,p+1} \quad \text{with} \quad x_{i,0} = 1 \ \forall i \in \{1, \dots, n\}. \tag{2.5}$$

The Ordinary least-squares (OLS) estimator is the vector of coefficients $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n,0} \dots \hat{\theta}_{n,p})^T \in \mathbb{R}^{p+1}$ such that

$$\hat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} (t_i - \boldsymbol{x}_i^T \boldsymbol{\theta})^2, \tag{2.6}$$

or in matrix notation

$$\hat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} ||(\boldsymbol{t} - \boldsymbol{x}\boldsymbol{\theta})||^2. \tag{2.7}$$

In this case $y(\boldsymbol{x})$ in Eq .(2.1) is the linear combination $\boldsymbol{x}\boldsymbol{\theta}$. The vector $\hat{\boldsymbol{\theta}}_n$ is such that

$$\mathbb{X}^T \mathbb{X} \hat{\boldsymbol{\theta}}_n = \mathbb{X}^T \boldsymbol{t}. \tag{2.8}$$

The solution is uniquely defined if and only if the Gram matrix $\mathbb{X}^T\mathbb{X}$ is invertible, in which case

$$\hat{\boldsymbol{\theta}}_n = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \boldsymbol{t}, \tag{2.9}$$

otherwise Eq. (2.8) has an infinite number of solutions. Often some constraints are added to the minimization problem: $||\boldsymbol{\theta}||_q \leq s$ [38, 91] where some of the $\theta_{i,j}$ are shrunk to exactly zero, resulting in a regression model that's easier to interpret. The tuning hyperparameter, s controls the strength of the q-norm penalty.

2.2.2 Polynomial methods

Polynomial regression fits a non-linear relationship between $\boldsymbol{x}$ and $\boldsymbol{t}$. Various choices of the polynomial basis are available. Here we present the Wiener-Hermite polynomial chaos expansion (PCE) which is an infinite series expansion of a square-integrable random variable involving orthogonal polynomials basis $\{\Psi_j, j \in \mathbb{N}\}$.

Consider a model with independent input variables gathered in a random vector $\boldsymbol{X}$ with a joint probability density function $p_{\boldsymbol{X}}$. Suppose the corresponding response T is a second-order random variable $\mathbb{E}[T^2] = 0$, then T can be expressed as follows:

$$T \simeq y(\boldsymbol{X}) = \sum_{j=0}^{P-1} a_j \Psi_j(\boldsymbol{X}). \tag{2.10}$$

In practice, the PCE is truncated after P terms. a_j are unknown deterministic coefficients for multi-index j; Ψ_j are multivariate polynomials of the PC basis which are orthogonal with respect to the joint PDF $p_{\boldsymbol{X}}$ of the input random vector $\boldsymbol{X}$. For instance, if the components of the input random vector $\boldsymbol{X}$ follow a uniform distribution over $[-1, 1]$, the orthogonal polynomials of the PC basis are the Legendre polynomials. Table 2.1 above shows the suitable orthogonal polynomials for three examples of input random variables.

Table 2.1: Random variables and corresponding polynomial basis functions.

Distribution	Polynomial basis
Uniform $\mathcal{U}(a, b)$	Legendre
Gaussian $\mathcal{N}(a, b)$	Hermite
Gamma Γ	Laguerre

To determine the coefficients a_j there are two main-stream methods, either using projection methods where the expansion is projected onto the polynomial space, or by casting a least-squares minimization problem.

- **Projection methods**: Multiplying Eq. (2.10) by $\Psi_j(\boldsymbol{X})$ and by taking the expectation, one gets:

$$\mathbb{E}\left[y(\boldsymbol{X})\Psi_j(\boldsymbol{X})\right] = \sum_i a_j \mathbb{E}\left[\Psi_i(\boldsymbol{X})\Psi_j(\boldsymbol{X})\right], \tag{2.11}$$

where $\mathbb{E}\left[\Psi_i(\boldsymbol{X})\Psi_j(\boldsymbol{X})\right] = \mathbb{1}_{\{i=j\}}$. As a consequence of the orthogonormality of the polynomial basis, each coefficient is the projection of the response onto the j-th Ψ_j.

$$a_j = \mathbb{E}\left[y(\boldsymbol{X})\Psi_j(\boldsymbol{X})\right]. \tag{2.12}$$

The calculation of the coefficients is therefore reduced to the calculation of the expectation value i.e. solving the integration problem via quadrature schemes [96].

- **Regression methods**: Assessing the coefficients of the truncated expansion can be cast as a regression problem then solved as a least square minimization problem:

$$\{\hat{a}_j\}_{j=0}^{p-1} = \underset{a_j}{\operatorname{argmin}} ||y(\boldsymbol{X}) - \sum_j a_j \Psi_j(\boldsymbol{X})||^2. \tag{2.13}$$

As in Section 2.2.1 we recover a solution similar to Eq. (2.9).

Sparce PCE A full PCE model is a model where all polynomials with all multi-index $j \leq P$ are considered in the expansion. As the number of input variables increases, the number of configurations of interest grow exponentially. This phenomenon is known as the curse of dimensionality. In sparce PCE approaches [92, 28], only the polynomials among possible candidates $\Psi_j(\boldsymbol{X})$ that have the greatest impact on the model response $y(\boldsymbol{X})$ are selected.

Model validation Using a cross validation procedure, the error can be evaluated using the cost function as in Eq. (2.4) on each validation set. Denoting the test set (input vector and response) as $\boldsymbol{x}_{test}$ and $\boldsymbol{t}_{test}$, respectively, the mean square error of data discrepancy is given as:

$$\epsilon_{test} = \frac{1}{n}\sum_1^n (y(\boldsymbol{x}_{test}) - \boldsymbol{t}_{test})^2, \tag{2.14}$$

for which we associated a coefficient of determination R^2:

$$R^2_{test} = 1 - \frac{\epsilon_{test}}{\mathrm{Var}\left[\boldsymbol{t}_{test}\right]}; \tag{2.15}$$

A value of R^2_{test} of 1 indicates that the predictions perfectly fit the data. With a large test set, R^2_{test} can be obtained by Monte Carlo simulations. Otherwise the cross validation procedure enables to reuse the same data for training and for validation. Leave-one-out cross validation (LOOCV) consists of leaving the i-th observation out for validation. With the remaining data points, a surrogate model is built y^{-i}, and

LOO error is evaluated by repeating the described process for each point in the *DoE* set:

$$\epsilon_{LOO} = \frac{1}{n}\sum_{i=1}^{n}(y^{-i}(\boldsymbol{x}_i) - t_i)^2. \tag{2.16}$$

In the general case, the input variables $\boldsymbol{X}$ can be dependent. Using an isoprobabilistic transform G which is a diffeomorphism from $supp(\boldsymbol{X})$ into $\mathbb{R}^n$ [51, 75], the dependent input $\boldsymbol{X}$ can be mapped into an independent input $U = G(\boldsymbol{X})$.

2.2.3 Kernel methods

A kernel is a symmetric function: $k : (\boldsymbol{x}, \boldsymbol{x}^{'}) \in \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which represents a set of n data points $\boldsymbol{x}_i \in \mathcal{X}, i = 1, \ldots, n$ by the comparison function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Kernel methods are algorithms that take $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as input instead of the original data set. Consequently, kernel methods have the power to handle atypical types of data sets (vectors, strings, graphs, images, etc.). In the next sections we focus on Gaussian processes modeling (Kriging).

Kriging starts with a prior distribution over the covariance of the output $y(\boldsymbol{x})$. It treats the deterministic response of $y(\boldsymbol{x})$ as a particular realisation $\mathcal{F}(\boldsymbol{x}, \omega), \omega \in \Omega$ of a Gaussian stochastic process $\mathcal{F}(\boldsymbol{x})$ such as:

$$\mathcal{F}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{2.17}$$

where $\mu(\boldsymbol{x})$ is the global model mean. $Z(\boldsymbol{x})$ is assumed to be a zero-mean Gaussian random process with the following properties:

$$\mathbb{E}\left[Z(\boldsymbol{x})\right] = 0, \qquad \mathrm{Cov}[Z(\boldsymbol{x}), Z(\boldsymbol{x}')] = \sigma^2 k(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.18}$$

where σ^2 is the process variance and $k(\boldsymbol{x}, \boldsymbol{x}')$ is the correlation function between any two locations $\boldsymbol{x}$ and $\boldsymbol{x}'$ (a kernel). $k(\boldsymbol{x}, \boldsymbol{x}')$ is often defined as a function of the Euclidean distance $h = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ with a set of so-called hyperparameters θ. The kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ maybe represented for instance as a product of univariate correlation functions for each variable as follows:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \prod_{i=1}^{p} k(x_i, x_i') \quad \text{or as:} \quad k(\boldsymbol{x}, \boldsymbol{x}') = R(h), \quad h = \sqrt{\sum_{i=1}^{p}\left(\frac{x_i - x_i'}{\theta_i}\right)^2}. \tag{2.19}$$

Standard correlation functions (kernels) are the Gaussian, exponential, and Matern kernels [79].

Depending on the stochastic properties of the Gaussian process and the various degrees of stationarity assumed, different methods for calculating the hyperparameters of k can be deduced [79].

2.2.4 Artificial neural networks

Another widely used class of metamodels is artificial neural networks (ANN) where the main idea is to mimic the way a brain processes information to learn complex models and predict when new situations occur. They are called *network* because they are typically represented by composing several different functions. We might have, for example, three functions $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$ connected in a chain to form $y(x) = y^{(3)}(y^{(2)}(y^{(1)}(x)))$. $y^{(1)}$ is called the first layer of the network; $y^{(2)}$ is the second layer, and so on. In addition to these layers, a neural network also involves some coefficients w, biases between the layers and differentiable non-linear functions called activation functions between the layers and denoted by $h(\cdot)$. A two-layer neural net can be trained as follows:

- a linear combination of the input is first conducted $a_j^{(1)} = \boldsymbol{w}_j^{(1)} \cdot \boldsymbol{x}$
- a_j are transformed using an appropriate activation function $h(\cdot)$ to give $z_j^{(1)} = h(a_j^{(1)})$
- z_j are again linearly combined to give $a_j^{(2)} = \boldsymbol{w}_j^{(2)} \cdot \boldsymbol{z}^{(1)}$
- finally, *the activations* $a_j^{(2)}$ are the network's output.

The network here is said to be a two-layer network because it is the number of layers of adaptive coefficients needed to determine the network properties ($w_j^{(1)}$ and $w_j^{(2)}$). $h(\cdot)$ are generally chosen to be sigmoidal functions such as logistic sigmoid or the tanh function. Other reasonably common activation functions include radial basis function (RBF) [67], Softplus [27], and hard tanh [22]. The superscript appearing on $w_j^{(1)}$ and $z_j^{(1)}$ refers to the layer in question.

To determine the set of parameters governing the neural net model, a cost function $J(w)$ is at first defined then minimized. $p(\boldsymbol{t}|\boldsymbol{x};\boldsymbol{w})$ is the distribution of the network output. By maximum the likelihood function, one can recover the hyperparameter $\boldsymbol{w}$. In practice, $p(\boldsymbol{t}|\boldsymbol{x};\boldsymbol{w})$ is assumed to be $\mathcal{N}(\boldsymbol{t}|y(\boldsymbol{x},\boldsymbol{w},\beta^{-1}))$ where β is the inverse of the variance of the Gaussian distribution. Given that the data $\boldsymbol{x}$ is a set of independent, identically distributed observations, the likelihood function corresponds to

$$p(\boldsymbol{t}|\boldsymbol{x};\boldsymbol{w}) = \prod_{i=1}^{n} p(t_i|\boldsymbol{x}_i;\boldsymbol{w},\beta). \tag{2.20}$$

Taking the negative logarithm and discarding terms that do not depend on $\boldsymbol{w}$, we recover the mean squared error:

$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n} ||y(\boldsymbol{x}_i; \boldsymbol{w}) - t_i||^2. \tag{2.21}$$

The cost function $J(\boldsymbol{w})$ is typically a highly non-linear dependence on the weights $\boldsymbol{w}$ and bias parameters. Finding global minima of $J(\boldsymbol{w})$ usually involves numerically evaluating the gradient ∇J. The back-propagation algorithm [76] uses a simple and inexpensive procedure to computing the gradient.

Finally one has to determine the architecture of the network, which refers to the depth of the network, the width of each layer, and how the units of each layer should be connected to each other [33].

2.3 (Global) sensitivity analysis

Generally speaking, sensitivity analysis (SA) aims at studying how the uncertainty in the output of a model is related to the input variables. The sensitivity analysis is said to be global when the behavior of the input variables is considered all over the domain of definition, unlike local sensitivity analysis where only the local behavior around a reference point is investigated. Different approaches in the literature address sensitivity analysis of a model, namely variance-based methods presented in Section 2.3.1 and entropy-based methods presented in Section 2.3.2.

2.3.1 Variance-based methods

Sobol' indices is a well-known global SA approach, in which the variance of the output is decomposed into contributions related to each input parameters and combinations thereof.

Let $f \in L^2([0,1]^p)$, where d is the input $\boldsymbol{x}$ dimension. $f(\boldsymbol{x})$ can be decomposed in the following way [37]:

$$f(\boldsymbol{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j}^{d} f_{i,j}(x_i, x_j) + \cdots + f_{1\ 2\ldots d}(\boldsymbol{x}), \tag{2.22}$$

where f_0 is a constant and f_i is a function of x_i, $f_{i,j}$ a function of x_i and x_j and so on such as:

$$\int_0^1 f_{i_1,i_2,\ldots,i_s}(x_{i_1}, \ldots, x_{i_s})\mathrm{d}x_{i_k} = 0,\ 1 \leq k \leq s, \{i_1, \ldots, i_s\} \subseteq \{1, \ldots, d\}. \tag{2.23}$$

This expansion is unique and all the terms in the functional decomposition are orthogonal to each other [82]:

$$\int_{[0,1]^d} f_{i_1,i_2,\ldots,i_s}(x_{i_1},\ldots,x_{i_s}) f_{j_1,j_2,\ldots,j_t}(x_{j_1},\ldots,x_{j_t})\mathrm{d}x = 0,\ \{i_1,\ldots,i_s\} \neq \{j_1,\ldots,j_t\}. \tag{2.24}$$

Applying the functional decomposition (Eq. (2.22)), the variance is written as follows:

$$D = \mathrm{Var}[f(\boldsymbol{x})] = \int_{[0,1]^d} f(\boldsymbol{x})^2 d\boldsymbol{x} - f_0^2 = \sum_{i=1}^{d} D_i + \sum_{i<j}^{d} D_{ij} + \cdots + D_{1,\ldots,d}, \tag{2.25}$$

where

$$D_{i_1,i_2,\ldots,i_s} = \int_{[0,1]^s} f^2_{i_1,i_2,\ldots,i_s}(x_{i_1},\ldots,x_{i_s})\mathrm{d}x_{i_1}\ldots\mathrm{d}x_{i_s},\ 1 \leq i_1 < \cdots < i_s \leq d,\ s \in \{1,\ldots,d\}. \tag{2.26}$$

The main effect Sobol' index is defined as follows: $S_i = \dfrac{D_i}{D}$. The total Sobol' index of the i^{th} input variable is denoted as S_i^{Tot} and quantifies the total effect of X_i on the variance of $f(\boldsymbol{x})$.

To evaluate the importance of each input variable, usually only the main effect and the total effect Sobol' indices are evaluated. Both should provide reliable information about the sensitivities of the computational model.
These indices are computed by means of Monte Carlo sampling methods, though these methods remain quite time consuming. Disposing of a surrogate model to the computational model is much more efficient in this case. Sobol' indices are analytically computed from the PCE coefficients. The PCE surrogate model thus offers a practical shortcut to compute the Sobol' indices [89].

2.3.2 Entropy-based methods

Since Shannon introduced entropy in 1948 [81] as a measure of uncertainty of a random variable, it did not stop from being wildly present in many engineering algorithms, including sensitivity analysis [5]. The main idea is to evaluate the conditional entropy of the output given the input of interest. Intuitively when the value of the conditional entropy is important it infers that the output does not depend on the input considered, and vice versa.

let H denote the Shannon entropy (also discrete entropy), and let us consider a model where $\boldsymbol{X}$ is the input and Y the response of the model as it was consistently denoted throughout this chapter. $\boldsymbol{X}$ and Y are two random variables with $p_{\boldsymbol{X}}(\boldsymbol{x})$

and $p_Y(y)$ respectively the corresponding probability density functions. The Shannon entropy of a random vector writes as follows:

$$H(\boldsymbol{X}) = -\int_{\boldsymbol{x}\in\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{2.27}$$

The entropy only depends on the probability distribution of the random variable, and not on the values. It achieves its maximum value if the random variable is uniform (highlighting the fact that the uniform variable is the "most uncertain" one, in the sense that all values have the same probability of appearance) and is at it minimum for the Dirac distribution. The conditional entropy writes as follows:

$$H(Y|\boldsymbol{X}) = -\int_{\boldsymbol{x}\in\mathcal{X}} \int_{y\in\mathcal{Y}} p_{(Y,\boldsymbol{X})}(y,\boldsymbol{x}) \log p_Y(y|\boldsymbol{X}=\boldsymbol{x}) dy d\boldsymbol{x}. \tag{2.28}$$

The mutual entropy between two random variables represents the information explained by $\boldsymbol{X}$ in Y and vise versa, and is as follows:

$$I(\boldsymbol{X}, Y) = H(\boldsymbol{X}) - H(\boldsymbol{X}|Y) = H(Y) - H(Y|\boldsymbol{X}). \tag{2.29}$$

Finally, the Krzykacz-Hausmann [49] sensitivity indices can be defined as:

$$\mu_i = \frac{I(X_i, Y)}{H(Y)} = 1 - \frac{H(Y|X_i)}{H(Y)}. \tag{2.30}$$

2.3.3 Other methods

In addition to the methods briefly introduced in Section 2.3.1 and 2.3.2, the sensitivity analysis of models behaving like linear models can be performed using regression-based indices or using Pearson correlation coefficients. Unfortunately, when the model is non-linear these indices fail to capture the sensitivity of the output to the input variables.

Another class of sensitivity indices based on dependence measures is the δ sensitivity measure of [13]. δ compares the distribution of the output $p_Y(y)$ and the conditional one $p_{Y|X_i}(y)$. The shift between the two PDFs is measured as follows:

$$\delta_i = \frac{1}{2}\mathbb{E}_{X_i}\left[\int |p_Y(y) - p_{Y|X_i}(y)| dy\right]. \tag{2.31}$$

Finally, graphical methods can end up being useful in situations where the input dimension is small. A cobweb plot, for example, enables the user to capture trends of dependence easily.

2.4 Conclusions

In this first chapter we merged concepts from both fields of statistical learning as well as uncertainty quantification basics. First, we introduced some of the most well-known machine learning algorithms to perform a statistical analysis and make predictions based on a dataset or a *DoE* from running a heavy numerical code. Section 2.3 addressed the sensitivity analysis of a model and introduced the reader to the most common methods for sensitivity analysis.

There is of course much more to statistical learning and uncertainty quantification than what is addressed here; other aspects such as reliability assessment, robustness of the models, sampling methods are either omitted or barely mentioned here to keep the manuscript concise.

Subject to a deterministic context, several books are dedicated to machine learning algorithms and to statistical learning in general [36, 10, 33]. Surrogate modeling in a stochastic context is a much less frequently explored field of research.

Chapter 3

Surrogate modeling of stochastic simulators

Contents

3.1 Context

Simulators (also called computational models) are mathematical models that mimic the behaviour of physical phenomena. Finite element models, for instance, simulate fluid dynamics equations in different applications ranging from blood flow to aerodynamics. Those simulators allow for solving the governing equations of the systems components and predict the changes of performance of the system when some parameters vary.

Some simulators may contain internal sources of randomness on top of uncertain input variables. Carrying out deterministic numerical operations without considering uncertainties leads to unreliable designs.

Simulators that describe uncertain model outputs, for a given input vector, are called stochastic simulators. In contrast to the deterministic ones which yield a unique output for each set of input parameters, stochastic simulators inherently contain some source of randomness, more precisely, the output at a given input is a random variable with a probability density function to be characterized. The mathematical object suitable to represent stochastic models is stochastic processes.

A stochastic process is a family of random variables indexed by a mathematical set. Let's consider $D \in \mathbb{R}^n$ the space of the input parameters, $\boldsymbol{x}$ the input variable such as $\boldsymbol{x} \in D$. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is a sample space, $\mathcal{F}$ is a σ-algebra and $\mathbb{P}$ the probability measure.

$H(\boldsymbol{x}, \omega),\ \omega \in \Omega$ denotes a stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by $\boldsymbol{x} \in D$. At a fixed $\boldsymbol{x}$, $H(\boldsymbol{x}, \omega)$ is a random variable, for a fixed ω, $H(\boldsymbol{x}, \omega)$ is a deterministic function of $\boldsymbol{x}$ and is called a trajectory. The covariance function of the process reads as follow:

$$C(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[H(\boldsymbol{x}, \omega)H(\boldsymbol{y}, \omega)], \tag{3.1}$$

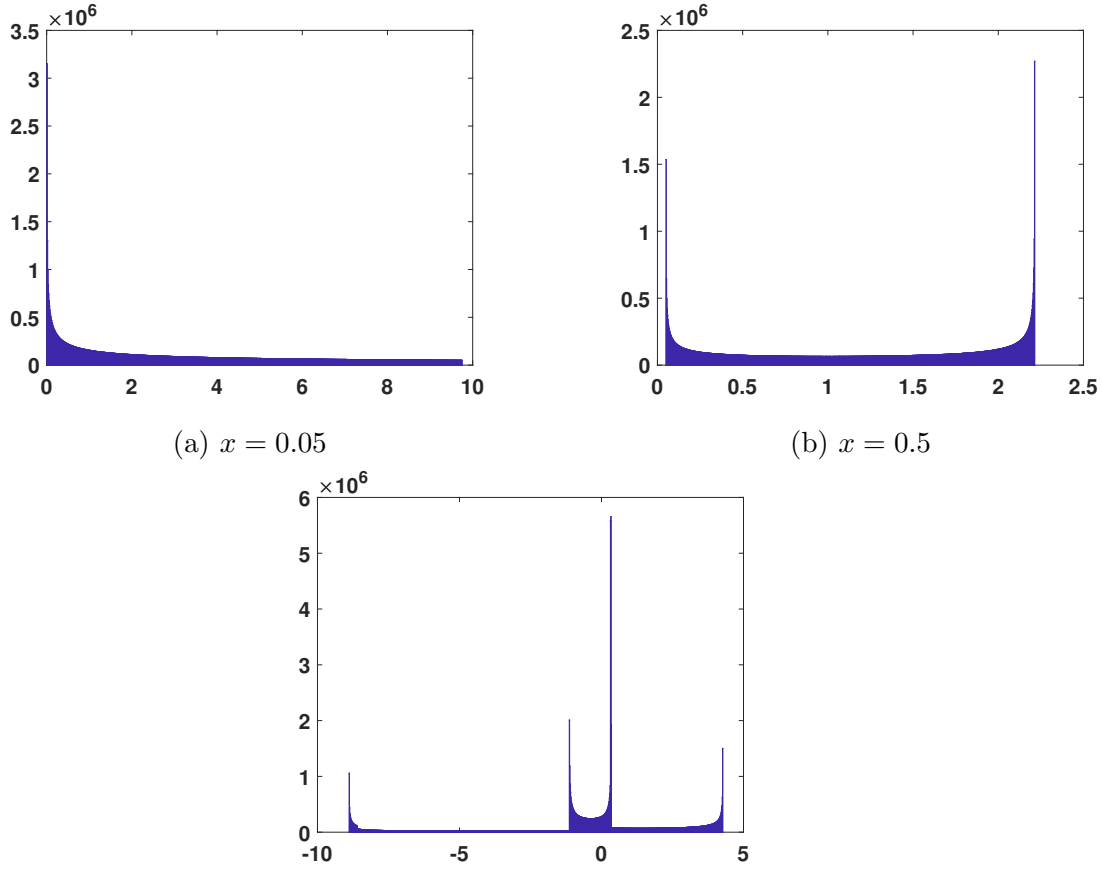
where $\boldsymbol{x}$ and $\boldsymbol{y}$ are in D, and $\mathbb{E}$ is the mean function of H.

Gaussian processes are a particular kind of stochastic process; every finite linear combination of random variables from this stochastic process is normally distributed, i.e $H(\boldsymbol{x}, \omega)$ is Gaussian if and only if for every finite set of indices $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ in the index set D, $(H(\boldsymbol{x}_1, \omega), \ldots, H(\boldsymbol{x}_k, \omega))$ is a multivariate Gaussian random variable. A nice feature of Gaussian processes is the fact that they are fully characterized given their mean and covariance functions.

Eq. (3.2) is an example of a dummy stochastic simulator, where $x \in [-\pi, \pi]$ $w \sim \mathcal{U}([-\pi, \pi])$. The output on three different points is plotted in Figure 3.1. The

probability density function (PDF) for each point form the DoE set do not necessarily have nice properties such as unimodality or being symmetrical. Eq. (3.2) presents an example where the output PDF can be unimodal, bimodal and multimodal, depending on the input points.

$$H(x,\omega) = ax\cos(w)^2 + w^2\cos(wx). \tag{3.2}$$



(a) $x = 0.05$

(b) $x = 0.5$

(c) $x = \pi$

Figure 3.1: The output PDF for three points, $a = 0.1$.

The simulators are often times computationally prohibitive due to the use of high-fidelity computations. This is where surrogate modeling becomes handy for the user. By substituting the heavy simulator by a mathematical function, quantities of interest and in general simulations can be affordably evaluated.

We aim at building a stochastic process $H(\boldsymbol{x},\omega)$ as a surrogate for the original stochastic simulator. The conventional first step is to design a sampling set $DoE = \{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(M)}\}$, run the simulator on the sampling set to then gather a training

data set $\{(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}), \ldots, (\boldsymbol{x}^{(M)}, \boldsymbol{t}^{(M)})\}$, where $\boldsymbol{t}^{(k)}$ is the target random vector for each $k \in \{1, \ldots, M\}$.

In the literature, the two types of simulators (deterministic and stochastic) are dealt with differently. Whereas the literature is abundant and very diverse for deterministic models metamodeling (Chapter 2), stochastic simulators metamodeling is a less mature field. Section 3.2 introduces the existing methods applied to surrogate model stochastic simulators.

3.2 State of the art

This section briefly summarizes the existing methods in the literature. Most of the methods dealing with predictions in a stochastic context focus on a quantity of interest (QoI) (or a set of QoI). Among the different QoI appearing frequently in the applications are, in the first place, the mean and the variance.

In [4] the quantities of interest considered were the mean and the variance, the proposed approach is an extension to Gaussian process modeling in the sense that the target is supposed to be a realization of a Gaussian process with a heteroscedastic variance. In other words instead of assuming that the function emulating the model $y(\boldsymbol{x})$ (same notation as Eq. (2.1)) is such as

$$y(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2 k(\boldsymbol{x}, \boldsymbol{x})), \tag{3.3}$$

it is rather supposed that

$$y(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\delta(\boldsymbol{x}) + k(\boldsymbol{x}, \boldsymbol{x}))). \tag{3.4}$$

The term $\delta(\boldsymbol{x})$ is estimated at each point $\boldsymbol{x}_i$ using replications. The design of experiment considered in [4] is $(\boldsymbol{x}_i, N_i)$. The number of replications allocated to each point $\boldsymbol{x}_i$ depends on the current point. The approach is based on two steps: a random replication number is first used for all the DoE, and likelihood equations are solved to estimate the parameters. The second step consists in using Eq. (29) from [4] to update the number of replications for each $\boldsymbol{x}_i \in DoE$.
This parametric approach has been used successfully in game theory simulations [66] and in measuring portfolio risk in finance [53].

Also based on select statistical indicators (here quantiles Q_α of level α), authors in [65] emulated the quantile function of the stochastic simulator. $Q_\alpha(\boldsymbol{x})$ is considered unknown, but M replicates can be drawn $(\tilde{Q}_\alpha(\boldsymbol{x}^{(1)}), \tilde{Q}_\alpha(\boldsymbol{x}^{(2)}), \ldots, \tilde{Q}_\alpha(\boldsymbol{x}^{(M)}))$, where $\tilde{Q}_\alpha$ corresponds to the estimate quantile of Q_α on the M points of the DoE. A prior

distribution of Q_α is a Gaussian process and a Kriging metamodel with a nugget parameter is used to get $\tilde{Q}_\alpha(\boldsymbol{x}_i)$. (The nugget parameter in the Kriging is used to avoid numerical instability in the computation of the inverse of the covariance matrix and to include noisy data [64].) Practically, the following empirical estimator has been used in [65]:

$$\tilde{Q}_\alpha(\boldsymbol{x}) = \inf\{s; \sum_{i=1}^{M} \mathbb{1}(t_i \leq s) \leq M\alpha\}, \tag{3.5}$$

where M is the size of DoE. The metamodel of Q_α is the same as in Eq. (3.4) except that the term δ is here constant and independent of $\boldsymbol{x}$. It represents the variation of $Q_\alpha(\boldsymbol{x}) - \tilde{Q}_\alpha(\boldsymbol{x})$.

Generalized additive models (GAM) [35] were used to predict the mean of the stochastic simulator. GAM is a generalized approach to linear models, where a linear predictor is replaced by an additive predictor, which allows more flexibility. The mean predictor for example can be written

$$\mu = \sum_j \rho_j(\boldsymbol{x}_j), \tag{3.6}$$

whereas a linear predictor would be $\sum_j \beta_j \boldsymbol{x}_j$. The functions $\rho(.)$ are obtained by fitting a smoother to the data (e.g splines). They are here univariate but can be multivariate.

In [44] the mean and the variance were simultaneously fitted based on two interlinked GAM. Meanwhile in [56, 47] the joint model for the mean and the variance is built based on Gaussian process.

Other methods share the same outlook as [4], [65] and [44], usually the interest is focused on the same summary statistics (mean, variance and quantiles) but the surrogating techniques may differ; in [73] GAM models were used to predict the mean and the variance of the conditional distribution. In [72] the mean function was predicted by assuming that the output is a mixture of normal distributions. Some works approached the problem from a different angle; the PDF is assumed to belong to a certain family fully determined by some coefficients. For instance in [61], the PDF f^* to be predicted at a new point $\boldsymbol{x}^* \in D$ such as $\boldsymbol{x}^* \notin DoE$ is approximated by:

$$\hat{f}^*(\boldsymbol{x}) = \sum_{i=1}^{q} \phi_i(\boldsymbol{x}^*)\rho_i(\boldsymbol{x}), \tag{3.7}$$

where ϕ_i are functions from D to $\mathbb{R}$, called *coefficient functions* such as for $\boldsymbol{x} \in D$

$$\begin{cases} \phi_i(\boldsymbol{x}) \geq 0 \\ \sum_{i=1}^{q} \phi_i(\boldsymbol{x}) = 1 \end{cases}, \tag{3.8}$$

and ρ_i are a set of basis PDF. To choose the basis, the author used three different approaches: an adaptation to functional principal component analysis [69], magic points method [54] and through minimizing the approximation error [61]. The three approaches can be written as optimization problems.
All the estimated PDF functions in [61] (ϕ_i and $\hat{f}_i$, the PDF function from the sampled random variable $\boldsymbol{t}_i$) have the extra constraint to be a PDF .i.e. to be non-negative and of integral equal to 1 (Eq. (3.8)), which is difficult to achieve in practice, especially in high dimension.
In the same line of reasoning, authors in [99] assumed that the PDF belongs to the generalized lambda distribution (GLD) .i.e.

$$H(\boldsymbol{x}, \omega) \sim GLD(\lambda_1(\boldsymbol{x}), \lambda_2(\boldsymbol{x}), \lambda_3(\boldsymbol{x}), \lambda_4(\boldsymbol{x})). \tag{3.9}$$

For all $\boldsymbol{x} \in DoE$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are estimated from the replications using the method of moments [50] or the maximum likelihood estimation [86]. A PCE is then used to surrogate the distribution parameters $\boldsymbol{\lambda}(\boldsymbol{x})$.

To illustrate the different schools when it comes to surrogate modeling stochastic simulators, a couple of papers from each school were briefly detailed. Namely the approaches based on selected statistics like the mean and the variance, and the approaches based on a functional decomposition of the PDF. Both approaches can be considered as parametric, in the sense that the PDF is reduced to its first moments, quantiles, or to its representation in a basis with predetermined number of coefficients.

The next section 3.3.1 introduces the approach developed in this thesis which was published in [6]. The approach is based on a non-parametric representation of the stochastic process $H(\boldsymbol{x}, \omega)$ using Karhunen-Loève (KL) expansion. First the KL theorem is recalled and the proposed method that makes use of the KL spectral expansion is then presented. The method has two different approaches. Each one is detailed in Sections 3.3.3 and 3.3.4, and they are compared in Section 3.3.5. The evaluation of the method is presented in Section 3.3.7. Finally discussions and conclusions are provided in the last Section 3.3.9.

3.3 Surrogate modeling of stochastic simulators based on KL decomposition

As a start, the general framework is briefly reminded. The first step is to design a sampling set $DoE = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\}$, run the simulator on the sampling set to

then gather a training data set $\{(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}), \ldots, (\boldsymbol{x}^{(M)}, \boldsymbol{t}^{(M)})\}$, where $\boldsymbol{t}^{(k)}$ is the target random vector for each $k \in \{1, \ldots, M\}$. The target $\boldsymbol{t}^{(k)}$ is a vector with N replications $\boldsymbol{t}^{(k)} = \{\boldsymbol{t}^{(k,1)}, \ldots, \boldsymbol{t}^{(k,N)}\}$. The objective is to fit a stochastic process to the training data set using KL theorem recalled in Section 3.3.1. Traditionally KL decomposition is used to simulate and represent a stochastic process analogous to a Fourier series representation of a function. In this work we use the KL expansion as a surrogate model of the stochastic simulator.

3.3.1 Karhunen-Loève decomposition

Also known as proper orthogonal decomposition, the KL decomposition essentially involves the representation of a stochastic process according to a spectral decomposition of its correlation operator.

Let $\{H(\boldsymbol{x}, \omega),\ \boldsymbol{x} \in D\}$ be a zero mean second order stochastic process. Its covariance function is continuous in the mean square sense and denoted as $C(\boldsymbol{x}, \boldsymbol{y})$. The eigenvalue problem related to the covariance function reads:

$$\int_D C(\boldsymbol{x}, \boldsymbol{y})\phi_i(\boldsymbol{y})d\boldsymbol{y} = \lambda_i \phi_i(\boldsymbol{x}), \tag{3.10}$$

where $\{\phi_i, i \in \mathbb{N}\}$ and $\{\lambda_i, i \in \mathbb{N}\}$ are the eigenvectors and the eigenvalues respectively. Furthermore, choose

$$\xi_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D H(\boldsymbol{x}, \omega)\phi_i(\boldsymbol{x})d\boldsymbol{x}. \tag{3.11}$$

Then the KL expansion reads:

$$H(\boldsymbol{x}, \omega) = \sum_{i=1}^{+\infty} \sqrt{\lambda_i}\xi_i(\omega)\phi_i(\boldsymbol{x}). \tag{3.12}$$

The random variables ξ_i have zero mean, unit variance, and are mutually uncorrelated, i.e. orthogonal with respect to the underlying probability measure. They are generally not independent, except for the case of Gaussian processes.

$$\mathbb{E}[\xi_i] = 0, \qquad \mathbb{E}[\xi_i \xi_j] = \delta_{ij}. \tag{3.13}$$

The KL expansion is optimal in the mean square sense; when truncated after a finite number p of terms, the resulting approximation minimizes the mean square error.

3.3.2 The proposed method for stochastic emulators

For the sake of simplicity, it is assumed in this chapter, that the random processes are of zero-mean. It is also assumed that it is possible to *freeze* the randomness ω and hence simulate trajectories by sampling, *for a frozen* ω, the model response at different values of $\boldsymbol{x}$. In other words, we are able to generate $H(\boldsymbol{x}^{(1)}, w_k)$ and $H(\boldsymbol{x}^{(2)}, w_k)$ with the same w_k, where w_k is an internal source of stochasticity in the simulation tool. This assumption enables to compute the empirical covariance function of the model output, and thus apply the KL expansion that can be used to model the random process. The design of the experiments consists of $DoE = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\right\} \subset D$ and the corresponding response. The process is simulated in each $\boldsymbol{x} \in DoE$ with the same random seed. For each point of DoE (M points), simulations have been carried out using N different random seeds, which corresponds to generate N trajectories of the random process at M discrete points. Using KL expansion, the random process can be modeled as [32]:

$$H(\boldsymbol{x}, \omega) \simeq \sum_{i=1}^{M} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}), \tag{3.14}$$

where $\boldsymbol{\phi_i}$ and λ_i are respectively the eigenvectors and the eigenvalues of $C(\boldsymbol{x}, \boldsymbol{y})$. ξ_i are uncorrelated random variables with unit variance (detailed in Section 3.3.6) and are given by,

$$\xi_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D H(\boldsymbol{x}, \omega) \phi_i(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.15}$$

Eq. (3.14) provides a potential surrogate to the stochastic simulator. let $\boldsymbol{x}^*$ be a new point such as $\boldsymbol{x}^* \in D$ and $\boldsymbol{x}^* \notin DoE$, the aim is to predict the random response. Eq. (3.14) requires the knowledge of $\xi_i(\omega)$ and $\phi_i(\boldsymbol{x})$ for $\boldsymbol{x}^* \in D$. However the eigenvectors $\phi_i(\boldsymbol{x})$ are only known at the sampled points of the DoE after solving the discrete KL problem. In order to get the value of the eigenvectors over the domain of interest, we proceed in two different ways, either by:

- metamodeling the eigenvectors via usual surrogate modeling methods;

- or surrogating the empirical covariance, then find the new eigenvectors on the domain of interest.

As far as the random variables $\xi_i(\omega)$ are concerned, they can be obtained as the discrete projection of the random process over the $\phi_i(\boldsymbol{x})$ (see Section 3.3.6).

The following sections describe the two approaches as well as the way the random variables $\xi_i(\omega)$ are characterized. For an overview of the multiple steps conducted in this method, the flowchart in Figure 4.1 describes the sequence of steps.

3.3.3 Surrogate model of the underlying covariance function

As mentioned above, the values of the eigenvectors on $\boldsymbol{x}^*$ are not available since $\boldsymbol{x}^* \notin DoE$. In this subsection a surrogate model of the covariance is used to predict the covariance not only over the DoE points but also over the whole domain of interest D (in particular for other points like $\boldsymbol{x}^*$). For the sake of simplicity we assume that the DoE and the points where predictions are to be made add up to M^* points hence $M \leq M^*$. Let $\hat{C}$ be the metamodel of the empirical covariance function C, built using a polynomial chaos expansion for instance. $\hat{C}$ allows one to have a predicted covariance for the M^* new points of interest as follows.

$$\hat{C}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=0}^{P} a_j \psi_j(\boldsymbol{x}, \boldsymbol{y}), \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in D^2. \tag{3.16}$$

The surrogate modeling technique is not the focus of this section, all the approaches presented in Chapter 2 can be applied as long as they provide a surrogate to the covariance function of H. Either way, the surrogated covariance matrix is now a $M^* \times M^*$ matrix, hence the number of eigenvectors $\hat{\boldsymbol{\phi}}_i$ of $\hat{\boldsymbol{C}}$ is M^*. Note that when metamodeling the covariance, the input dimension is doubled (the covariance is defined on the product set $D \times D$). For instance, for an input dimension of three, the covariance function has an input dimension of six, and so does its surrogate. Learners that tend to perform poorly in higher dimensions are to be avoided in this step.

3.3.4 Surrogate model of the eigenvectors

A more straight forward approach to get the eigenvectors all over D would be to interpolate the eigenvectors $\boldsymbol{\phi}_i$. Let $\hat{\boldsymbol{\phi}}_i$ be a surrogate function of the true eigenvector $\boldsymbol{\phi}_i(.),\ i = \{1, \cdots, M\}$ based on the DoE. The KL expansion (Eq. (3.14)) will then read as follow:

$$H(\boldsymbol{x^*}, w) = \sum_{i=1}^{M} \sqrt{\lambda_i} \xi_i(\omega) \hat{\phi}_i(\boldsymbol{x^*}). \tag{3.17}$$

The interpolation of $\phi_i(\boldsymbol{x})$ can be done with any surrogating technique that interpolates the data, i.e techniques where the predicted value is identical to the simulated value at the points of the DoE. As an example, cubic spline interpolation can be

used for one or two dimensional models. When considering higher dimension we can use Kriging, linear interpolation or decompose onto radial basis functions.

Starting from eigenvectors known over the DoE, a surrogate model of $\phi_i(\boldsymbol{x})$ enables us to build $\hat{\phi}_i(\boldsymbol{x})$ $\forall \boldsymbol{x} \in D$, hence evaluate $\boldsymbol{\phi}_i$ over all D. This approach is intuitive: following the eigendecomposition, we predict the new point's coordinates with the adequate exact interpolator and as shown in Eq. (3.17) deduce the stochastic process response.

3.3.5 Conclusion on the two approaches and outlook

This section discusses the properties of the covariance surrogate. When the second approach is considered (the eigenvectors are surrogated one-by-one as in Section 3.3.4), there is actually no guarantee that the new eigenvectors form an orthogonal base in the new set. Meanwhile when applying the first approach (surrogate modeling the covariance as in Section 3.3.3), to perform next the eigendecomposition of the covariance surrogate, the eigenvectors form an orthogonal base of the covariance surrogate. To overcome this lack of rigour, Gram-Schmidt orthogonalizing process can be applied in future related works.

A covariance operator is symmetrical and positive definite, in particular the surrogated covariance from the first approach (Section 3.3.3). The covariance surrogate $\hat{\boldsymbol{C}}$ has been symmetrized, meaning that if the obtained metamodel of the covariance is denoted C_* (which is not necessarily a symmetric function of its inputs $(\boldsymbol{x}, \boldsymbol{y})$) then we consider $\hat{\boldsymbol{C}} = \dfrac{C_* + C_*^\intercal}{2}$. However the surrogated covariance is not systematically a positive definite matrix. To overcome this limitation one can think of imposing a constraint on the learning method. A clever choice of the parameters of the metamodel can guarantee that the surrogate covariance is a positive definite matrix. In such case the method will no more be agnostic to the choice of the learners, but rather depend on the flexibility of the learning method. This idea was not applied in this work and is only mentioned for the sake of future exploration.

Both problems can be seen in a slightly different way. The aim here is to mimic a function (either the eigenbase or the covariance operator), based on a test set; somehow the surrogate model is supposed to reproduce the properties of the function (either orthogonal base or covariance operator) as accurately as possible.

3.3.6 Random variable evaluation

First the trivial case is considered, in the case where $H(\boldsymbol{x}, \omega)$ is a Gaussian process, the ξ_i appearing in the KL expansion in Eq. (3.14) are zero-mean, unit-variance, independent Gaussian random variables [32], so no computation is needed.

When dealing with more general random processes, ξ_i are the projection of H onto the base of the eigenvectors $\hat{\boldsymbol{\phi}}_i$ and given by Eq. (3.15). The integral in Eq. (3.15) cannot be calculated since H is only known over the M points of the DoE and over the N trajectories. To overcome this limitation, the integral is approximated with a sum involving the M known values of H:

$$\hat{\xi}_i(\omega_k) = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{M} \nu_j H(\boldsymbol{x}^{(j)}, \omega_k) \hat{\phi}_i(\boldsymbol{x}^{(j)}), \tag{3.18}$$

where ν_j is the volume of the i^{th} partition of D, $k \in \{1, \ldots, N\}$ is the trajectory index and $j \in \{1, \ldots, M\}$ indicates the M points where H was simulated. If the covariance is surrogated and the base is expended then $k \in \{1, \ldots, M^*\}$, otherwise if only the M eigenvectors were surrogated then $k \in \{1, \ldots, M\}$.

There are as many random variables $\hat{\xi}_i$ as basis vectors $\hat{\boldsymbol{\phi}}_i$. When the eigenvectors are interpolated, the cardinality of $\hat{\boldsymbol{\phi}}_i$ and $\hat{\xi}_i$ is M. In the second option (when the covariance matrix is interpolated) the cardinality of $\hat{\boldsymbol{\phi}}_i$ and $\hat{\xi}_i$ is M^*. Either way, the projection is only computed using the M points simulated because H is known only on the M points of the DoE. Therefore, even when the base is extended and M^* eigenvectors are available, $\hat{\xi}_i(\omega_k)$ only depends on the M points of the DoE.

3.3.7 Error evaluation

Once the surrogate model is built, it is of interest to evaluate the accuracy of the prediction. Because the comparison here is between two random variables (true PDF response and the predicted PDF in $\boldsymbol{x}^*$), metrics from the probabilistic framework take over.

3.3.7.1 Probabilistic metrics comparing the PDFs

When dealing with Gaussian processes, and since we only consider centred processes in this work, the error estimation will boil down to comparing the variance of the original stochastic simulator with that of the emulator. For a new point $\boldsymbol{x}^*$ where the surrogate is to be evaluated, one gets:

$$\hat{H}(\boldsymbol{x}^*, \omega) = \sum_{i=1}^{M} \sqrt{\lambda_i} \xi_i(\omega) \phi_i(\boldsymbol{x}^*), \tag{3.19}$$

where $(\xi_i, i = 1, \cdots, p)$ are independent standard normal variables in this case. Then, the associated variance reads:

$$\sigma^2(\boldsymbol{x}^*) = \sum_{i=1}^{p} \lambda_i \phi_i(\boldsymbol{x}^*)^2. \tag{3.20}$$

For non-Gaussian processes, statistical tests can be applied to quantify the error of the metamodel. The Kolmogorov Smirnov (KS) test has been used to test if two drawn samples are from the same distribution (null hypothesis). The null hypothesis is rejected at level α if

$$KS_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| > c(\alpha) \sqrt{\frac{n+m}{nm}}, \tag{3.21}$$

where n, F_1 and m, F_2 are respectively the size of the samples and their empirical distribution functions.

A more intuitive and graphical approach to compare two distributions is using histogram intersection: when it is equal to 0, no overlap exists between the two of them, and when it is equal to 1, they are identical (Figure 3.2). The drawback of this approach is the influence of the selection of the bins, especially for long tailed distributions. In practice, a bin number is defined for both PDFs, the space is decomposed into the defined number of bins, and discrete probabilities are evaluated for each PDF on each bin. The histogram intersection error is defined as the minimum probability of the PDFs in question summed on all the bins.

In addition to the KS test and the histogram intersection, we introduce two more metrics, namely the Hellinger distance [9] and the Jensen-Shannon divergence [29].

3.3.7.2 Hellinger distance

Let p and q be two discrete probability measures. The Hellinger distance reads as follows:

$$H(p, q) = \frac{1}{\sqrt{2}} ||\sqrt{p} - \sqrt{q}||_2. \tag{3.22}$$

Hellinger distance forms a bounded ($\in [0, 1]$) metric on the space of probability distribution.

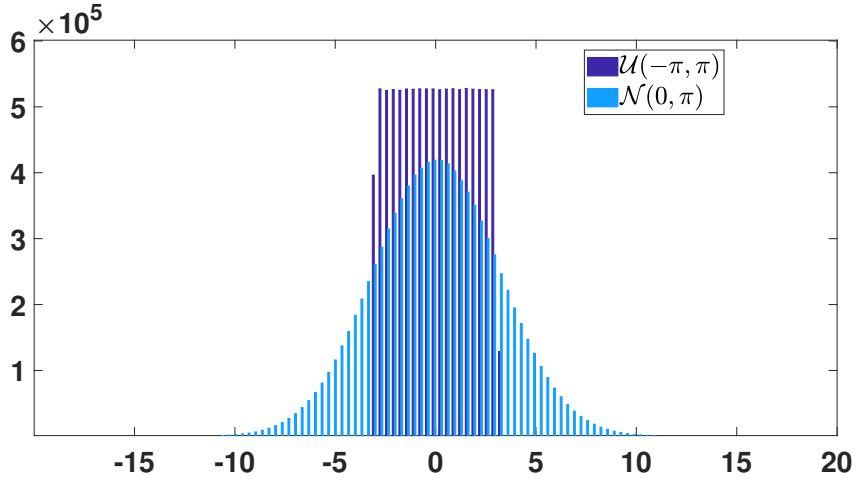


Figure 3.2: Visualization of histograms intersection.

3.3.7.3 Jensen-Shannon divergence

Based on the Kullback-Leibler divergence, the Jensen-Shannon (JS) divergence is a statistical method of measuring the behaviour of two different distributions. A Jensen-Shannon divergence equal to 1 indicates that the two distributions are totally different. If the Jensen-Shannon divergence is equal to 0, the two distributions are the same almost everywhere. We first introduce the Kullback-Leibler divergence. Let p and q be two discrete probability measures. Then:

$$D_{KL}(p||q) = -\sum_i p(i) \log \frac{q(i)}{p(i)}. \tag{3.23}$$

Let $r = (p+q)/2$ then the Jensen-Shannon divergence reads as follow

$$JSD(p,q) = \frac{D_{KL}(p||r) + D_{KL}(q||r)}{2}. \tag{3.24}$$

The Jensen-Shannon divergence is symmetric, finite and $0 \leq JSD(p,q) \leq 1$. The different error metrics stated above provide a different information on how the real and the surrogated PDFs are similar. The histogram intersection metric does not provide information about the shape-similarity of two PDFs. To cover up this limitation, JS divergence provides an idea on how much the compared PDFs belong to a same probability family but tends to be non-discriminant.

3.3.7.4 Cross validation

To estimate the accuracy of the surrogate prediction, we perform a k-fold cross-validation: the data is partitioned onto k subsets of equal size. At each step a single

subsample is retained as the validation set for testing the model, and the remaining data are used to build the surrogate model. The k-fold validation is repeated for several partitions of the data (Figure 3.3). The error is evaluated using the error metrics defined above, namely the KS test, the histogram intersection, the Hellinger distance and the JS divergence.

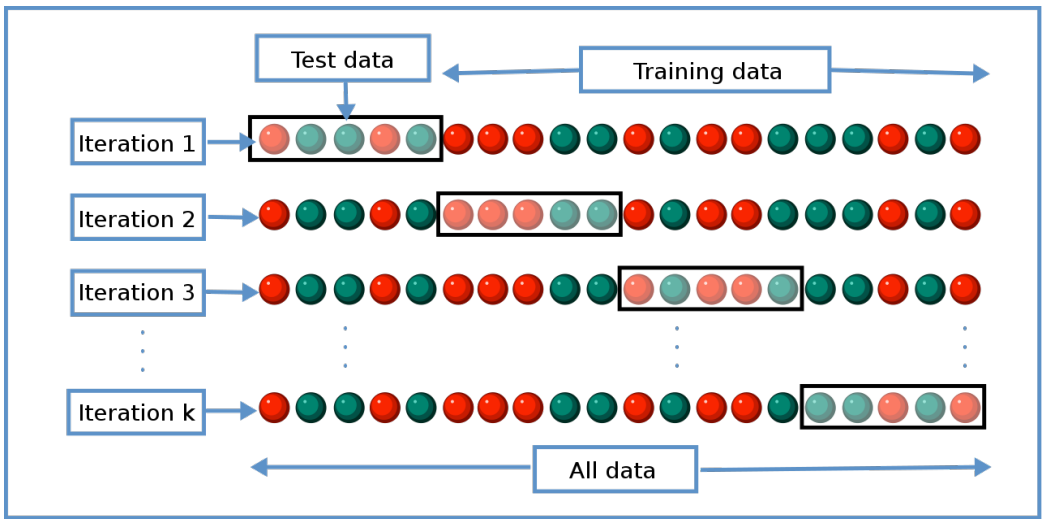


Figure 3.3: Visualization of the k-fold cross validation. Figure from wikipedia.org.

3.3.8 Application on an analytical 3-dimensional example

For demonstration purposes, the method described in the previous sections have been tested on a dummy stochastic simulator consisting of an analytical, 3-dimensional function. We remind that we are considering only centered processes in this chapter. When considering the simulated data, this is achieved by removing the empirical mean prior to any treatment. The surrogate models (PCE and Kriging) are obtained with the Matlab package UQLab [55].

The stochasticity is introduced to the process through known distributions. By means of simulations on different points of the design of the experiments and numerous replications, the empirical covariance of the process is assessed. Let H be a random process on $D = [0,2]^3 \times \Omega$:

$$\begin{aligned} H(\boldsymbol{x},\omega) &= 100\ \omega_1\left(\frac{1}{10}\exp(x_1\omega_2) + x_2x_3\omega_3\right), \\ \boldsymbol{x} = (x_1,x_2,x_3) \in [0,2]^3 \text{ and } \omega_1 &\sim \mathcal{N}(0,1),\ \omega_2 \sim \mathcal{U}([1,2]),\ \omega_3 \sim \mathcal{U}([0,1]). \end{aligned} \tag{3.25}$$

Based on a Latin hypercube sampling (LHS), the design of experiments (DoE) is 30 points in $[0, 2]^3$, and 50 realizations on each point, which makes a total computational cost of $1,500$ calls to the random function. These numbers are selected as if it was a real costly simulator. The trajectories are the same for all 30 points of the DoE. The empirical covariance is $C(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left[H(\boldsymbol{x}, \omega)H(\boldsymbol{y}, \omega)\right]$.

Following the simulations and the covariance computation, two options are tested (Figure 4.1). In the first approach we interpolate the basis vectors independently using linear interpolation at first, then using Kriging. The aim is to test the impact of the interpolation technique on the process surrogate, hence the choice of linear metamodel ('basic' interpolator) and Kriging metamodel ('advanced' interpolator).

For the second approach a PCE surrogate model $\hat{C}$ is built to surrogate the covariance function:

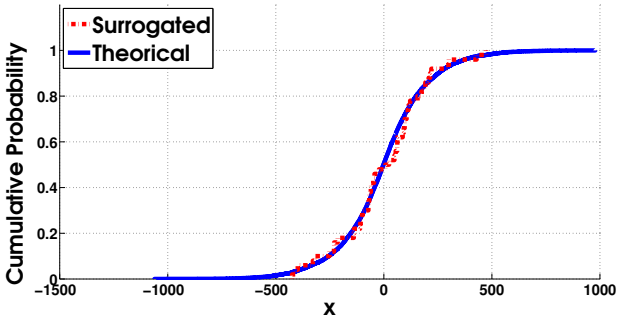
$$(x_1, x_2, x_3, y_1, y_2, y_3) \in [0, 2]^6 \to \hat{C}(x_1, x_2, x_3, y_1, y_2, y_3) \in \mathbb{R}. \tag{3.26}$$

The covariance metamodel has $2 \times 3 = 6$ inputs, and has a training set of size up to 29×29, depending on the size of the test set. Results from both approaches are presented in Table 3.1. The mean value of the three error metrics evaluated over $3,000$ test points shows that surrogating the eigenvectors using Kriging performs best for this toy example. Three examples are plotted in Figure 3.4, the surrogated density is computed respectively by interpolating the eigenvectors using linear model, interpolating the eigenvectors using Kriging and finally interpolating the covariance using PCE (Figure 4.1). The histogram intersection error in the three cases is respectively 0.89, 0.96 and 0.55 (equal to the mean error (Table 3.1)).

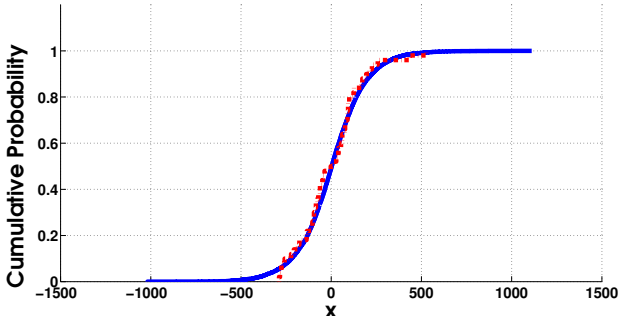
Table 3.1: Mean error over 3,000 test points.

Method	Histogram intersection	Hellinger distance	JS divergence
Linear interpolation of eigenvectors	0.89	0.06	0.004
Kriging surrogate of eigenvectors	0.96	0.025	0.001
PCE covariance surrogate	0.55	0.27	0.03

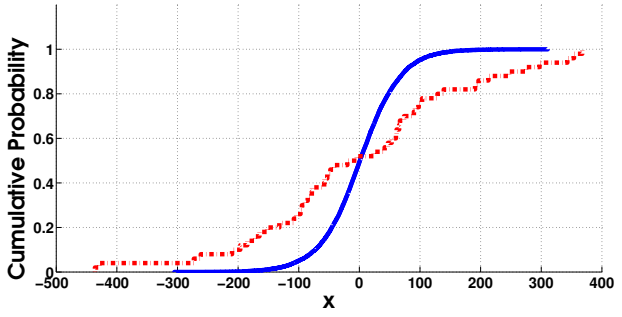
To characterize the dependence of the method on the surrogate model used, the size of input data M and the number of realizations, the histogram intersection error is estimated, and results are presented in Table 3.2. For this comparison, only the histogram intersection metric is used. Hellinger distance varies in the same way as the histogram intersection and JS divergence did not seem to be discriminant.



(a) Linear interpolation of eigenvectors (mean histogram intersection is equal to 0.89).



(b) Kriging surrogate of the eigenvectors (mean histogram intersection is equal to 0.96).



(c) PCE surrogate of the covariance (mean histogram intersection is equal to 0.55).

Figure 3.4: Surrogated and true CDFs plotted in the three approaches.

Table 3.2: Parametric study of the histogram intersection error by varying the size M of the DoE and the number of realizations N.

	M the size of DoE											
	30			60			100			200		
	lin	Krig	PCE	lin	Krig	PCE	lin	Krig	PCE	lin	Krig	PCE
$N = 50$	0.89	0.96	0.55	0.92	0.97	0.56	0.95	0.99	0.65	0.96	0.99	0.63
$N = 100$	0.9	0.96	0.63	0.94	0.98	0.62	0.95	0.99	0.66	0.96	0.99	0.59
$N = 1000$	0.96	0.98	0.7	0.96	0.99	0.77	0.97	0.99	0.72	0.98	0.99	0.71

Table 3.2 shows that the performance increases when M and/or N increases. That said, increasing N seems to grant a better accuracy compared with increasing M.

The poor performance of the PCE surrogate points out to the dependence of the overall method on the eigenvectors and their computation and is probably due to the following reasons:

- For a data set of size $M = 30$, there is $30 \times 30 = 900$ covariance terms. Hence the surrogate model of the covariance will have 900 inputs (as in Eq. (3.26)), the PCE model might get noisy and over-fitted.

- The covariance surrogate $\hat{\boldsymbol{C}}$ has been symmetrized, meaning that if the obtained metamodel of the covariance is denoted C_* (which is not necessarily a symmetric function of its inputs $(\boldsymbol{x}, \boldsymbol{y})$) then the following surrogate is considered $\hat{\boldsymbol{C}} = \dfrac{C_* + C_*^\intercal}{2}$. This step may contribute to the noisy results. Surrogate modeling C only on a triangular domain has been tested, yet the performance on the same test points did not improve.

The error is always evaluated between the simulated and the surrogated PDF (using one of the three options), mainly because in case studies the real PDF is usually unknown, hence comparing the surrogate and the original simulator is impossible.

3.3.9 Conclusions

The flowchart in Figure 4.1 is a reminder of the multiple steps conducted in the proposed methods, namely the two possible approaches: either surrogating the covariance or the eigenvectors.

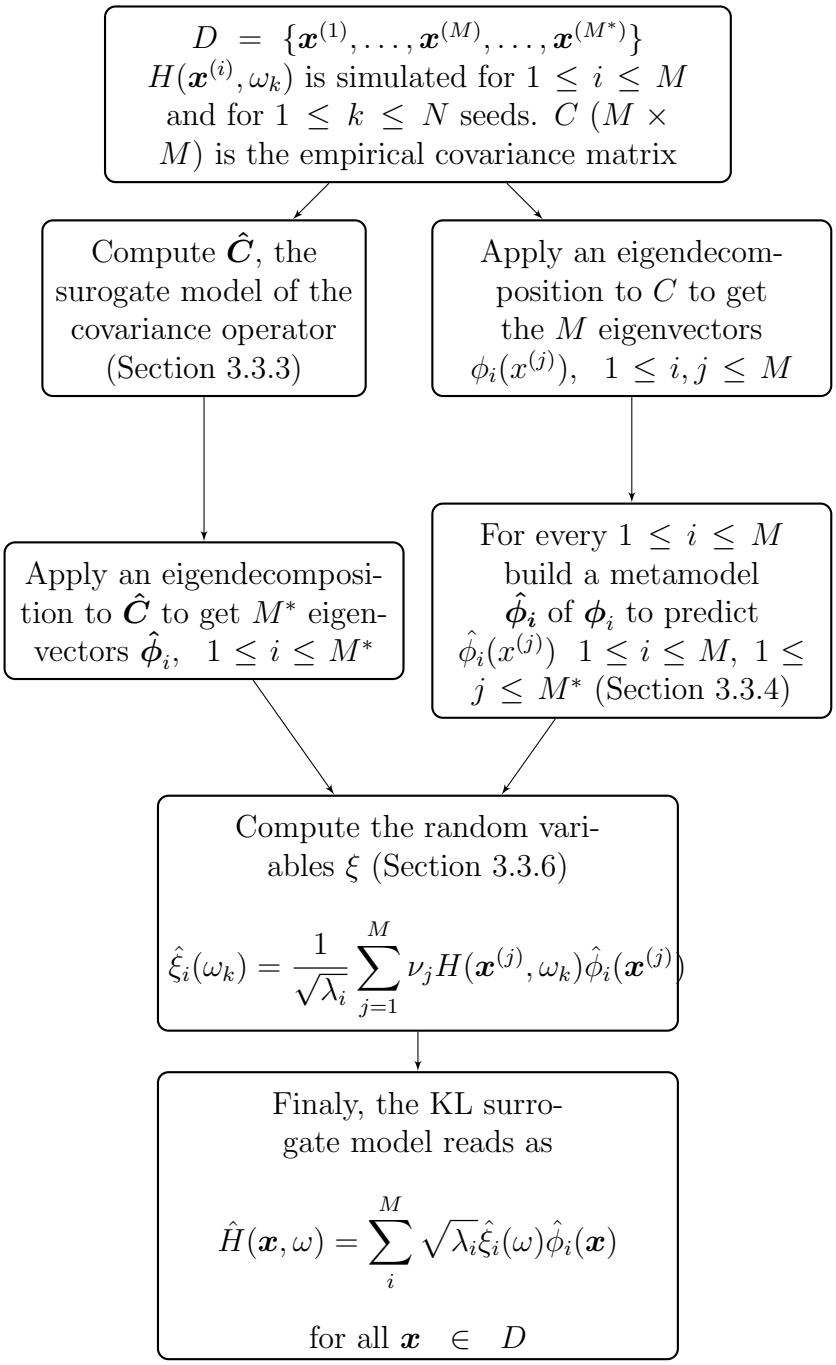


Figure 3.5: Flowchart summarizing the method and the two possible options (surrogate modeling the covariance -right, surrogate modeling the eigenvectors -left) for building up a surrogate model of H.

This study describes a non-parametric surrogate model of stochastic simulators based on Karhunen-Loève (KL) expansion. The approach has been tested first on closed-form processes in order to validate the method, and after that applied to a full scale problem linked to the assessment of a population exposure induced by base

station antennas (Chapter 5).

When dealing with Gaussian processes, the surrogate can be built in a simple way; the random variables computation is reduced to generating independent Gaussian random variables. The evaluation of the model accuracy is also simplified, it is reduced to comparing two deterministic quantities of interest; mean and variance. In the non-Gaussian case, there is much more to discuss.

The eigenvectors of the KL expansion in the domain of interest has been predicted in two different ways : at first a surrogate model of the process covariance operator using polynomial chaos expansions (PCE) has been used. The second approach consists in directly surrogating the eigenvector. In terms of performance, the error evaluation on the toy example shows better results when the eigenvectors are surrogated using the Kriging.

In this work, the KL expansion was not truncated. The M eigenvectors are all summed in Eq. (3.14) and no truncation is made. A perspective of improvement would be to explore the effect of a truncation scheme, for example, based on the most important eigenvalues.

For the demonstration example, and when the eigenvectors are interpolated using either Kriging or a linear interpolator, the tests performed do not show a significant difference in the overall performance. This is mainly due to the multiple steps governing the stochastic metamodeling procedure. Hence the eigenvector interpolation error fades away into the global error. Nonetheless the empirical covariance and its eigenvectors play a crucial role in the precision of the expansion (the PCE surrogate performed poorly).

Considering the error, the size of the DoE M, and the number of realizations N, impact the accuracy of the covariance matrix and the precision of its surrogate but also the accuracy of the random variables appearing in the KL expansion. The central limit theorem can be used to evaluate the error of the covariance matrix, but once the covariance or its eigenvectors are surrogated, we lose track of the analytical error, since errors from the surrogate model of the covariance, its parameters and the sampling over M points were added.

The fact that the randomness in the case study was 'controllable' (through freezing the same seed ω_k for different points of the DoE) is a key characteristic, since it enabled us to compute all the terms of the expansion.

Chapter 4

Global sensitivity analysis on stochastic simulators

Contents

4.1 General introduction

An important aspect when analysing computational models is the sensitivity analysis (SA), which consists on quantifying the influence of the input variables onto the output of the model. SA is usually performed directly on the computational model, or on its surrogate, and is a highly useful tool mainly to identify the most contributing inputs (or combination of inputs) that explain at best the variability of the output, but also to spot non-influential inputs in order to fix them to nominal values, hence reduce the dimension of the problem. Therefore, SA is essential to understand and explore the complex behaviour of the modeled system.

In a deterministic context, SA is performed most commonly using the variance-based approach where the variance of the output is expanded as a sum of contributions of each input variable, or their combinations [82]. Regression-based measures (like Pearson correlation coefficient) are also used for models with linear behavior. Alternative global SA methods are available such as the Morris method [60] as well as the moment independent indicators [13]. Some of those methods are presented in more details in Chapter 2.

In a more complex context, i.e., for stochastic simulators, where the output of the simulator in a given point is a random variable, sensitivity measures have been developed lately and are briefly presented in Section 4.2. In Section 4.3, our approach to perform sensitivity analysis on stochastic simulators is detailed; the stochastic simulator is represented as a stochastic process and the sensitivity analysis is performed on the differential entropy of this stochastic process. The approach is published and will be appearing soon [7]. The performance of the method is illustrated on a toy example (Section 4.3.4).

4.2 Literature review

When the computational model is more complex, other approaches are considered. For a computational model with functional outputs, for example, the objective is to detect input variables that impact the curve of the functional output; in [16], the SA is conducted on the coefficients of the expansion of the functional output in an appropriate set of basis functions. The basis functions can be either predefined like Legendre polynomials or a data-adaptative functions (e.g. principal components or partial least squares).

For an output that is a random variable, sensitivity analysis was applied on the mean and the dispersion of the random output [56]. In this case the sensitivity analysis results does not take into account the influence of the input on higher moments of the random variable output. Sobol' indices were evaluated for the mean and variance functions by performing two independent ANOVA (analysis of variance) decompositions on the two functions. As mentioned by the authors, it is not possible to combine both functional ANOVA decompositions of the mean and variance functions and forming merged SA indices remains an open problem. Authors from [56], through an example, concluded that the SA on the mean and the variance functions does not bring enough information to quantitatively estimate all the Sobol' indices. A similar approach was used in [44, 45].

Sobol' indices applied on stochastic models are not defined in a unique way. Mazo [58] introduces two kinds of Sobol' indices for stochastic simulators namely a first and a second kind. At first, Sobol' indices formula from Section 2.3 is reminded. Let $f \in L^2([0,1]^p)$, where p is the input $\boldsymbol{x}$ dimension. Sobol' indices are defined as

$$S_j = \frac{\text{Var}\,\mathbb{E}\left[f(\boldsymbol{x})|\boldsymbol{x}_j)\right]}{\text{Var}\left[f(\boldsymbol{x})\right]}, \quad j = 1, \ldots, p. \tag{4.1}$$

For a stochastic simulator $H(\boldsymbol{X}, \omega)$, $\omega \in \Omega$ and $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_d\} \in D \subset \mathbb{R}^d$, first kind Sobol' indices [58] are defined as

$$S'_j = \frac{\text{Var}\,\mathbb{E}\left[H(\boldsymbol{X}, \omega)|\boldsymbol{X}_j\right]}{\text{Var}\left[H(\boldsymbol{X}, \omega)\right]}, \quad j = 1, \ldots, p. \tag{4.2}$$

By supposing that ω in $H(\boldsymbol{x}, \omega)$ is just another input of the stochastic simulator, S'_j in Eq. (4.2) is a direct application of Eq. (4.1) for $H(\boldsymbol{x}, \omega)$. The second kind Sobol' indices defined in [58] arises from the definition of Sobol' indices (Eq. (4.1)) applied to the mean function of $H(\boldsymbol{x}, \omega)$: $\boldsymbol{x} \to \mathbb{E}[H(\boldsymbol{x}, \omega)|\boldsymbol{X} = \boldsymbol{x}]$. The second kind Sobol' indices are defined as

$$S''_j = \frac{\text{Var}\,\mathbb{E}\left[\mathbb{E}[H(\boldsymbol{X}, \omega)|\boldsymbol{X}]|\boldsymbol{X}_j\right]}{\text{Var}\mathbb{E}[H(\boldsymbol{X}, \omega)|\boldsymbol{X}]}, \quad j = 1, \ldots, p. \tag{4.3}$$

In [58], the indices are estimated using Monte Carlo simulations, and an optimal number of realization N and exportation M is introduced.

The lack of approaches that address the sensitivity of the output as a random variable to the model inputs, is obvious. In the following sections, a parametric method is proposed to evaluate the sensitivity of the stochastic model output to the inputs as detailed in Section 4.3. The performance of the method is evaluated in Section 4.3.4 through a toy example. Finally Section 4.4 concludes and discusses the approach and its results .

4.3 The method

The interest is to quantify the sensitivity of the random output to the model input variables. This is achieved by reducing the output random variable to its differential entropy. Thus instead of considering the sensitivity of the stochastic model, the sensitivity of the differential entropy of the stochastic model is considered. In practice, following the sampling of the stochastic model on a predefined design of experiments set (DoE), differential entropy is evaluated on each DoE point. The next step consists of building a surrogate model of the differential entropy of the stochastic process to then apply standard methods of sensitivity analysis (SA), in this case, via evaluating Sobol' indices. Subsections 4.3.1, 4.3.2 and 4.3.3 give detailed explanations of the techniques used and provide further insight on the choice of differential entropy. The method is summarized on the flowchart (Figure 4.1), where DoE is the design of experiments set with M points, $H(x, \omega)$ is the stochastic model and ω is the random seed. To be able to compare with state-of-the-art methods [56], mean and variance of the stochastic model are also evaluated, on which SA is performed.

Monte Carlo simulations are often used to evaluate the Sobol' indices [83] using direct calls to the simulator. However, in the context of this thesis, for the following reasons, it was preferable to lean toward using a surrogate model:

- Surrogates offer a much cheaper option to the expensive calls to the true models. The mathematical function mimics the model, and predicts behaviours of inputs of interest with good accuracy.

- Often times, such surrogates offer the possibility to compute the Sobol' indices with a mere post processing step (e.g. Sobol' indices can be computed analytically from polynomial chaos expansions [87]).

4.3.1 Differential entropy

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the event space, $\mathcal{F}$ its σ-algebra and $\mathbb{P}$ its probability measure. Let Y be a random variable, $p(y)$ its probability density function (PDF) and S its support set i.e a set where $p(y) > 0$. Differential entropy $h(Y)$ of a continuous random variable Y (if it exists) is:

$$h(Y) = -\int_S p(y) \ \log \ p(y) dy. \tag{4.4}$$

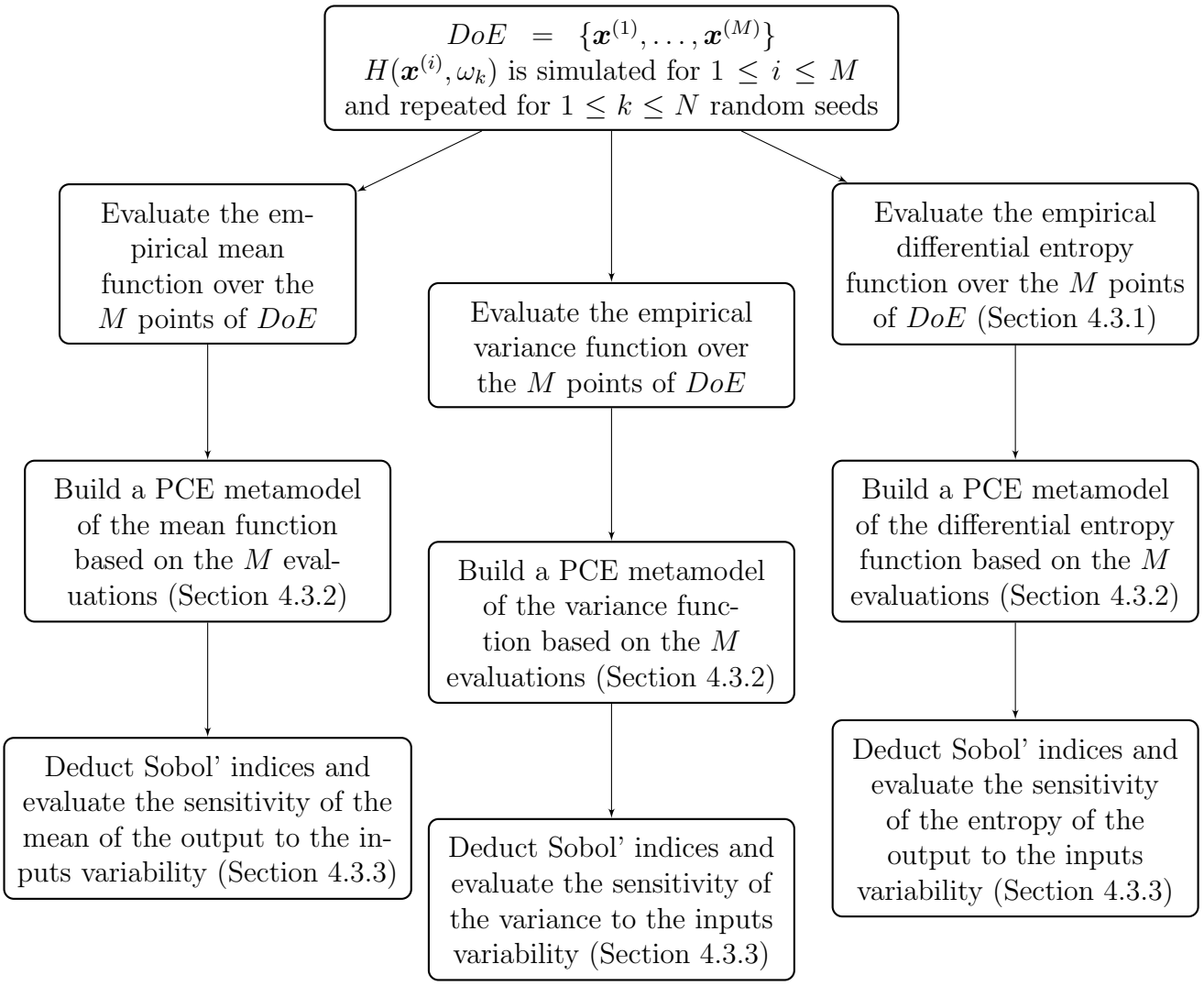


Figure 4.1: Flowchart summarizing the SA method for stochastic simulators.

Differential entropy can be negative, is translation invariant, but is, in general, variant to any transformation from a random variable to another [26], for example:

$$h(aY) = h(Y) + \log|a|. \tag{4.5}$$

The Gaussian distribution maximizes h over all distributions with the same mean and variance as proven in [10] by maximizing the Lagrange function under constraints of mean and variance. Similarly, the maximum entropy of a continuous random variable having values in a finite-length interval $[a, b]$ is attained for a uniform distribution on $[a, b]$ [74].

Entropy is wildly used in statistics, with the maximum entropy principle being one of the well known applications. It states that the PDF that best represents a given set

of data is the one with maximum entropy among all those that satisfy the constraints of prior knowledge. In [68], entropy was also used as a space filling technique where the entropy of the drawn sample is estimated then maximized. Finally, in [5], entropy was used to evaluate the global SA of a deterministic model, in [49] the indices were based on conditional entropy whereas in [52] they were based on Kullback-Leibler entropy. All in all, other research areas highlighted interest in this measure mainly due to the meaning that entropy contained. Originally, Shannon called it 'uncertainty', before changing the name to entropy [74]. In our case differential entropy infers how confined or spread the random variable is. Hence low entropy implies that the random variable is confined to a small effective volume and high entropy indicates that the random variable is widely dispersed.

In practice, differential entropy is evaluated by dividing the range of the sampled random variable $\boldsymbol{Y} = (y_1, \ldots, y_N)$ into bins of length Δ, then assuming that the density is continuous within the bins, the mean value theorem tells us that, for each bin, there must exist a value y_i such that:

$$p(y_i) = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} p(y)dy. \tag{4.6}$$

The quantized random variable $\boldsymbol{Y}^{\Delta}$ is defined as $\boldsymbol{Y}^{\Delta} = y_i$. If $i\Delta \leq \boldsymbol{Y} \leq (i+1)\Delta$, then

$$\mathbb{P}(\boldsymbol{Y}^{\Delta} = y_i) = \int_{i\Delta}^{(i+1)\Delta} p(y)dy = p(y_i)\Delta. \tag{4.7}$$

If $p(y)\log p(y)$ is Riemann integrable, then as $\Delta \to 0$, $-\sum \Delta p(y_i)\log p(y_i)$ approaches $h(Y)$.

In practice, there are many rules that can be used to decide the bin number (e.g. Sturges rule [85], Scotts rule [80]). With Sturges rule $\lceil \log_2(N) \rceil + 1$ bins are advised. For a data set where $N = 50$, 7 bins are recommended with this rule.

4.3.2 Surrogating the entropy

Following the computation of entropy over the *DoE* set, a surrogate model is built to emulate the entropy of the stochastic output. Here, the PCE metamodel is used (see Section 2.2.2 for more details). For completeness sake, polynomial chaos expansion is briefly reviewed.

Polynomial chaos expansion

Polynomial chaos expansion (PCE) approximates the dependence of the model outputs on the model inputs by expansion in an orthogonal polynomial basis $\{\Psi_{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{N}^d\}$ with respect to the joint PDF of the input parameters [32] [88].

Consider a deterministic numerical model g with independent inputs gathered in a random vector $\mathbf{X} \in D \subset \mathbb{R}^d$ with a joint probability density function $p_{\boldsymbol{X}}$. Suppose that the model response $g(\mathbf{X})$ has a finite variance, i.e. $\mathbb{E}\left[g(\mathbf{X})\right]^2 < \infty$, $g(\mathbf{X})$ can be expressed as follows :

$$g(\mathbf{X}) = \sum_{\boldsymbol{\beta} \in \mathbb{N}^d} a_{\boldsymbol{\beta}} \Psi_{\boldsymbol{\beta}}(\mathbf{X}), \tag{4.8}$$

where $a_{\boldsymbol{\beta}}$ are unknown deterministic coefficients and $\Psi_{\boldsymbol{\beta}}$ are multivariate polynomials obtained as tensor products of univariate polynomials of degree $(\beta_1, \ldots, \beta_d)$.

Several methods exist to calculate the coefficients $a_{\boldsymbol{\beta}}$ of the PCE for a given basis, namely using projection methods [96] where the expansion is projected onto the polynomial space, or by casting a least-squares minimization problem [8, 11]. A nice feature of PCE is the simplicity with which one obtains the most commonly used statistics of the quantities of interest: mean, variance as well as Sobol' sensitivity indices [82, 78], which can be computed analytically from the estimated coefficients [89].

4.3.3 Sensitivity analysis of the entropy

Following the construction of the surrogate model of differential entropy, the next step consists of evaluating the SA of the input to the entropy of the output by estimating the Sobol' indices, a well-known global SA approach, in which proportional values of the variance of the inputs to the output are evaluated.

Sobol' indices are based of the ANOVA decomposition of the variance function [37] and defined as in Eq. (4.1). Sobol' indices are described in details in Section 2.3.

Remark. *The function f from Section 2.3 is a function describing the stochastic process, it can be either the mean, variance or entropy of the stochastic process $H(x, \omega)$.*

Remark. *The use of differential entropy in this chapter can be seen as a parametric representation of the random process; instead of considering the random output, its entropy is rather considered. Not to be confused with entropy-based sensitivity analysis introduced in Section 2.3.2.*

4.3.4 4-dimensional analytic example

For demonstration purposes, sensitivity analysis is evaluated on a four-dimensional analytic example based on the rotated hyper ellipsoid function,

$$H(\boldsymbol{x}, \omega) = x_1^2 + \omega_1(x_1^2 + x_2^2) + \omega_2(x_1^2 + x_2^2 + x_3^2) + \omega_3(x_1^2 + x_2^2 + x_3^2 + x_4^2), \qquad (4.9)$$

where $\omega_1 \sim \Gamma(10, 1),\ \omega_2 \sim \Gamma(7.5, 1),\ \omega_3 \sim \Gamma(2, 2)$. $\Gamma(\alpha, \beta)$ refers to the gamma distribution with shape parameter α and rate parameter β. A $DoE \subset [-65, 65]^4$ of size $M = 50$ is generated using Latin hypercube sampling. The simulations were repeated $N = 10^5$ times for each point of DoE. Differential entropy is empirically evaluated over the 50 points. For each point from DoE, 10 bins were used to evaluate differential entropy based on the 10^5 samples. A PCE metamodel of the mean, variance and entropy is then built and the respective leave-one-out errors are $4.2 \cdot 10^{-28}$, $1.8 \cdot 10^{-28}$ and 0.012 (LARS was used to evaluate the PCE coefficients [28, 12]).

The Sobol' indices were evaluated using the PCE surrogate model.

Variable	Sobol' indices	Mean	Variance	Entropy
X_1	total	0.4534	0.4784	0.4003
	first order	0.4534	0.4307	0.3371
X_2	total	0.4139	0.4381	0.3617
	first order	0.4139	0.3915	0.2819
X_3	total	0.1184	0.1262	0.2455
	first order	0.1184	0.1058	0.1962
X_4	total	0.0143	0.0162	0.1087
	first order	0.0143	0.0132	0.0783

Table 4.1: Total and first order Sobol' indices for the mean, variance and entropy of $H(x, \omega)$ from the analytic example.

Since differential entropy is translation invariant, the sensitivity of the model to the mean of the stochastic process needs to be explored as well. Sobol' indices are also evaluated for the variance function, the objective being to compare the SA from (mean, variance) to (mean, entropy). As detailed in Table 4.1, the three approaches rank the variables similarly. The main difference is that the interactions between variables are more pronounced in entropy case, namely 8% of the variance of entropy is due to interactions between X_2 and other variables. Unlike the mean and variance-based SA, variables such as X_3 and X_4 are not negligible for entropy-based SA.

As demonstrated through this example, the use of the entropy as a measure of interest complements the state of the art approaches, mainly based on evaluating the sensitivity on mean and variance.

4.4 Conclusions

This chapter presents a novel approach to assess the sensibility of a stochastic simulator by considering differential entropy as a measure of uncertainty on the output distribution given a set of inputs. When dealing with simulators with random outputs, and when the interest is focused on the probability density function of the output rather than on one of its moments (mean, variance ...), considering the entropy of the random output is useful. Any significant fluctuation of the value of entropy can be recognized as a sensitivity to the input variable causing it, but as any numerical approximation, entropy is sensitive to the choice of the metamodel and to the bin number.

The method proposed is agnostic as to the machine learning technique used, it is up to the user to choose metamodels that are adequate to the dimension of the inputs, the size of the design of experiments and to the properties of the models response.

Relying on selected moments such as variance to describe the random output can be restrictive since two different probability density functions can have the same value for a specific moment. It is though less likely for two different probability density functions to have the same mean, variance, and differential entropy at the same time, that's why more than one indicator is used.

Since entropy is translation invariant, the SA performed on entropy is paired with the one performed on the mean. The method is particularly efficient since differential entropy is more general than the variance. Not only does it enclose the variance, but it also demonstrates how the random variable scatters randomness in its support.

Chapter 5

Application to computational electromagnetic dosimetry

Contents

The scope of the thesis was briefly introduced in Chapter 1. This chapter now describes in more details the context and the tools used to address the problematic of the thesis, namely the study of the impact the morphological features of a city have on the resulting human exposure induced by a base station antenna. The first sections summarize the basic and advanced notions of electromagnetic dosimetry, introduced in a simplified way so that a non-experimented reader can follow the developments.

5.1 The human exposure

Wireless communication means were revolutionary, to say the least, to the telecommunications industry. The use of electromagnetic waves for wireless communication is not new; Guglielmo Marconi first began developing a wireless telegraph system using radio waves in 1894. His contributions to wireless communication led him to be awarded the Nobel prize in 1909 (shared with Karl Ferdinand Braun). For a long time, firefighters, hospitals and police used radio waves to communicate but it took until the 1990's for wireless telephone networks to proliferate and induce a social revolution.

Now wireless communications play a significant role in people everyday lives, and the extremely rapid technological evolution results in phenomenal changes in the usage of wireless devices, enabling voice traffic, digital data exchange, etc. (Figure 5.1). Along with this sweeping success, concerns about possible health side effects related to the exposure to the radio frequency (RF) electromagnetic radiations hve emerged, giving birth to electromagnetic dosimetry.

To protect people from overexposure induced by electromagnetic field (EMF) radiations, compliance tests and safety standards were defined by European and international bodies such as ICNIRP[1] [41], FCC[2], the European Council [71] and IEEE SCCs[3] [3]. The guidelines provided by such organizations define exposure levels above which the exposure becomes detrimental. For example, standards defined in [42] aim at putting 'safe' devices onto the market, in other words devices that satisfy the requirements imposed. Here the focus is rather on the actual exposure. Epidemiological studies emphasize the actual exposure assessment typically by computing the specific absorption rate (SAR) that is derived from the electric and magnetic fields induced by the wireless system.

[1]International commission on non-ionizing radiation protection
[2]Federal communication commission
[3]Standards coordinating committees

The vast array of wireless technologies and the different generations of cellular networks (GSM[4], UMTS[5], LTE[6]), stress the importance of monitoring and managing EMF exposure.

Figure 5.1: Day-to-day exposure of a population [1].

In order to further characterize the real exposure, the next Section 5.2 introduces a measure to evaluate the exposure to EMF, namely the specific absorption rate (SAR).

5.2 Computational dosimetry

The human exposure to RF EMF represents the power absorbed by the tissues of the human body and is characterized by the SAR value in watt per kilogram. SAR measures the rate at which the power is absorbed in a human body exposed to RF EMF. The whole-body SAR can be evaluated by assessing the power absorbed per unit mass of the body tissue.

$$SAR_{whole-body} = \frac{P_{abs}(body)}{m(body)} \quad (W/kg). \tag{5.1}$$

[4]Global system for mobile communications (2G)
[5]Universal mobile telecommunications system (3G)
[6]Long term evolution (4G)

SAR can be assessed locally, for a fetus [46] for instance, or for an organ like the brain.

The most popular formula of SAR directly links SAR to the magnitude of the incident electric field $E(V/m)$, to the conductivity $\sigma(S/m)$ and to the density of the tissue $\rho(kg/m^3)$

$$SAR = \frac{\sigma E^2}{2\rho} \quad (W/kg). \tag{5.2}$$

Eq. (5.2) is a consequence of Maxwell equations [57], a mathematical model describing how electric and magnetic fields are generated by charges, currents, and changes of the fields. The set of equations is described here for the sake of completeness.

Let $\boldsymbol{E}$ denote the electric field, $\boldsymbol{B}$ the magnetic field, J the electric current density, ϵ the permittivity and μ the permeability. The $\boldsymbol{\nabla\cdot}$ symbol denotes the divergence operator meanwhile $\boldsymbol{\nabla\times}$ denotes the curl operator. Maxwell's differential equations read:

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= \frac{\rho}{\epsilon}, \\ \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0, \\ \boldsymbol{\nabla} \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, \\ \boldsymbol{\nabla} \times \boldsymbol{B} &= \mu(\boldsymbol{J} + \epsilon \frac{\partial \boldsymbol{E}}{\partial t}). \end{aligned} \tag{5.3}$$

The reader can consult [95] for further details on how the SAR formula is derived from Maxwell equations.

SAR can be assessed using measurements. In this case the electric field is measured and the SAR can be deduced by Eq. (5.2). However performing measurements is not always possible in particular it is impossible to measure inside the human body for invasive concerns. To overcome this limitation, and based on realistic heterogeneous body models (e.g. Figure 5.2), EMF is assessed using numerical methods, mainly finite-difference in the time-domain (FDTD [90]). The main drawback of such numerical method is the prohibitive computational cost.

5.3 Exposure induced by base stations

The sources of electromagnetic field are various, and can be located in a near field like WiFi boxes or cell phones, very close to the human body. They can also be located in far locations like a radio base station on a building roof. The exposure is modelled differently depending on the location of the source, its characteristics and usage.

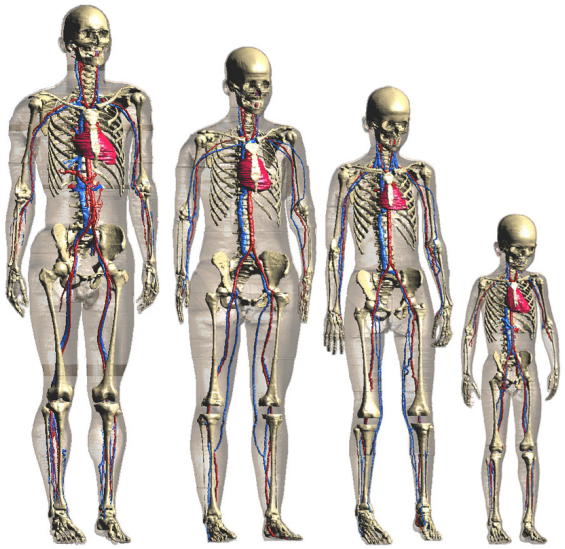


Figure 5.2: The virtual family model: Duke, Ella, Billie and Thelonious (from left to right) [21].

In this work, base station antennas are considered. During the last two decades base stations antennas have blown up in numbers in rural, urban, and sub-urban areas. They are typically 1 to 2 meter-long arrays of antennas with gains between 15 and 21 dBi placed in towers between 25 and 75 meters above ground. They have high aperture efficiencies and are able to handle a very high power (Figure 5.3).

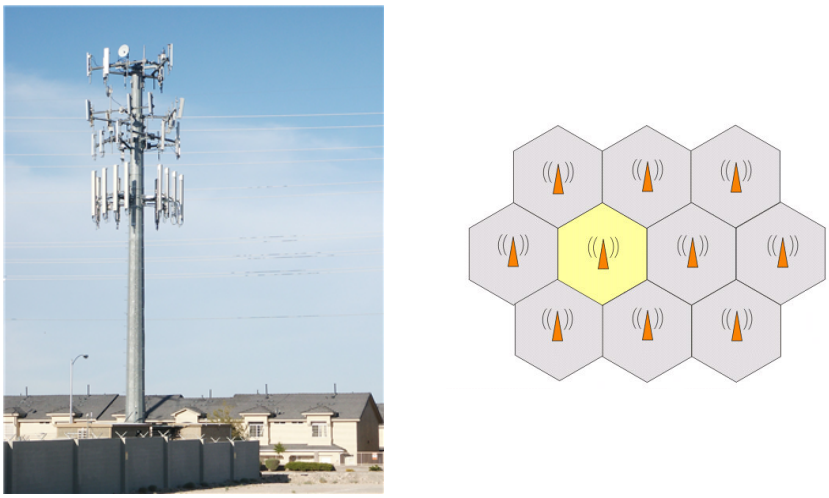


Figure 5.3: Cell phone base station antennas on a roof (left) - A small cellular network of uniform cell size (right).

Macrocell base stations are used in cellular networks to provide radio coverage over very large areas (several hundreds of meters in urban areas to several kilometres in rural areas). A large number of antennas each covering limited areas of land (called

cells) make a cellular network (Figure 5.3). When a mobile phone is within these cells it is possible to make voice calls and transfer data through these base stations. These base stations emit power at all time. The emitted power depends on different parameters such as technology, user traffic and cell capacity.
Since SAR assessment is a complex and time consuming matter, studies were conducted to define functions assessing the absorbed power by the human body from the incident field. These functions are called *transfer functions*. Many studies were carried in order to characterize these transfer functions in terms of variabilities of SAR through advanced numerical methods. The relationship between the incident field and the absorbed power depends on a few parameters [23, 95], namely the wave frequency (Figure 5.4) and the intensity of the incident field. In addition, the body posture, the body morphology and the conductivity of the organs impact the absorbed power. As a matter of fact the absorbed power varies (up to 40%) in terms of morphological parameters (Figure 5.4). When adults and children are exposed to the same incident field, the latter will have a higher absorbed power due to their body size (see Figure 5.5). The angle of incidence of the EM wave also impacts the absorbed power: more power is absorbed if the field directly hits the front or the back of the human body [23].

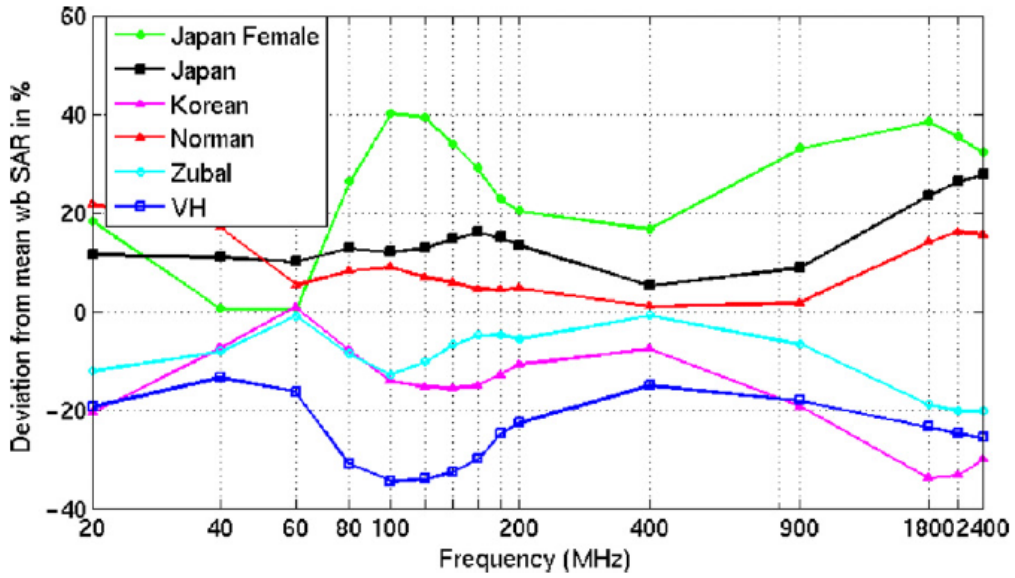


Figure 5.4: Deviation of the whole-body SAR versus frequency for different numerical human models [24].

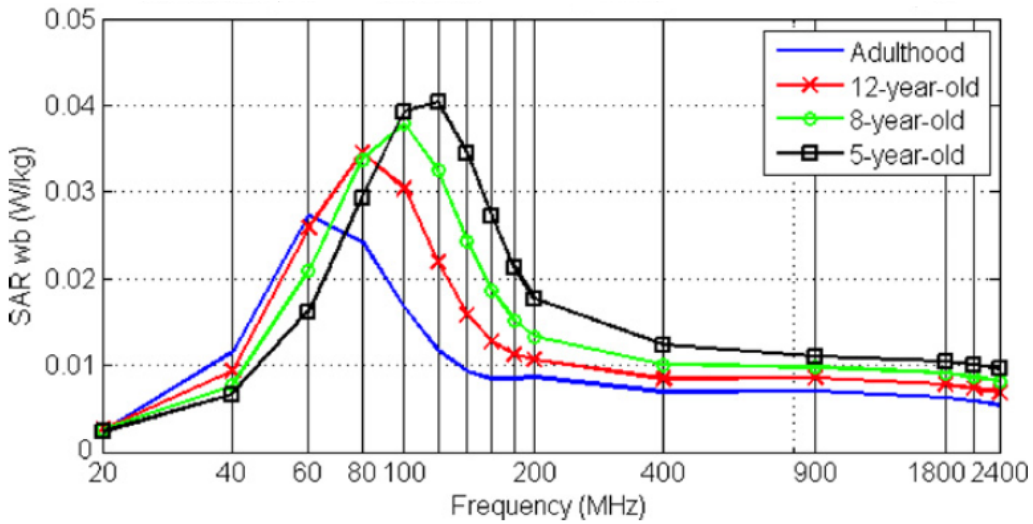


Figure 5.5: Whole-body SAR versus frequencies for different ages [24].

5.4 Path loss exponent

This section introduces tools to evaluate the exposure in a real context, i.e. the day-to-day exposure of a population to RF-EMF in a typical urban environment.

The SAR formulation in Eq. (5.1) requires the evaluation of the power absorbed, which is equal to the power emitted by the RF-EMF source minus any losses. The losses are modelled using propagation models. Propagation models should take into consideration the influence of building, field data and RF waves propagation. In this respect, advanced computational tools were used in many studies [40, 94, 39] to characterize the signal attenuation between the transmitter and the receiver. The propagation model can be approximated using a path loss model [70]:

$$P(d) = \beta - 10\alpha \log_{10}(d) + \mathcal{N}(0, \sigma^2). \tag{5.4}$$

In this equation P is the received power, d the distance between the transmitter and the receiver, α is the path loss exponent (PLE), β a constant and $\mathcal{N}(0, \sigma^2)$ is a Gaussian noise with zero mean and variance σ^2. α and β can be computed by linear regression for a minimum mean-squared estimate (see Section 2.2.1). The path loss exponent represents the attenuation of energy between the transmitter and the receiver, and depends on the frequency, antennas height, and propagation environments. In free space, i.e. a region free of all objects that might affect RF propagation by absorption, reflection, or refraction, it is shown that $\alpha = 2$. In the presence of a very strong guide wave phenomenon (tunnel effect) α can get lower that 2. In

general, when obstacles are present α is larger. Hence the PLE widely varies across propagation terrains. The aim of this chapter is to further study this variability.

5.5 Stochastic city generator

The path loss exponent presented in Section 5.4 is typically evaluated using measurements. Lately, and without any need for in-situ measurements, a statistical model [98, 40] based on virtual cities, generated using stochastic geometry is used to evaluate the PLE. Authors of [98] developed a framework implemented in C++ called *GeoStat*. Other features were added to GeoStat in [31].

To generate a city, some parameters are fixed in the first place. Table 5.1 represents some of the parameters that the user can select. To see the full list of city parameters the reader can refer to [31, 98].

Parameters	Definition domain
Street width (mean value)	$\mathbb{R}^+$
Building height (mean value)	$\mathbb{R}_0^+$
Building facade length (mean value)	$\mathbb{R}_0^+$
Anisotropy	$[0, 1]$
Edges of the simulation window	$\mathbb{N} \setminus \{0, 1, 2\}$

Table 5.1: Values for some morphological features of a typical urban city.

The skeleton of the city is represented by a polygon to which a tessellation is applied. A tessellation is a countable family of convex polygons partitioning $\mathbb{R}^2$ and whose interiors do not intersect. The most used tessellations in modeling street systems are Poisson Line [20] and Crack STIT [62] tessellations. This partitioning step is repeated in a recursive way (Figure 5.6). The following step consists on applying an erosion and dilatation to the edges of each polygon. The building footprints are then created by dividing the inner surface between the two polygons (between erosion and dilatation). A random height is then associated to each building. The value of the height of each building is drawn from a distribution, the mean value of which is picked by the user (Table 5.1). Figure 5.7 from [25] reviews the different steps to build a virtual city.

Figure 5.8 shows the final results for different values of anisotropy. The morphology depends on the choice of the parameters governing the city. However, due to the random processes involved in the construction of such virtual cities, even for two

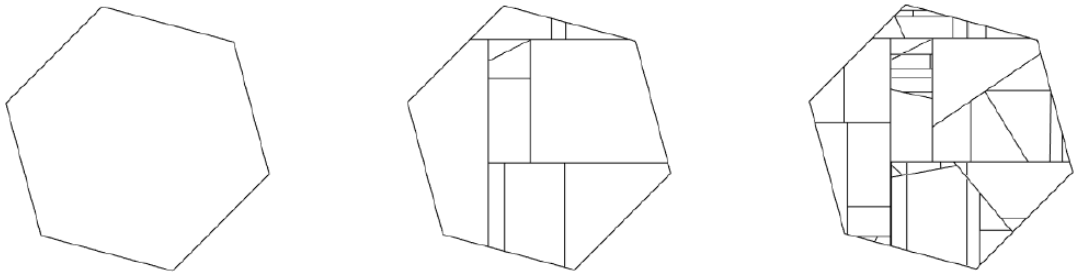


Figure 5.6: Realisation of tessellations on a virtual city.

cities governed by the same parameters, the morphology and the buildings in the city are not similar.

5.6 Ray tracing

Once the virtual city is finalized, the assessment of the exposure is performed using ray tracing techniques used to propagate EMF in urban areas. In a given stochastic city model, an antenna has been placed somewhere in the city (position fixed by the user) operating at a fixed frequency. N rays are then launched from the source (Figure 5.9). The launched rays produce reflections and diffractions and a portion of their power is also absorbed by the surface. The signal attenuation map can be obtained by assessing the received power in the measurement plane, 1.5 m above the ground to represent the human exposure.

GeoStat enables the user to choose parameters related to the ray tracing step, some of which are listed in Table 5.2. The full list can be found in [25] and the updated list in [31].

Parameters	Definition domain
Number of antennas	$\mathbb{N}^*$
Power gain after reflection	$\mathbb{R}^-$
Wavelength of the antenna	$\mathbb{R}^+$
Maximum reflections per ray	$\mathbb{N}^*$
Number of rays	$\mathbb{N}^*$

Table 5.2: Some parameters governing ray launching in a typical urban city.

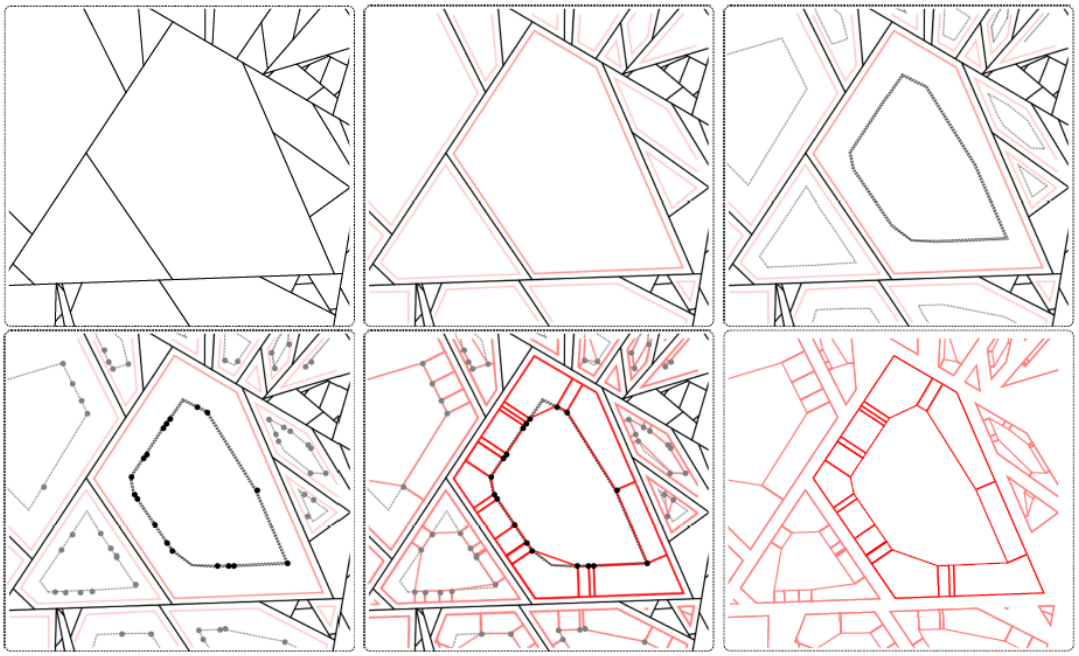


Figure 5.7: Steps to generate a virtual city: left to right: (1) a tessellation (2) an erosion is applied to each polygon (3) in each new cell, the dilated polygon with respect to its center of mass is computed (4) a Poisson point process is drawn on the edge of the polygon (5) those points are projected to create buildings footprints (6) the final result [25].

5.7 Statistical analysis of PLE in urban environment

Based on the tools presented above, i.e., the virtual city generator and the ray tracing technique, the power absorbed can be computed and the PLE can be fitted using Eq. (5.4). This process of computations (from a virtual city to the PLE evaluation) is what we call here a stochastic city simulator.

The aim is to further study the variability of the PLE. For a given city, determined by few parameters, the PLE α is evaluated with the help of the stochastic city simulator. By running the simulator for the same city, PLE will be different due to the stochasticity involved while building a city and launching rays. Hence, by running the simulator for the same city, multiple times, a distribution of α can be attributed to the city parameters.

The objective is to further explore the impact that the city characteristics has onto the exposure (or onto the PLE). This can be done by evaluating the sensitivity of the simulator to the variation of its inputs. First of all, the stochastic simulator input

Figure 5.8: Footprint of three virtual cities with different anisotropy: from left to right, anisotropy values are: 0, 0.5 and 1 also called a Manhattan-like city.

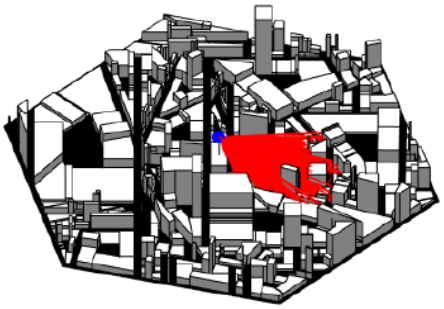


Figure 5.9: 3D view of ray tracing in a virtual city.

variables are selected among all the variables governing the stochastic city simulator, while the irrelevant ones are set to nominal values. Table 5.3 represents the considered variables. Since the aim here is to explore only the impact the geometry of a city can have onto exposure, the variables describing the EMF are set to nominal values.

Input variables	Range
Street width X_1	$[10\ m, 20\ m]$
Building height X_2	$[9\ m, 18\ m]$
Anisotropy X_3	$[0, 1]$

Table 5.3: Input variables for the stochastic city generator.

Let $H(\boldsymbol{x}, \omega)$ be the stochastic city generator, $\boldsymbol{x}$ is the set of variables governing the city (Table 5.3), and ω the randomness arising from building a city and launching rays. To generate data from the simulator, a design of experiments (DOE) is at first built using Latin hypercube sampling (LHS). This technique is briefly presented in

the next section.

5.7.1 Generating the design of experiments

LHS is a pseudo-random sampling method used to design a more efficient sample set than Monte Carlo methods. A Latin square is a square grid containing sample positions where there is only one sample in each row and each column. A Latin hypercube is the generalisation of this concept to an arbitrary number of dimensions, whereby exactly one sample is drawn in each axis-aligned hyperplane containing it. This strategy does not prevent possible bad space filling. To offset this limitation, the *maximin* criteria is advised. A number of LHS designs are sampled, and the one that maximises the minimum distance between the points is selected.

Figure 5.10 represents the LHS of the input variable $\boldsymbol{X} = \{X_1, X_2, X_3\}$ for the experiment. $M = 30$ points were selected, mainly because of the high computational cost of the stochastic simulator. Indeed one run takes more than one hour (by means of a computer of type Intel Xeon E5-2620v3 2.4 GHz 6 Core 15 Mo and Nvidia Tesla K80). In such situations building a surrogate model enable the use and exploration of the characteristics of the stochastic model. Section 5.7.3.1 tackles the building of a metamodel to the stochastic process using the method introduced in Chapter 3.

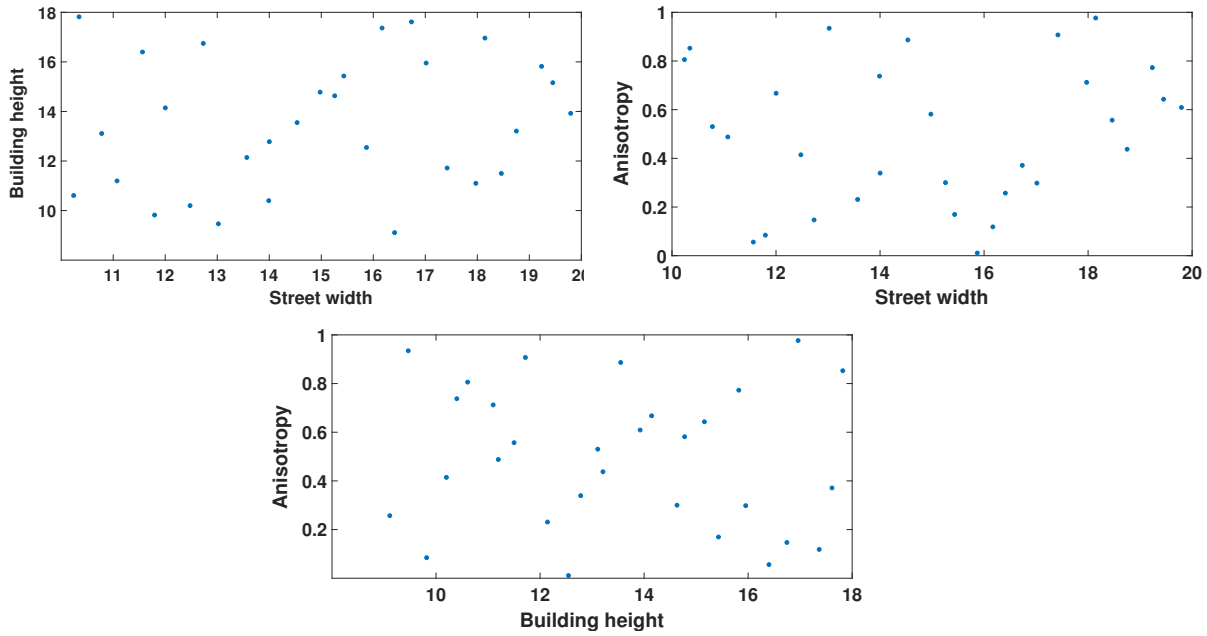


Figure 5.10: Projection onto the three dimensions of the 30 *DOE* points selected using LHS.

5.7.2 PLE distribution using stochastic cities

Following the sampling on the deterministic variables of the stochastic process, it is now time to run the stochastic simulator. As mentioned previously, the output (PLE) will not be the same if the call to the simulator is repeated for the same input $\boldsymbol{x}$, unless the random seed in the code is fixed, which is something feasible with GeoStat. In this case, it is possible to *freeze* the randomness (say ω_k) and regenerate $H(\boldsymbol{x}, \omega_k)$, it is also possible to run twice the simulator for two different points $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ with the same random seed ω_k. The seed is used to initialize a pseudo-random number generator in the stochastic city generator. By freezing the trajectory of the process, two cities with the same seed number and the same parameters are exactly the same, and accordingly their path loss exponents are identical. The random seeds are still unkown, and not controllable by the user. They can only be frozen. In this experiment 50 seeds were considered.

Using the stochastic city simulator, 8×10^5 rays have been generated and launched from an 30 m high antenna located in the center of a city measuring 360000 m^2 and fully determined by a seed number and three input variables detailed in Table 5.3 (other parameters such as ground and building properties were not addressed in this case study).

Following the ray tracing, the received power and the corresponding distances for each ray hitting the city are collected. They are then used to predict α for this city by casting a least-squares minimization problem on Eq. (5.4). So in total 1500 minimization problems were solved to get the values of α on different cities, and for different seeds.

The *DoE* is a LHS of 30 cities for the 50 seeds, meaning in total $30 \times 50 = 1,500$ simulations. The simulations lasted over three months.

5.7.3 Uncertainty quantification

Admittedly, the objective of the thesis was to explore the impact a city has on the exposure of the population, which is done by performing sensitivity analysis on the stochastic simulator. Yet for all the reasons mentioned before (the code is costly and stochastic), a user could not get hold of any kind of data or statistical characteristics of the model easily. In addition to the 1.5 hour per run, preparing input files, the use of adequate computational tools, and the post processing of the data (computation of PLE after the ray tracing) is just not practical to perform any kind of inferential statistics, let alone predictions.

The need to build a metamodel came from the stated reason, but also to establish a direct link between any city parameters and the corresponding PLE probability density function (PDF).

5.7.3.1 Metamodel of PLE

The non-parametric surrogate model based on Karhunen-Loève (KL) expansion was used here. The eigenvectors of the KL expansion in the domain of interest have been predicted in two different ways: at first a surrogate model of the covariance operator has been used using polynomial chaos expansions (PCE). The second approach, consists in directly surrogating the eigenvector.

Among the 30×50 runs, 10% of the data are for testing. A k-fold cross-validation was carried out by dividing the data onto $k = 10$ subsets. At each step a surrogate model of the stochastic simulator (Chapter 3) is built using nine out of the ten subsets. The remaining subset is used to evaluate the performance of the model (Figure 5.11). This procedure is then repeated for 100 different partitions of the data set. To evaluate the performance of the metamodel on the test points, metrics from Section 3.3.7 are used. The results are presented in Table 5.4.

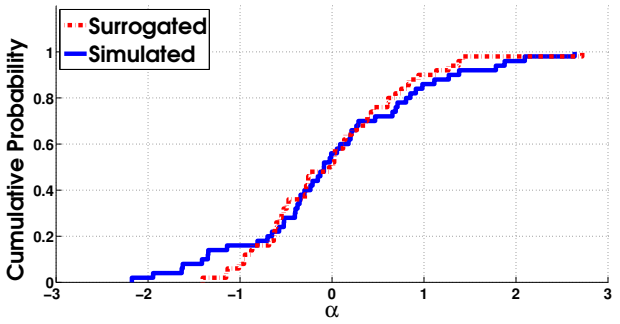
Method	Histogram intersection	Hellinger distance	JS divergence
Linear interpolator of eigenvectors	0.74	0.15	0.02
Kriging surrogate of eigenvectors	0.71	0.17	0.03
PCE covariance surrogate	0.76	0.14	0.02

Table 5.4: Mean error estimators over 3,000 test points.

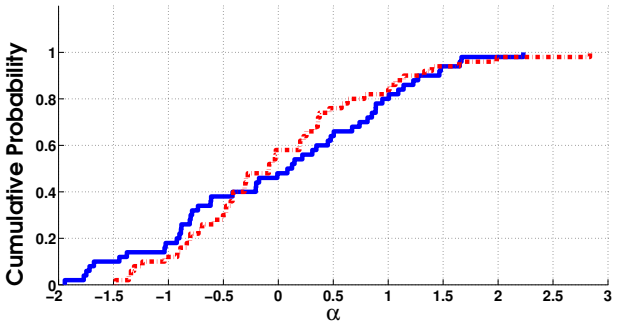
In this case, and for $M = 30$, $N = 50$, the PCE performs slightly better than the other options. The ranking of the three approaches depends on the process. For instance, for the toy example in Section 3.3.8, the ranking was the other way around and the performance of PCE was the worst among the tested interpolators.

For this example, the dependence of the method on M and N could not be evaluated since only 30 points were simulated (due to the high computational cost).

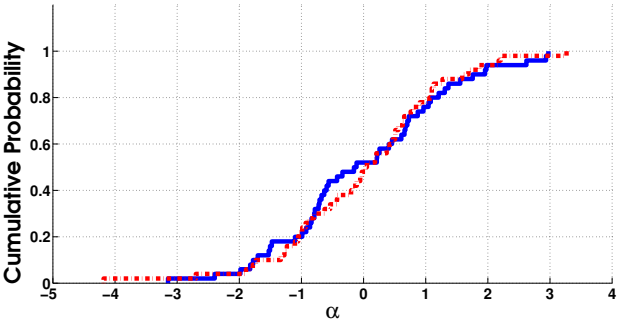
The KS test is used to test the null hypothesis: the predicted and the sampled PDFs come from the same distribution. Results are 4.8%, 12.6%, 1.46% of rejection of the null hypothesis (at level 5%) for the respectively linear surrogate of $\boldsymbol{\phi}_i$, the Kriging surrogate of $\boldsymbol{\phi}_i$, and the PCE surrogate of the covariance C.



(a) Linear interpolation of eigenvectors (mean histogram intersection is equal to 0.74), $X = (18, 13, 0.4)$.



(b) Kriging surrogate of the eigenvectors (mean histogram intersection is equal to 0.71), $X = (19, 15, 0.6)$.



(c) PCE surrogate of the covariance (mean histogram intersection is equal to 0.76), $X = (12, 16, 0.14)$

Figure 5.11: Surrogated and simulated CDFs plotted in the three approaches, α is centred.

The KS test allows to rank these three approaches, something that none of the three error metrics could provide, since they all showed that the performance of the three approaches is more or less the same. Considering the fact that the size of the training set was small because of the high computational cost of the original stochastic simulator, the error has been considered acceptable.

5.7.3.2 Sensitivity analysis

In this section we apply the method presented in Chapter 4 to the stochastic city generator. Following the generation of the DOE, differential entropy is then evaluated on each of the 30 random variables $\alpha(X_1^{(i)}, X_2^{(i)}, X_3^{(i)})$, $1 \leq i \leq 30$ (the bin number is 7 in this example). A PCE surrogate model of differential entropy is built, from which Sobol' indices are drawn (Table 5.5). They are compared to Sobol' indices of the mean and the variance computed from the PCE surrogate models of the stochastic process, both based on the $M = 30$ points of the DoE. The respective leave one out error of the mean, variance and entropy metamodel is $0.1, 0.076$ and 0.08. The respective Sobol' indices are drawn in Table 5.5.

Input variables	X_1		X_2		X_3	
PCE Sobol	total	first order	total	first order	total	first order
Mean of α	0.5086	0.4902	0.5098	0.4914	0.0184	0
Variance of α	0.5000	0.4485	0.5367	0.5000	0.0148	0
Entropy of α	0.2083	0.1939	0.6112	0.5969	0.195	0.195

Table 5.5: Total and first order Sobol' indices for the mean, variance and entropy for the exposure example.

Sobol' indices on the mean, variance and entropy do not lead to the same conclusions. The variance of (the mean, variance, and entropy of) α is mostly sensitive to the variance of the building height. Depending on the building height in a given city, rays generated by the antenna can be propagated in a wider or more confined space. For both, the mean and the variance, the variable X_3 (i.e the streets anisotropy) does not seem to impact the variance of the output. Meanwhile, in recent studies, anisotropy has proved to impact the propagation. In fact, the higher the anisotropy, the larger α [77, 43], which is more consistent with the case where the SA is performed on entropy.

The total indices are the summation of the first order Sobol' indices and any interactions. In this study total and first order Sobol' indices are rather close. Thus

second and third order Sobol' indices do not seem to have significant impact on the mean, variance and entropy. In this respect we can deduce that the relationship seems linear between the city parameters and (the mean, variance and entropy) of the PLE.

5.8 Conclusions

In this chapter, the methods developed in the thesis were applied to a full scale problem dealing with the impact the morphological characteristics of a city have onto the exposure of the population to RF-EMF.

The case study was at first described in a simplified way. A non-expert reader can easily follow the drive to perform such a study, the computational obstacles and the use of the surrogate model as well as the sensitivity analysis to overcome them.

The main tools to tackle the case study, i.e. concepts such as human exposure, the path loss exponent, GeoStat and the ray tracing method were introduced. The stochastic simulator is a model combining all the stated concepts to evaluate the human exposure in a city. It takes the city morphological characteristics as an input, and computes the path loss exponent (the reduction of power density from a source to a receiver). The simulator is called stochastic since each input parameters set is not tied to a unique output, but rather to a probability density function (PDF).

The planning of the experiments was an important task. With a limited budget, the challenge was not only to efficiently explore the domain, but also to ensure a sufficient number of repetitions per point for a good grasp of the PDF.

The approach presented in Chapter 3 was successfully applied to build a learner to the stochastic city generator. Using the metamodel, the PDF of the path loss exponent can be evaluated for each city 'type', with a computational cost reduced to a few seconds.

The original question about which of the city morphological characteristics impacts the most the exposure has been addressed using the sensitivity analysis method presented in Chapter 4. Methods from the literature were also implemented. The comparison showed that the proposed approach brought the most accurate ranking. The height of buildings impacted the most the exposure. The proposed method revealed that the variance of the anisotropy considerably impacted the exposure, whereas methods from the literature judged the anisotropy as insignificant.

Chapter 6

Conclusion

The methods presented in this thesis arise from a full scale problem. It is about exploring the human exposure and its possible links to the urban characterization of a city. The principal objective is to identify the most influencing parameters that could serve as levers to control the population levels of EMF exposure in urban cities (building height, street width, anisotropy). To this end, dosimetric tools and statistical methods were practical. However the code used to evaluate the exposure in such cities was expensive to run and inherently stochastic.

Towards this objective, a first method was proposed in this thesis to build a surrogate model to the original model to alleviate the computational burden. The literature was not that diverse when it comes to surrogate modeling stochastic simulators such as the one we had at hand. To this end the method proposed filled a void. The method is based on Karhunen-Loève expansion and two approaches were considered. The eigenvectors of the KL expansion in the domain of interest have been predicted in two different ways: at first a surrogate model of the process covariance operator using polynomial chaos expansions (PCE) has been used. The second approach consists in directly surrogating the eigenvectors. The method is non-parametric: the output is considered as a probability density function, and not reduced to its first moments.

To gather data and apply the mentioned method, a computer experiment was planned. $1,500$ runs of the simulator were performed, split between exploration and repetitions. Three months were needed to gather the necessary data to build the surrogate model. The method involved multiple steps. At each one, options were available and the performance of the surrogate depended on the selected options, on the data set and its size.

To evaluate how sensitive the exposure is to the city characteristics, a method was proposed to evaluate the impact the variability of the city has onto the entropy of the exposure. Evaluating the sensitivity via variations of differential entropy offered

a new perspective to assess uncertainty. Results of the proposed method brought out interesting conclusions about the impact city parameters (building height, street width, anisotropy) have on the exposure.

The two proposed methods from Chapters 3 and 4 are independent from each other. Both chapters are presented in a general context and can be applied to a wide range of industrial problems entailing stochastic simulators. The methods are agnostic to the tools used, and often many possibilities are available. The objective here is to stay as generic as possible and allow the user to adapt the methods to the problem at hand.

Some perspectives are detailed in 3.3.5 concerning the surrogate modeling of the stochastic simulator. For the sensitivity analysis on stochastic simulators, few options can be explored in future works. Instead of choosing the differential entropy as a measure of uncertainty, *f-divergence* functions can be explored, by fixing a usual probability density function (PDF) as a reference (e.g. uniform PDF), and computing the f-divergence between the output PDF and the reference PDF.

This work raised the well-known dilemma of exploration versus repetition. For budgetary reasons, M and N were heuristically chosen. A perspective lies in providing a criteria enabling the update and the optimality of M and N.

To conclude, the objectives achieved in this thesis aim at featuring a particular, yet very compelling type of computational models called stochastic simulators. The urge to study such simulators is pressing, now more than never, mainly due to their adequacy to model complex problems. No doubt that upcoming advances will shed more light onto related subjects.

Bibliography

[1] www.lexnet.fr/pages-sp/search.html.

[2] www.statista.com/statistics/467177/forecast-of-smartphone-users-in-france/.

[3] 28, I. S. C. C. *IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3kHz to 300 GHz*. Institute of Electrical and Electonics Engineers, Incorporated, 1992.

[4] Ankenman, B., Nelson, B., and Staum, J. Stochastic Kriging for simulation metamodeling. *Operations Research 58* (2009), 371–382.

[5] Auder, B., and Iooss, B. Global sensitivity analysis based on entropy. In *Safety, reliability and risk analysis-Proceedings of the ESREL 2008 Conference* (2008), pp. 2107–2115.

[6] Azzi, S., Huang, Y., Sudret, B., and Wiart, J. Surrogate modeling of stochastic functions - application to computational electromagnetic dosimetry. *Int. J. Uncer. Quant. 9*, 4 (2019).

[7] Azzi, S., Sudret, B., and Wiart, J. Sensitivity analysis for stochastic simulators using differential entropy. *Int. J. Uncer. Quant. 10*, 1 (2020).

[8] Berveiller, M., Sudret, B., and Lemaire, M. Stochastic finite elements: a non intrusive approach by regression. *Eur. J. Comput. Mech. 15*, 1-3 (2006), 81–92.

[9] Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics* (1946), 401–406.

[10] Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

[11] Blatman, G., and Sudret, B. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique 336*, 6 (2008), 518–523.

[12] Blatman, G., and Sudret, B. Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *J. Comput. Phys 230* (2011), 2345–2367.

[13] Borgonovo, E. A new uncertainty importance measure. *Reliab. Eng. Sys. Safety 92* (2007), 771–784.

[14] Browne, T., Iooss, B., Le Gratiet, L., Lonchampt, J., and Rémy, E. Stochastic simulators based optimization by Gaussian process metamodels – Application to maintenance investments planning issues. *Quality and Reliability Engineering International 32*, 6 (2016), 2067–2080.

[15] Bursztyn, D., and Steinberg, D. Comparison of designs for computer experiments. *J. Stat. Planning. Infer. 136* (2006), 1103–1119.

[16] Campbell, K., McKay, M. D., and Williams, B. J. Sensitivity analysis when model outputs are functions. *Reliability Engineering & System Safety 91*, 10 (2006), 1468 – 1472.

[17] Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

[18] Chevreuil, M., Lebrun, R., Nouy, A., and Rai, P. A least-squares method for sparse low rank approximation of multivariate functions. *SIAM/ASA J. Uncertainty Quantification 3*, 1 (2015), 897–921.

[19] Chiaramello, E., Parazzini, M., Fiocchi, S., Ravazzani, P., and Wiart, J. Stochastic dosimetry based on low rank tensor approximations for the assessment of children exposure to wlan source. *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology 2*, 2 (June 2018), 131–137.

[20] Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[21] Christ, A., Kainz, W., Hahn, E. G., Honegger, K., Zefferer, M., Neufeld, E., Rascher, W., Janka, R., Bautz, W., Chen, J., et al. The virtual family—development of surface-based anatomical models of two adults and two children for dosimetric simulations. *Physics in Medicine & Biology 55*, 2 (2009), N23.

[22] Collobert, R. Large scale machine learning. Tech. rep., Université de Paris VI, 2004.

[23] Conil, E., Hadjem, A., Gati, A., Wong, M.-F., and Wiart, J. Influence of plane-wave incidence angle on whole body and local exposure at 2100 MHz. *IEEE Transactions on electromagnetic compatibility 53*, 1 (2011), 48–52.

[24] Conil, E., Hadjem, A., Lacroux, F., Wong, M., and Wiart, J. Variability analysis of sar from 20 mhz to 2.4 ghz for different adult and child models using finite-difference time-domain. *Physics in Medicine & Biology 53*, 6 (2008), 1511.

[25] Courtat, T., Decreusefond, L., and Martins, P. Stochastic simulation of urban environments. application to path-loss in wireless systems. *arXiv:1604.00688,* (2016). [Online].

[26] Cover, T. M., and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

[27] Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. Incorporating second-order functional knowledge for better option pricing. In *Advances in neural information processing systems* (2001), pp. 472–478.

[28] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. Least angle regression. *The Annals of statistics 32*, 2 (2004), 407–499.

[29] Endres, D. M., and Schindelin, J. E. A new metric for probability distributions. *IEEE Transactions on Information theory* (2003).

[30] Fisher, R. A. The arrangement of field experiments. In *Breakthroughs in statistics*. Springer, 1992, pp. 82–91.

[31] Frederic, S., Huang, Y., and Wiart, J. *Programme GeoStat : Fiche technique*, 2016.

[32] Ghanem, R., and Spanos, P. *Stochastic Finite Elements: A Spectral Approach*, 2nd ed. Courier Dover Publications, Mineola, 2003.

[33] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

[34] Halton, J. H. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik 2*, 1 (1960), 84–90.

[35] Hastie, T., and Tibshirani, R. *Generalized additive models.* Chapman & Hall, 1990.

[36] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: Data mining, inference and prediction.* Springer, New York, 2001.

[37] Hoeffding, W. A class of statistics with asymptotically normal distributions. *Ann. Math. Stat. 19* (1948), 293–325.

[38] Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[39] Huang, Y., Varsier, N., Niksic, S., Kocan, E., Pejanovic-Djurisic, M., Popovic, M., Koprivica, M., Neskovic, A., Milinkovic, J., Gati, A., Person, C., and Wiart, J. Comparison of average global exposure of population induced by a macro 3G network in different geographical areas in France and Serbia. *Bioelectromagnetics, 37* (2016), 382–390.

[40] Huang, Y., and Wiart, J. Simplified assessment method for population RF exposure induced by a 4G network. *IEEE Journal of Electromagnetics, RF, and Microwaves in Medicine and Biology 1* (2017), 34–40.

[41] ICNIRP. Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). *Health Phys 74*, 4 (1998), 494–522.

[42] IEC. Procedure to measure the Specific Absorption Rate (SAR) in the frequency range of 300 MHz to 3 GHz – Part 1: hand-held mobile wireless communication devices. *International Electrotechnical Commission, committee draft for vote, IEC 62209* (2001).

[43] Ikegami, F., Yoshida, S., Takeuchi, T., and Umehira, M. Propagation factors controlling mean field strength on urban streets. *IEEE Transactions on Antennas and Propagation 32*, 8 (1984), 822–829.

[44] Iooss, B., and Ribatet, M. Global sensitivity analysis of computer models with functional inputs. *Reliab. Eng. Syst. Saf. 94* (2009), 1194–1204.

[45] Iooss, B., Ribatet, M., and Marrel, A. Global sensitivity analysis of stochastic computer models with generalized additive models. *Technometrics, submitted* (2008).

[46] Kersaudy, P., Sudret, B., Varsier, N., Picon, O., and Wiart, J. A new surrogate modeling technique combining Kriging and polynomial chaos expansions – Application to uncertainty analysis in computational dosimetry. *J. Comput. Phys 286* (2015), 103–117.

[47] Kleijnen, J. P. Design and analysis of simulation experiments.

[48] Konakli, K., and Sudret, B. Polynomial meta-models with canonical low-rank approximations: Numerical insights and comparison to sparse polynomial chaos expansions. *J. Comput. Phys. 321* (2016), 1144–1169.

[49] Krzykacz-Hausmann, B. Epistemic sensitivity analysis based on the concept of entropy. *Proceedings of SAMO* (2001), 31–35.

[50] Lakhany, A., and Mausser, H. Estimating the parameters of the generalized lambda distribution. *Algo research quarterly 3*, 3 (2000), 47–58.

[51] Lebrun, R., and Dutfoy, A. A generalization of the nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics 24*, 2 (2009), 172–178.

[52] Liu, H., Chen, W., and Sudjianto, A. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design 128*, 2 (2006), 326–336.

[53] Liu, M., and Staum, J. Stochastic Kriging for efficient nested simulation of expected shortfall. *Journal of Risk 12*, 3 (2010), 3.

[54] Maday, Y., Nguyen, N. C., Patera, A. T., and Pau, S. A general multipurpose interpolation procedure: the magic points. *Communications on Pure & Applied Analysis 8*, 1 (2009), 383.

[55] Marelli, S., and Sudret, B. UQLab: A framework for uncertainty quantification in Matlab. In *Vulnerability, Uncertainty, and Risk (Proc. 2nd Int. Conf. on Vulnerability, Risk Analysis and Management (ICVRAM2014), Liverpool, United Kingdom)* (2014), pp. 2554–2563.

[56] Marrel, A., Iooss, B., Da Veiga, S., and Ribatet, M. Global sensitivity analysis of stochastic computer models with joint metamodels. *Stat. Comput. 22* (2012), 833–847.

[57] Maxwell, J. C. *A treatise on electricity and magnetism*, vol. 1. Clarendon press, 1881.

[58] Mazo, G. An optimal tradeoff between explorations and repetitions in global sensitivity analysis for stochastic computer models. Preprint hal-02113448, 2019.

[59] McKay, M. D., Beckman, R. J., and Conover, W. J. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 2 (1979), 239–245.

[60] Morris, M. Factorial sampling plans for preliminary computational experiments. *Technometrics 33*, 2 (1991), 161–174.

[61] Moutoussamy, V., Nanty, S., and Pauwels, B. Emulators for stochastic simulation codes. *ESAIM: Mathematical Modelling and Numerical Analysis 48* (2015), 116–155.

[62] Nagel, W., and Weiss, V. Crack stit tessellations: characterization of stationary random tessellations stable with respect to iteration. *Advances in applied probability 37*, 4 (2005), 859–883.

[63] Opinion, T., and Social. Electromagnetic fields report, 2010.

[64] Peng, C.-Y., and Wu, C. F. J. On the choice of nugget in Kriging modeling for deterministic computer experiments. *Journal of Computational and Graphical Statistics 23*, 1 (2014), 151–168.

[65] Plumlee, M., and Tuo, R. Building accurate emulators for stochastic simulations via quantile Kriging. *Technometrics 56*, 4 (2014), 466–473.

[66] Pousi, J., Poropudas, J., and Virtanen, K. Game theoretic simulation metamodeling using stochastic Kriging. In *Proceedings of the Winter Simulation Conference* (2010), Winter Simulation Conference, pp. 1456–1467.

[67] Powell, M. J. Radial basis functions for multivariable interpolation: a review. *Algorithms for approximation* (1987).

[68] Pronzato, L., and Müller, W. G. Design of computer experiments: space filling and beyond. *Statistics and Computing 22*, 3 (May 2012), 681–701.

[69] Ramsay, J., and Silverman, B. *Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2006.

[70] Rappaport, T. S., et al. *Wireless communications: principles and practice*, vol. 2. prentice hall PTR New Jersey, 1996.

[71] Recommendation, C., et al. Limitation of exposure of the general public to electromagnetic fields (0 hz to 300 ghz). *Official Journal of the European Communities 199* (1999).

[72] Reich, B. J., Kalendra, E., Storlie, C. B., Bondell, H. D., and Fuentes, M. Variable selection for high dimensional Bayesian density estimation: application to human exposure simulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*, 1 (2012), 47–66.

[73] Rigby, R., and Stasinopoulos, D. Construction of reference centiles using mean and dispersion additive models. *Journal of the Royal Statistical Society: Series D (The Statistician) 49*, 1 (2000), 41–50.

[74] Rioul, O. This is it: A primer on Shannon's entropy and information. *Progress in Mathematical Physics, to appear* (2018).

[75] Rosenblatt, M. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*, 3 (1952), 470–472.

[76] Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. Learning representations by back-propagating errors. *Nature 323* (1986), 533–536.

[77] Sakagami, S., and Kuboi, K. Mobile propagation loss prediction for arbitrary urban environments. *Electronics and Communications in Japan (Part I: Communications) 74*, 10 (1991), 87–99.

[78] Saltelli, A., Tarentola, S., Campolongo, F., and Ratto, M. *Sensitivity analysis in practice – A guide to assessing scientific models*. J. Wiley & Sons, 2004.

[79] Santner, T., Williams, B., and Notz, W. *The Design and Analysis of Computer Experiments*. Springer, New York, 2003.

[80] Scott, D. W. On optimal and data-based histograms. *Biometrika 66*, 3 (1979), 605–610.

[81] Shannon, C. E. A mathematical theory of communication. *Bell system technical journal 27*, 3 (1948), 379–423.

[82] Sobol', I. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling & Comp. Exp. 1* (1993), 407–414.

[83] Sobol', I. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul. 55*, 1-3 (2001), 271–280.

[84] Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 7*, 4 (1967), 784–802.

[85] Sturges, H. A. The choice of a class interval. *Journal of the American Statistical Association 21*, 153 (1926), 65–66.

[86] Su, S. Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics & Data Analysis 51*, 8 (2007), 3983–3998.

[87] Sudret, B. Global sensitivity analysis using polynomial chaos expansions. In *Proc. 5th Int. Conf. on Comp. Stoch. Mech (CSM5), Rhodos, Greece* (2006), P. Spanos and G. Deodatis, Eds.

[88] Sudret, B. *Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods.* Université Blaise Pascal, Clermont-Ferrand, France, 2007. Habilitation à diriger des recherches, 173 pages.

[89] Sudret, B. Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Sys. Safety 93* (2008), 964–979.

[90] Taflove, A., and Hagness, S. C. Computational electromagnetics: the finite-difference time-domain method. *Artech House, Norwood* (2005).

[91] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[92] Tropp, J. A., and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory 53*, 12 (2007), 4655–4666.

[93] Uguen, B., Plouhinec, E., Lostanlen, Y., and Chassay, G. A deterministic ultra wideband channel modeling. In *2002 IEEE Conference on Ultra Wideband Systems and Technologies (IEEE Cat. No.02EX580)* (May 2002), pp. 1–5.

[94] Varsier, N., Plets, D., Corre, Y., Vermeeren, G., Joseph, W., Aerts, S., Martens, L., and Wiart, J. A novel method to assess human population exposure induced by a wireless cellular network. *Bioelectromagnetics 36* (2015), 451–463.

[95] Wiart, J. *Radio-frequency human exposure assessment: from deterministic to stochastic methods*. John Wiley & Sons, 2016.

[96] Xiu, D. *Numerical methods for stochastic computations – A spectral method approach*. Princeton University press, 2010.

[97] Xiu, D., and Karniadakis, G. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput. 24*, 2 (2002), 619–644.

[98] Yu, X., Courtat, T., Martins, P., Decreusefond, L., and Kelif, J. Crack STIT tessellations for city modeling and impact of terrain topology on wireless propagation. In *Proc. 10th Int. Workshop on Wireless Network Measurements and Experimentation (WiNMeE 2014)* (2014).

[99] Zhu, X., and Sudret, B. Replication-based emulation of the response distribution of stochastic simulators using generalized lambda distributions. *Int. J. Uncer. Quant. 10*, 3 (2020).

Publications

Journal

- Azzi, S., Sudret, B. and Wiart, J. Sensitivity analysis for stochastic simulators using differential entropy. *Int. J. Uncer. Quant. 10*, 1 (2020).

- Azzi, S., Huang, Y., Sudret, B. and Wiart, J. *Surrogate modelling of stochastic functions - application to computational electromagnetic dosimetry. Int. J. Uncer. Quant. 9*, 4 (2019).

Conferences

- Azzi, S., Sudret, B. and Wiart, J. Sensitivity analysis on dosimetry simulator. In 33rd General Assembly and Scientific Symposium of the International Union of Radio Science, URSI GASS, Rome (Italy), 29 August–5 September, (2020) Submitted.

- Azzi, S., Sudret, B. and Wiart, J. Sensitivity analysis for stochastic simulators using differential entropy. In SAMO2019, 9th International Conference on Sensitivity Analysis of Model Output, Barcelona (Spain), October 28-30 (2019) (poster).

- Azzi, S., Sudret, B. and Wiart, J. Surrogate modelling of stochastic simulators based on Karhunen-Loève expansion - Application to population RF exposure. In Proc. 3rd Int. Conf. Uncertainty Quantification in Computational Sciences and Engineering (UNCECOMP), Crete Island (Greece), June 24-26 (2019) (oral).

- Azzi, S., Sudret, B. and Wiart, J. Surrogate modelling of stochastic simulators using Karhunen-Loève expansions. In MascotNum workshop, IFPEN, Rueil-Malmaison (France), March 18-20 (2019) (poster).

- Azzi, S., Sudret, B. and Wiart, J. Random processes metamodeling using Karhunen-Loève expansion Application to dosimetry. In UMEMA 2018, 4th Workshop on Uncertainty Modelling for Engineering Computational Sciences Applications, Split (Croatia), December 10-11 (2019) (oral).

- Azzi, S., Huang, Y., Sudret, B. and Wiart, J. Random processes metamodelling applied to dosimetry. In Proc. 2nd URSI Atlantic Radio Science Meeting, Gran Canaria (Spain), May 28- June 1 (2018) (oral).

- Azzi, S., Sudret, B. and Wiart, J. Stochastic metamodeling applied to dosimetry. In MascotNum workshop, Centrale Nantes, Nantes, (France), March 18-20 (2018) (poster).

- Azzi, S., Huang, Y., Sudret, B. and Wiart, J. Surrogate modelling of random functions linked to population RF Exposure. In UMEMA 2017, 3th Workshop on Uncertainty Modelling for Engineering Applications, Turin (Italy), November 23-24 (2017) (oral).

ECOLE DOCTORALE

Titre : Emulateurs de simulateurs stochastiques

Mots clés : Processus stochastiques, Apprentissage statistique, Métamodèle, Exposition aux ondes électromagnétiques

Résumé : Cette thèse propose des outils statistiques pour étudier l'impact qu'a la morphologie d'une ville sur l'exposition des populations induite par un champ électromagnétique provenant d'une station de base. Pour cela l'exposition a été évaluée numériquement en propageant (via des techniques de lancer de rayons) les champs émis dans une antenne dans des villes aléatoires. Ces villes aléatoires ont les mêmes caractéristiques macroscopiques (e.g. hauteur moyenne des immeubles, largeur moyenne des rues et anisotropie) mais sont distinctes les unes des autres. Pour les mêmes caractéristiques de nombreuses villes aléatoires ont été générées et l'exposition induite a été calculée pour chacune. Par conséquent, chaque combinaison de variables correspond à plusieurs valeurs d'exposition. L'exposition est décrite par une distribution statistique non nécessairement gaussienne. Ce comportement stochastique est présent en plusieurs problèmes industriels et souvent les nombreuses simulations menées ont un cout de calcul important.
Les travaux de cette thèse étudient la modélisation de substitution des fonctions aléatoires. Le simulateur stochastic est considéré comme un processus stochastique. On propose une approche non paramétrique basée sur la décomposition de Karhunen-Loève du processus stochastique. La fonction de substitution a l'avantage d'être très peu coûteuse à exécuter et à fournir des prédictions précises.
En effet, l'objective de la thèse consiste à évaluer la sensibilité de l'exposition aux caractéristique morphologiques d'une ville. On propose une approche d'analyse de sensibilité tenant compte de l'aspect stochastique du modèle. L'entropie différentielle du processus stochastique est évaluée et la sensibilité est estimée en calculant les indices de Sobol de l'entropie. La variance de l'entropie est exprimée en fonction de la variabilité de chacune des variables d'entrée.

Title : Surrogate modeling of stochastic simulators

Keywords : Stochastic processes, Statistical learning, Surrogate modeling, Electromagnetic dosimetry

Abstract : This thesis is a contribution to the surrogate modeling and the sensitivity analysis on stochastic simulators. Stochastic simulators are a particular type of computational models, they inherently contain some sources of randomness and are generally computationally prohibitive. To overcome this limitation, this manuscript proposes a method to build a surrogate model for stochastic simulators based on Karhunen-Loève expansion.
This thesis also aims to perform sensitivity analysis on such computational models. This analysis consists on quantifying the influence of the input variables onto the output of the model. In this thesis, the stochastic simulator is represented by a stochastic process, and the sensitivity analysis is then performed on the differential entropy of this process.
The proposed methods are applied to a stochastic simulator assessing the population's exposure to radio frequency waves in a city. Randomness is an intrinsic characteristic of the stochastic city generator. Meaning that, for a set of city parameters (e.g. street width, building height and anisotropy) does not define a unique city. The context of the electromagnetic dosimetry case study is presented, and a surrogate model is built. The sensitivity analysis is then performed using the proposed method.

Institut Polytechnique de Paris
91120 Palaiseau, France