A multivariate non-parametric kernel estimator for global sensitivity analysis

Lamia Djerroud, Tristan Senga Kiessé, Smail Adjabi

A multivariate non-parametric kernel estimator for global sensitivity analysis (*2nd revised version*)

Lamia Djerroud [a*], Tristan Senga Kiessé [b, c] and Smail Adjabi [a]

[a] University of Bejaia, LAMOS Laboratory, Algeria

[b] UMR SAS, INRA, Agrocampus Ouest, Rennes, France

[c] University of Nantes, GeM Laboratory, CNRS UMR 6183,
Chair of civil engineering and eco-construction, Saint-Nazaire, France

Abstract

To estimate how a model output is influenced by the variations of inputs has become an important problematic in reliability and sensitivity analyses. This paper is interested in estimating sensitivity indices useful to quantify the contribution of inputs to the variance of model output. A multivariate mixed kernel estimator is investigated since, until now, discrete and continuous inputs have been separately considered in kernel estimation of sensitivity indices. To illustrate the differences between the influence of mixed, discrete and continuous inputs, analytical expressions of Sobol sensitivity indices are expressed in these three cases for the Ishigami test function. Besides, the performance of mixed kernel estimator is illustrated through simulations in which the Bayesian procedure is applied for bandwidth parameter choice. An application is also realized on a real example. Finally, to use an appropriate kernel estimator according to the type of inputs is found to be influential on the accuracy of sensitivity indice estimates.

*Corresponding author

Key Words: Analysis of variance; Associated kernel; Nonparametric regression; Sensitivity indices.

1 Introduction

Global Sensitivity Analysis (GSA) methods are useful to identify sources of variability/uncertainty in a model and to quantify the influence of inputs on model output [20]. GSA has found many applications in various domains such as civil engineering [1], renewable energy industry [13] and maritime industry [22]. In different applications, the quantitative approaches based on the ANOVA (analysis of variance) decomposition of model output are among the more popular GSA methods. On the basis of ANOVA decomposition, sensitivity indices quantifying the influence of inputs $X_{i,i=1,2,\dots,d} \in \mathbb{T}$ on the output $Y \in \mathbb{R}$ can be calculated as

$$S_i = \frac{\mathrm{Var}\{\mathbb{E}(Y|X_i)\}}{\mathrm{Var}(Y)}, \quad S_{ij} = \frac{\mathrm{Var}\{\mathbb{E}(Y|X_i, X_j)\}}{\mathrm{Var}(Y)} - S_i - S_j, \quad \cdots, \tag{1}$$

where the index S_i measures the main effect of input parameter X_i on output Y and, the index S_{ij} measures the interaction effect between X_i and X_j by excluding their individual effect [20]. Besides, the total sensitivity index (or total effect) introduced by Homma and Saltelli [6] measures the individual effect of one input by including its interactions with all other inputs such that

$$ST_i = S_i + \sum_{j \neq i} S_{ij} + \sum_{j \neq i, k \neq i, j<k} S_{ijk} + \cdots = 1 - \frac{\mathrm{Var}\{\mathbb{E}(Y|X_{-i})\}}{\mathrm{Var}(Y)}, \tag{2}$$

with $\mathrm{Var}\{\mathbb{E}(Y|X_{-i})\}$ being the variance of the expectation of Y conditionally to all variables except X_i.

One of the main objectives of current studies on GSA remains to quickly and efficiently estimate sensitivity indices. To obtain accurate estimations of Sobol sensitivity indices requires a large number of model evaluations [18]. Thus, various statistical tools were investigated in the literature to accurately estimate conditional expectation $\mathbb{E}(Y|\cdot)$ and, consequently, sensitivity indices in Equation (1) (refer to [8] for a review on GSA methods). Among other approaches, non-parametric smoothing methods as the continuous kernel-based estimation [17] and the State-Dependent Pa-

rameter estimation [15] are good choices for estimating $\mathbb{E}(Y|\cdot)$. The former estimation method has been shown to be competing with the latter in term of performance [12]. Then, continuous [12] and discrete [19] kernel approaches have been investigated for estimating the sensitivity indices of continuous and discrete inputs, respectively.

While practically many data sets are mixed, i.e. contain both continuous and discrete inputs, to our knowledge estimating sensitivity indices by a mixed kernel approach has not been yet considered in the literature. Table 1 illustrates the difference that may induce the nature of inputs when quantifying their effect, for the Ishigami test function $Y = f(\mathbf{X}) = \sin(X_1) + a\sin^2(X_2) + bX_3^4\sin(X_1)$ with $\mathbf{X} = (X_1, X_2, X_3) \in \mathbb{T}^3$ [10]. The sensitivity indices values calculated analytically vary depending on the nature (continuous, discrete or mixed) of the random vector $\mathbf{X}$, even if the ranking of the influence of parameters remains the same. Thus, it is worth studying mixed kernel estimation of sensitivity indices in a situation like this one in which discrete and continuous estimators abound. An appropriate mixed kernel estimation must provide an accurate estimation of the sensitivity indices for mixed inputs.

Table 1 about here

This work considers simultaneously continuous and discrete input variables X_i, contrary to previous ones [12, 19]. Two important issues of both discrete and continuous kernel estimations are the kernel and bandwidth choices. About the kernel choice, two associated kernels are considered: a continuous Gaussian kernel and a discrete symmetric triangular kernel [7]. About the bandwidth selection, Bayesian techniques are considered which enable to choose the variance of the Gaussian error in the non-parametric multivariate count regression. The Bayesian formalism is characterized by treating the bandwidth and the variance error as parameters with prior distributions. For the multivariate continuous kernel regression estimation, the Bayesian bandwidth selector is comparable to the cross-validation method and more accurate than bootstrapping method and normal rule reference bandwidth selector [26]. One of the advantage of the proposed Bayesian approach over the cross-validation is its ability to estimate the error density.

An illustration is proposed by calculating analytical expressions of Sobol sensitivity indices of first, second and total order of mixed inputs for the Ishigami function, in comparison with cases

of discrete and continuous inputs. Then, simulations are conducted by using a multivariate non-parametric mixed kernel estimator of the conditional expectation $\mathbb{E}(Y|\cdot)$ for estimating sensitivity indices. Such Nadaraya-Watson (NW) type estimator with different mixed kernel functions is also investigated in Zhang et al. [24]. An application is also proposed on a real example with a model having mixed inputs.

2 Multivariate non-parametric estimator of ANOVA decomposition

This part presents the kernel estimator of ANOVA decomposition of a model $Y = f(X_1, X_2, \ldots, X_d)$ with some asymptotic properties.

Let us consider the realizations $(x_{ij}, y_i)_{j=1,2,\ldots,d}^{i=1,2,\ldots,n}$ of independent and identically distributed (iid) random variables defined on $\mathbb{T}^d \times \mathbb{R}$ such that we have the regression model $m(\cdot) = \mathbb{E}(Y^k|\mathbf{X}^k = \cdot)$. Based on an original idea of Luo et al. [12], the ANOVA decomposition of the model $Y = f(X_1, X_2, \ldots, X_d)$ is given by

$$f(x_1, x_2, ..., x_d) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j}^{d} f_{ij}(x_i, x_j) + ... + f_{12...d}(x_1, x_2, ..., x_d), \tag{3}$$

where each term is defined by

$$f_0 = \mathbb{E}(Y),\ f_i = \mathbb{E}(Y|X_i) - f_0,\ f_{ij} = \mathbb{E}(Y|X_i, X_j) - f_i - f_j - f_0, \ldots$$

The kernel estimators of elementary terms f_0, f_i, ..., are defined by

$$\begin{aligned}
\widehat{f_0} &= \tfrac{1}{n}\textstyle\sum_{l=1}^{n} y_l \\
\widehat{f_i}(x_i, h_{ii}) &= \tfrac{1}{n}\textstyle\sum_{l=1}^{n} \mathbb{K}_{x_i,h_{ii}}(x_{il}) y_l \\
\widehat{f_{ij}}(x_i, x_j, h_{ii}, h_{jj}) &= \tfrac{1}{n}\textstyle\sum_{l=1}^{n} \mathbb{K}_{\mathbf{x};\mathbf{H}}(x_{il}, x_{jl}) y_l \\
\widehat{f_{ijk}}(x_i, x_j, x_k; h_{ii}, h_{jj}, h_{kk}) &= \tfrac{1}{n}\textstyle\sum_{l=1}^{n} \mathbb{K}_{\mathbf{x};\mathbf{H}}(x_{il}, x_{jl}, x_{kl}) y_l
\end{aligned}$$

where the functions $\mathbb{K}_{\mathbf{x};\mathbf{H}}$ are given by

$$\begin{aligned}
\mathbb{K}_{x_i,h_{ii}}(x_{il}) &= K_{x_i,h_{ii}}(x_{il}) - 1 \\
\mathbb{K}_{\mathbf{x};\mathbf{H}}(x_{il},x_{jl}) &= K_{\mathbf{x};\mathbf{H}}(x_{il},x_{jl}) - K^{[i]}_{x_i,h_{ii}}(x_{il}) - K^{[i]}_{x_j,h_{jj}}(x_{jl}) - 1, \\
\mathbb{K}_{\mathbf{x};\mathbf{H}}(x_{il},x_{jl},x_{kl}) &= K_{\mathbf{x};\mathbf{H}}(x_{il},x_{jl},x_{kl}) - K^{[i]}_{x_i,h_{ii}}(x_{il}) - K^{[j]}_{x_j,h_{jj}}(x_{jl}) - K^{[k]}_{x_k,h_{kk}}(x_{kl}) - 1,
\end{aligned}$$

with

- $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{T}^d$ being a target vector;
- $\mathbf{H} = \mathbf{Diag}(h_{11}, \ldots, h_{dd})$ being a bandwidth matrix with $h_{jj} > 0$ such as $\mathbf{H} \equiv \mathbf{H}_n$ tend to the null matrix $\mathbf{0_d}$ as $n \to \infty$;
- $K_{\mathbf{x},\mathbf{H}}(\cdot)$ being a multivariate associated kernel defined as a product of univariate associated kernel $K^{[j]}_{x_j,h_{jj}}$ with its corresponding random variable $\mathcal{K}^{[j]}_{x_j,h_{jj}}$, i.e $K^{[j]}_{x_j,h_{jj}}(y) = \Pr(\mathcal{K}^{[j]}_{x_j,h_{jj}} = y)$, on support $\mathbb{S}_{x_j,h_{jj}}$ such that

$$x_j \in \mathbb{S}_{x_j,h_{jj}} \;\; (A1), \;\; \lim_{h_{jj}\to 0} \mathbb{E}(\mathcal{K}^{[j]}_{x_j,h_{jj}}) = x_j \;\; (A2), \;\; \lim_{h_{jj}\to 0} \mathrm{Var}(\mathcal{K}^{[j]}_{x_j,h_{jj}}) = 0 \;\; (A3).$$

Then, the multivariate kernel $K_{\mathbf{x},\mathbf{H}}$ associated with the random variable $\mathcal{K}_{\mathbf{x},\mathbf{H}}$ of support $\mathbb{S}_{\mathbf{x},\mathbf{H}} = \times_{j=1}^{d} \mathbb{S}_{x_j;h_{jj}}$ is a probability mass function (pmf) [21] satisfying

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \;\; \mathbb{E}(\mathcal{K}_{\mathbf{x},\mathbf{H}}) = \mathbf{x} + \mathbf{U}(\mathbf{x},\mathbf{H}), \;\; \mathbb{C}\mathrm{ov}(\mathcal{K}_{\mathbf{x},\mathbf{H}}) = \mathbf{B}(\mathbf{x},\mathbf{H}),$$

where $\mathbf{U}(\mathbf{x},\mathbf{H}) = (u_1(\mathbf{x},\mathbf{H}), \ldots, u_d(\mathbf{x},\mathbf{H}))^\top$ and $\mathbf{B}(\mathbf{x},\mathbf{H}) = (b_{ij}(\mathbf{x},\mathbf{H}))_{i,j=1,\ldots,d}$ tend to null vector $\mathbf{0}$ and null matrix $\mathbf{0_d}$ as $\mathbf{H} \to \mathbf{0_d}$, respectively.

Basic asymptotic properties of estimator $\widehat{f_{\mathbf{I}}}$ of $f_{\mathbf{I}}$ are established. For instance, the bias of $\widehat{f_{\mathbf{I}=\{i\}}}$ is obtained as

$$\mathrm{Bias}\{\widehat{f_i}(x_i;h_{ii})\} = u_i(x_i,h_{ii})(mg)^{(1)}(x_i) + \frac{1}{2}\mathrm{Var}(\mathcal{K}_{x_i,h_{ii}})(mg)^{(2)}(x_i) + o(h_{ii}^2),$$

with $(mg)^{(k)}$ being the finite difference of k-order of product function mg and $g(x_1, x_2, \cdots, x_d)$ the joint probability distribution of inputs $\mathbf{X}^i$. Moreover, the variance of $\widehat{f_i}$ is given by

$$\mathrm{Var}\{\widehat{f_i}(x_i;h_{ii})\} = \frac{1}{n}m^2(x_i)g(x_i)\{g(x_i)-1\} + \frac{1}{n}\mathrm{Var}\{m(\mathbf{X}^i)\} + o(h_{ii}).$$

Thus, under assumptions (A2)-(A3), the pointwise consistency of $\widehat{f_i}$ is deduced by showing that the mean squared error (MSE) tends to 0 as $h_{ii} \to 0$ and $n \to \infty$ since

$$\text{MSE}(x_i) = \text{Bias}^2\{\widehat{f_i}(x_i; h_{ii})\} + \text{Var}\{\widehat{f_i}(x_i; h_{ii})\}.$$

Hereafter, a mathematical result on the almost sure (a.s.) consistency of estimator $\widehat{f_i}$ is formulated. We assume the continuity of the function $f_i : \mathbb{T} \mapsto \mathbb{R}$ at x_i in the sense that

$$\forall\, \epsilon,\ \exists\, \eta > 0 : \forall\, t_i \in (x_i - \eta; x_i + \eta) \cap \mathbb{T} \Rightarrow |f(t_i) - f(x_i)| < \epsilon.$$

Note that, considering a discrete support $\mathbb{T}$, the discrete neighborhood $(x_i - \eta; x_i + \eta) \cap \mathbb{T}$ of x_i may be reduced to the point $\{x_i\}$, for $\eta > 0$.

Proposition 1 For any fixed $x_i \in \mathbb{T}$ and $h_{ii} > 0$, under assumptions (A2) and (A3), the non-parametric kernel estimator $\widehat{f_i}$ of f_i satisfies:

$$\widehat{f_i}(x_i; h_{ii}) \xrightarrow{a.s.} f_i(x_i) \text{ as } \ n \to \infty \text{ and } \ h_{ii} \to 0.$$

The details of calculations are postponed to the Appendix.

We then get the estimated terms

$$\widehat{\mathbb{V}}(Y) = \frac{1}{n}\sum_{l=1}^{n} y_l^2 - \widehat{f_0}^2,\ \widehat{\mathbb{V}_i} = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_i}(x_i; h_{ii})\}^2,\ \widehat{\mathbb{V}_{ij}} = \mathbb{E}_{\mathbf{X}^k}\{\widehat{f_{ij}}(x_i, x_j; h_{ii}, h_{jj})\}^2, \ldots$$

coming from the decomposition of the total variance of Y given by

$$\mathbb{V}(Y) = \sum_{i=1}^{k} \mathbb{V}_i + \sum_{i<j} \mathbb{V}_{ij} + \ldots + \mathbb{V}_{12\ldots k}, \tag{4}$$

where each variance term is defined such as

$$\mathbb{V}_i = \text{Var}\{\mathbb{E}(Y|X_i)\}, \quad \mathbb{V}_{ij} = \text{Var}\{\mathbb{E}(Y|X_i, X_j)\} - \mathbb{V}_i - \mathbb{V}_j, \ldots$$

That leads to the estimated sensitivity indices

$$\widehat{S_i} = \frac{\widehat{\mathbb{V}_i}}{\widehat{\mathbb{V}}(Y)},\ S_{ij} = \frac{\widehat{\mathbb{V}_{ij}}}{\widehat{\mathbb{V}}(Y)}, \quad \cdots, \widehat{ST_i} = 1 - \frac{\widehat{\mathbb{V}_{-i}}}{\widehat{\mathbb{V}}(Y)},$$

such that $\widehat{S_{\mathbf{I}}}$ goes to $S_{\mathbf{I}}$ as n tends to ∞ (see [12] for the continuous case and [19] for the discrete case).

3 Bandwidth choice

In this section, we consider the multivariate estimator $\widehat{m}_n^d$ of the regression function m such that

$$\widehat{m}_n^d(\mathbf{x}, \mathbf{H}) = \sum_{i=1}^{n} \frac{y_i K_{\mathbf{x},\mathbf{H}}(\mathbf{x}^i)}{\sum_{l=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{x}^l)}, \tag{5}$$

with $K_{\mathbf{x},\mathbf{H}}$ being a multivariate associated kernel, $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{T}^d$ a fixed target and $\mathbf{H} = \mathbf{Diag}(h_{11}, \ldots, h_{dd})$ a bandwidth matrix with $h_{jj} > 0$. The Bayesian approach for deriving the selectors of bandwidth matrices is presented in the context of multivariate kernel regression estimation.

3.1 Bayesian bandwidth selection

We propose a bayesian sampling approach for the bandwidth estimation for the NW estimator involving mixed types of regressors. For such an approach in the non-parametric regression, one can refer to [25] with continuous regressors and [24] with mixed types continuous and categorical data.

We consider the multivariate non-parametric regression model given by $y_i = m(\mathbf{x_i}) + \epsilon_i$, with ϵ_i, for $i = 1, 2, \ldots, n$ assumed to be iid such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_m^2$. The previous regression model can be expressed as

$$y_i - m(\mathbf{x_i}) \sim N(0, \sigma_m^2),$$

since ϵ_i follows the Gaussian distribution $N(0, \sigma_m^2)$ with σ_m^2 an unknown parameter. Let us estimate $(\mathbf{H}, \sigma_m^2)$ the unknown parameters in this model. The estimator of the likelihood of the data $(y_i)_{i=1}^n$ knowing the parameters $\mathbf{H}$ and σ_m^2 is given by

$$LCV(y_1, \ldots, y_n; \mathbf{H}, \sigma_m^2) = (2\pi\sigma_m^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma_m^2} \sum_{i=1}^{n} \Big\{ y_i - \widehat{m}_{n,-k}^d(\mathbf{x}; \mathbf{H}) \Big\}^2 \Big],$$

where $\widehat{m}_{n,-k}^d(\mathbf{x}; \mathbf{H})$ is the estimator of the regression function calculated from all observations except $\mathbf{x}^k$.

Priors. For the dispersion parameter σ_m^2, the prior generally adopted in the Bayesian framework is the inverse gamma distribution, denoted by $IG(\alpha, \beta)$, whose distribution is given by

$$\pi(\sigma_m^2) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}\Big(\frac{1}{\sigma_m^2}\Big)^{\alpha+1} \exp\Big(-\frac{\beta}{\sigma_m^2}\Big), \quad \sigma_m^2 \in [0, \infty[,$$

with α and β being the hyper parameters. The prior of $\mathbf{H}$ is also a inverted Gamma with scale parameters α_1 and β_1 denoted by $IG(\alpha_1, \beta_1)$.

An approximate posterior. The posterior distribution of the parameters $(\mathbf{H}, \sigma_m^2)$ given the data is

$$\pi(\mathbf{H}, \sigma_m^2|y) \propto \prod_{i=1}^{d} \pi(h_{jj})\Big(\frac{1}{\sigma_m^2}\Big)^{\frac{n+2\alpha}{2}+1} \exp\Big[-\frac{1}{2\sigma_m^2}\Big\{\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(\mathbf{x}_i))^2 + 2\beta\Big\}\Big].$$

Note that $\pi(\mathbf{H}, \sigma_m^2|y) = \pi(\mathbf{h}|y)\pi(\sigma_m^2|\mathbf{H}, y)$, where $\pi(\sigma_m^2|\mathbf{H}, y)$ is an inverted Gamma density

$$\sigma_m^2 \sim IG\left(\frac{n+2\alpha}{2}, \frac{1}{2}\sum_{i=1}^{n}\Big\{y_i - \widehat{m}_{n,-k}^d(\mathbf{x}; \mathbf{H})\Big\}^2 + \beta\right).$$

The law of $\mathbf{H}$ given the data is as follows:

$$\pi(\mathbf{H}|y_1, ..., y_n) \propto \prod_{j=1}^{d} \pi(h_{jj})\Big[\frac{1}{2}\sum_{i=1}^{n}\Big\{y_i - \widehat{m}_{n,-k}^d(\mathbf{x}; \mathbf{H})\Big\}^2 + \beta\Big]^{-\frac{n+2\alpha}{2}}.$$

3.1.1 MCMC method

We use the Metropolis-Hastings algorithm proposed by [14] and generalized by [4]. This algorithm generates a Markov chain $\mathbf{H}^{(i)}$, $i \in \{1, ..., T\}$, using the proposal distribution $q(.|\mathbf{H}^T)$ and an arbitrary initial value $\mathbf{H}^{(0)}$. After a number of iterations T, sufficiently large, the Markov chain converges to the interest density $\pi(\mathbf{H}|y_1, ..., y_n)$. A random-walk Metropolis-Hastings algorithm and Gibbs sampling procedure are given as follows:

- *Step* 1. For t=0, initialize the first vector $\mathbf{H}^{(0)}$;
- *Step* 2. For $t \in \{1...T\}$
  (a) generate $\tilde{\mathbf{H}}$ following $q(.|\mathbf{H}^{(t-1)})$

(b) calculate $\rho(\mathbf{H}^{(t)}, \tilde{\mathbf{H}}) = \min\left\{1, \frac{\pi(\tilde{\mathbf{H}})}{\pi(\mathbf{H}^{(t)})} \frac{q(\mathbf{H}^{(t)}|\tilde{\mathbf{H}})}{q(\tilde{\mathbf{H}}|\mathbf{H}^{(t)})}\right\}$

(c) take

$$\mathbf{H}^{(t+1)} = \begin{cases} \tilde{\mathbf{H}}, & \textit{with the probability} \quad \rho(\mathbf{H}^{(t)}, \tilde{\mathbf{H}}) \quad \textit{if} \quad u < \rho(\mathbf{H}^{(t)}, \tilde{\mathbf{H}}) \\ \\ \mathbf{H}^{(t)}, & \textit{with the probability} \quad 1 - \rho(\mathbf{H}^{(t)}, \tilde{\mathbf{H}}) \quad \textit{else} \end{cases}$$

(d) generate $\{\sigma^2\}^{(t)}$ from $IG\left(\frac{n+2\alpha}{2}, \frac{1}{2}\sum_{i=1}^{n}\{y_i - \hat{m}_{-i}(\mathbf{x}_i)\}^2 + \beta\right)$

- *Step* 3. t=t+1 and returns to step 2;
- *Step* 4. calculate the Bayes estimator $\hat{I}_H = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}^{(t)}$.

4 Ishigami test function analysis

In this section, analytical expressions of Sobol sensitivity indices of first, second and total order are expressed for a test function with mixed input variables, and compared to discrete and continuous cases. Moreover, we propose to evaluate the application of multivariate mixed (discrete and continuous) kernel estimation procedures through simulations.

Consider a mixed random vector with one discrete random variable $x_1 \in \mathbb{T}$ and two continuous random variables $x_2, x_3 \in \mathbb{T}$, the terms of ANOVA decomposition in Equation (3) are calculated as follows:

$$\begin{aligned} f_0 &= \sum_{x_1\in\mathbb{T}} \int_{x_2\in\mathbb{T}} \int_{x_3\in\mathbb{T}} f(x_1, x_2, x_3) \prod_{i=1}^{3} \Pr(X_i = x_i) dx_2 dx_3 \\ f_1(x_1) &= \int_{x_2\in\mathbb{T}} \int_{x_3\in\mathbb{T}} f(x_1, x_2, x_3) \prod_{i=2}^{3} \Pr(X_i = x_i) dx_2 dx_3 - f_0 \\ f_2(x_2) &= \sum_{x_1\in\mathbb{T}} \int_{x_3\in\mathbb{T}} f(x_1, x_2, x_3) \Pr(X_1 = x_1) \Pr(X_3 = x_3) dx_3 - f_0 \\ f_{12}(x_1, x_2) &= \int_{x_3\in\mathbb{T}} f(x_1, x_2, x_3) \Pr(X_3 = x_3) dx_3 - f_0 - f_1 - f_2 \\ f_{13}(x_1, x_3) &= \int_{x_2\in\mathbb{T}} f(x_1, x_2, x_3) \Pr(X_2 = x_2) dx_2 - f_0 - f_1 - f_3 \\ f_{123}(x_1, x_2, x_3) &= f(x_1, x_2, x_3) - f_0 - f_1 - f_2 - f_3 - f_{12} - f_{13} - f_{23}, \end{aligned} \tag{6}$$

where $\Pr(X_i = x_i)$ is the (discrete or continuous) uniform distribution. We omit the terms $f_3(x_3)$ and $f_{23}(x_2, x_3)$ which can easily be deduced from the expressions of $f_2(x_2)$ and $f_{13}(x_1, x_3)$, respectively.

Then, variance terms in the decomposition give

$$\begin{aligned}
\mathbb{V}(Y) &= \mathrm{Var}\{f(X)\} = \textstyle\sum_{x_1\in\mathbb{T}} \int_{x_2\in\mathbb{T}} \int_{x_3\in\mathbb{T}} \{f(x_1,x_2,x_3)\}^2 \prod_{i=1}^{3} \Pr(X_i = x_i) dx_2 dx_3 - f_0^2 \\
\mathbb{V}_1 &= \mathrm{Var}\{f_1(X_1)\} = \textstyle\sum_{x_1\in\mathbb{T}} \{f_1(x_1)\}^2 \Pr(X_1 = x_1) \\
\mathbb{V}_2 &= \mathrm{Var}\{f_2(X_2)\} = \textstyle\int_{x_2\in\mathbb{T}} \{f_2(x_2)\}^2 \Pr(X_2 = x_2) dx_2 \\
\mathbb{V}_{12} &= \mathrm{Var}\{f_{12}(X_1,X_2)\} = \textstyle\sum_{x_1\in\mathbb{T}} \int_{x_2\in\mathbb{T}} \{f_{12}(x_1,x_2)\}^2 \Pr(X_1 = x_1) \Pr(X_2 = x_2) dx_2 \\
&\vdots
\end{aligned} \tag{7}$$

Herein the Ishigami test function is used: the parameter x_1 is assumed to be discrete on support $\mathbb{T} = \{-3, -2, -1, 0, 1, 2, 3\}$ while the two other parameters x_2 and x_3 are assumed to be continuous on a compact support $\mathbb{T} = [-\pi, \pi]$. The values of the constants a and b are 5 and 0.1, respectively.

4.1 Analytical expression of Sobol indices

4.1.1 First and second order Sobol indices

The ANOVA decomposition of y is

$$y = f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1,x_2) + f_{13}(x_1,x_3) + f_{23}(x_2,x_3) + f_{123}(x_1,x_2,x_3).$$

According to the expressions in Equation (6), we get the terms in the decompostion of f as

$$f_0 = 2.5,\ f_1(x_1) = \sin(x_1)\Big(1+\frac{\pi^4}{50}\Big),\ f_2(x_2) = 5\,\sin^2(x_2) - 2.5, f_{13}(x_1,x_3) = 0.1\,x_3^4\,\sin(x_1) - \frac{\pi^4}{50}\,\sin(x_1)$$

and $f_3(x_3) = f_{12}(x_1,x_2) = f_{23}(x_2,x_3) = f_{123}(x_1,x_2,x_3) = 0$.

Then, expressions in Equation (7) result in the terms of the variance decomposition given by

$$\begin{aligned}
\mathrm{V} &= \frac{25}{8} + \frac{2}{7}\left\{\left(1+\frac{\pi^4}{50}\right)^2 + \frac{4\pi^8}{5625}\right\}\sum_{x_1=1}^{3}\sin^2(x_1) \\
\mathrm{V}_1 &= \frac{2}{7}\left(1+\frac{\pi^4}{50}\right)^2\sum_{x_1=1}^{3}\sin^2(x_1) \\
\mathrm{V}_2 &= \frac{25}{8} \\
\mathrm{V}_{13} &= \frac{2}{7}\left(\frac{4\pi^8}{5625}\right)\sum_{x_1=1}^{3}\sin^2(x_1)
\end{aligned}$$

and $V_3 = V_{12} = V_{23} = V_{123} = 0$. Ultimately, analytical values of sensitivity indices expressed in Equation (1) are given by

$$S_1 = 0.3867516,\ S_2 = 0.3130144,\ S_{13} = 0.3002341,\ S_3 = S_{12} = S_{23} = S_{123} = 0.$$

Table 2 summarizes the terms of the decomposition of the model f and its total variance for continuous, discrete and mixed input variables in the Ishigima test function. Looking at the analytical values of sensitivity indices, the important individual effect of X_1 on Y is pointed out in discrete, mixed and continuous cases, with the greatest value of S_1 in the discrete case. The effect of the interaction between X_1 and X_3 is also pointed with the most important value in the discrete case, while the value of S_1 is greater in continuous and mixed cases than in the discrete case.

Table 2 about here

4.1.2 Total order Sobol indices

This paragraph is concerned with the total sensitivity index in Equation (2). For the input variable X_1, we get

$$ST_1 = S_1 + S_{12} + S_{13} + S_{123} = 1 - \frac{\mathrm{Var}\{\mathbb{E}(Y|X_{-1})\}}{\mathrm{Var}(Y)} = 0.686,$$

where the variance of the conditional expectation at the numerator is given by

$$\mathrm{Var}\{\mathbb{E}(Y|X_{-1})\} = \int\int f^2(x_2, x_3) \prod_{i=2}^{3} \Pr(X_i = x_i) dx_2 dx_3 = \frac{3a^2}{8} - a\ f_0 + f_0^2 = \frac{25}{8},$$

with

$$f(x_2, x_3) \ = \ \sum_{x_1 \in \mathbb{T}} f(x_1, x_2, x_3) \Pr(X_1 = x_1) - f_0 = a\ \sin^2(x_2) - f_0.$$

Similarly, for the input variable X_2, we get

$$ST_2 = S_2 + S_{23} = 1 - \frac{\mathrm{Var}\{\mathbb{E}(Y|X_{-2})\}}{\mathrm{Var}(Y)} = 0.313,$$

with

$$\begin{aligned} \mathrm{Var}\{\mathbb{E}(Y|X_{-2})\} &= \sum_{x_1 \in \mathbb{T}} \int f^2(x_1, x_3) \Pr(X_1 = x_1) \Pr(X_3 = x_3) dx_1 dx_3 \\ &= \frac{2}{7}\Big(1 + \frac{\pi^4}{25} + \frac{\pi^8}{900}\Big) \sum_{x_1 \in \mathbb{T}} \sin^2(x_1). \end{aligned}$$

And, for the input variable X_3, we get $ST_3 = 0.300$ with

$$\begin{aligned}\mathrm{Var}\{\mathbb{E}(Y|X_{-3})\} &= \frac{2}{7}\left(1+\frac{b\pi^4}{5}\right)^2\sum_{x_1=1}^{3}\sin^2(x_1)+\frac{4}{7}\left(\frac{5}{2}-f_0\right)\left(1+\frac{b\pi^4}{5}\right)\sum_{x_1=1}^{3}\sin(x_1)\\ &\quad+\frac{3a^2}{8}-af_0+f_0^2\\ &= \frac{2}{7}\left(1+\frac{\pi^4}{50}\right)^2\sum_{x_1=1}^{3}\sin^2(x_1)+\frac{25}{8}.\end{aligned}$$

Table 3 compares the total sensitivity index and the variance of the conditional expectations for continuous, discrete and mixed input variables in the Ishigima test function. As a logical consequence of the results in Table 2, the input parameter X_1 is found to be the most influential (including interaction with other parameters) on Y. Besides, the value ST_1 is greatest for the discrete case in comparison with the two other cases. In addition, total influences of X_2 and X_3, measured by ST_2 and ST_3, correspond to the main effect of X_2 and to the interaction effect between X_1 and X_3 on Y, respectively.

Table 3 about here

4.2 Simulation results

In this section we investigate the performance of the mixed kernel regression estimator for Sobol indices. The average of first order indice estimates are calculated as

$$\overline{\hat{S}_i} = \sum_{l=1}^{N}(1/N)\hat{S}_i^{(l)}, \quad i = 1,2,3.$$

The mixed regression estimator applied used two kernels: a continuous gaussian kernel and a discrete symmetric triangular kernel having its pmf given by

$$T_{p;x,h}(y) = \frac{(p+1)^h - |y-x|^h}{(2p+1)(p+1)^h - 2\sum_{k=0}^{p}k^h}, \forall y \in \mathbb{S}_{p;x} = \{x, x\pm 1, ..., x\pm p\}, \quad p \in \mathbb{N},$$

with $x \in \mathbb{T}$ being a fixed point and $h > 0$ the bandwidth parameter [7]. The fixed value $p = 1$ is considered since the global squarred error of estimator using discrete symmetric triangular kernels

was shown to increase with respect to $p \in \mathbb{N}$ for a fixed bandwidth $h > 0$. A comparison is realized with the continuous Gaussian kernel estimator of sensitivity indices for mixed inputs. For the bandwidth selection, the Bayesian approach is considered. Note that the boundary bias effect is not treated in this study but has been treated in [19].

The simulation study is carried out for sample sizes $n = \{100, 250, 500, 1000\}$ and the model error $N(0, 0.2)$. We use the MCMC technique for the Bayes estimation. The number of total iterations $N = 1000$ are run and the first $N_0 = 500$ iterations are set as the burn-in period, with the hyper-parameters of the priors dispersion of σ^2 and $\mathbf{H}$ being chosen as $\alpha = \alpha_1 = 1$ and $\beta = \beta_1 = 0.05$ [3]. All computations were done by using the R software. We employ the batch-mean standard error and the simulation inefficiency factor (SIF) to check the convergence performance of the sampling algorithm [23][27]. Both the batch-mean standard error and SIF indicate that all the simulated chains converge very well. Table 4 presents the estimated σ^2 and bandwidth parameters.

Table 4 about here

In order to evaluate the performance of studied estimators, the mean absolute error (MAE) is calculated over N_{sim} replications such that

$$\overline{MAE}(S_i) = (1/N_{sim}) \sum_{l=1}^{N_{sim}} |S_i - S_i^{(l)}|.$$

Results in Tables 5 and 7 give the estimated Sobol indices for numbers of simulations $N_{sim} = 100$ and 200. Estimations of sensitivity indices by mixed and continuous kernel approaches adequately reflect influences of mixed inputs, according to analytical values. The parameter X_1 is the most influential by taking into account the interaction with other variables. For the parameter X_2, we get $ST_2 \approx S_2$, since X_2 has a quasi null interaction with other parameter. And, for parameter X_3, we get $ST_3 \approx S_{13}$. However, the main effect of X_2 and the interaction between X_1 and X_3 are underestimated.

Looking now at $\overline{MAE}$ values calculated over $N_{sim} = 100$ and 200 simulations (Table 6), the mixed kernel estimator globally outperformed the continuous kernel estimator as the sample size n is increased from 100 to 1000, except for estimating the interaction between X_2 and X_3. We also

note that increasing the number of simulations from 100 to 200 confirms and stabilizes the results. The same tendency is observed in Table 7 which presents the average values of the total sensitivity indices calculated by using estimators with mixed and continuous kernels. We omit to present the $\overline{MAE}$ values for the estimated total sensitivity indices in Table 7, but the performance of estimators with mixed and continuous kernels are generally the same that for the sensitivity indices of first and second orders.

We have also used the cross-validation procedure to select the optimal bandwidth matrix [26]. Simulations have pointed out that the kernel estimation of sensitivity indices is better, in the sense of average $\overline{MAE}$, by using the Bayesian approach rather than the cross-validation procedure which does not always converge in many situations. However the computational time of the cross-validation procedure is smaller.

Tables 5, 6 and 7 about here

5 Application on a real case study

This real example serves just here to illustrate the sensitivity of the model output to the nature of input parameters. We apply the multivariate kernel estimator for simulating the height of flooding of a river, compared to the height of a dyke that protects a dwelling or an industrial site. We consider an academic model used for learning purposes to simulate the occurrence of the flooding when the river height exceeded the one of the dyke [2, 9]. The model is based on a crude simplification of the 1D hydro-dynamical equations of Saint Venant under the assumptions of uniform and constant flow rate and large rectangular sections. Input parameters of this model were originally assumed to be continuous [8], we propose here discretised input parameters in order to study how the influence of inputs on model output will be affected. The model consists of an equation that involves the characteristics of the river section upstream of the industrial site:

$$S = Z_v + H - H_d - C_b, \tag{8}$$

with S being the maximal annual overflow (in meters), H being the maximal annual height of the river (in meters) expressed by

$$H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6} ;$$

the other input parameters of the flood model are presented in Table 8 with their probability distribution.

Table 8 about here

Another model output was also considered in [8]: the associated cost (in million euros) of the dyke,

$$C_p = \mathbf{1}_{\{S>0\}} + [0.2 + 0.8(1 - \exp^{\frac{-1000}{S^4}})\mathbf{1}_{\{S\leq 0\}}] + \frac{1}{20}(H_d \mathbf{1}_{\{H_d>8\}} + 8\,\mathbf{1}_{\{H_d\leq 8\}})$$

with $\mathbf{1}_A(x)$ the indicator function which is equal to 1 for $x \in A$ and 0 otherwise. In this equation, the first term represents the cost due to a flooding ($S > 0$) which is 1 million euros, the second term corresponds to the cost of the dyke maintenance ($S \leq 0$) and the third term is the investment cost related to the construction of the dyke. The latter cost is constant for a height of dyke less than 8 m and is growing proportionally with respect to the dyke height otherwise.

Table 9 about here

Let us illustrate the estimation of Sobol indices of $d = 5$ random inputs Z_v, K_s, Q, H_d, C_b assumed to be independent and involved in the model output C_p. For a sample size $n = 1000$, we use the multivariate mixed kernel regression with the Bayesian bandwidth selector. Table 9 presents the results of the mixed case treated in this study and the continuous case in [8]. The influence of discrete parameters Z_v, H_d and continuous parameters C_b, Q and K_s on the variance of C_p changes in comparison with the case where all inputs are assumed to be continuous. Nevertheless, the ranking of influential parameters globally remains the same as in the continuous case. The requirement of an adapted estimation method is thus pointed out to evaluate sensitivity indices. But, a more deeper investigation would be realized to better understand and evaluate the practical signification and consequences of the parameters discretization, in particular for physical parameters.

6 Concluding remarks

In this work, the non-parametric mixed kernel method is applied to estimate sensitivity indices calculated from the ANOVA decomposition. For discrete, continuous and mixed cases, the analytical calculation of Sobol index for the Ishigami test function leads logically to the same conclusion regarding the effect of input parameters and their interaction. The difference between the three cases considered in this study essentially comes from the approximation of sensitivity index values. Thus, an appropriate kernel estimator depending on the type (discrete, continuous or mixed) of data is useful for an accurate estimation of sensitivity indices. We have investigated the Bayesian MCMC for bandwidth matrix selection as a competing alternative to the cross-validation method in the sense of the $\overline{MAE}$. Some interesting perspectives would consist of investigating more deeply the effects of kernel choice since nowadays a large choice of (symmetric and asymmetric) discrete and continuous kernels is available in the literature according to the type of support $\mathbb{T}$ (compact, semi-compact,...). Finally, multivariate theoretical properties of the kernel estimator would be more properly studied with the curse of dimensionality.

Acknowledgements

The research and education chair of civil engineering and eco-construction is funded by the Chamber of Trade and Industry of Nantes and Saint-Nazaire cities, the CARENE (urban agglomeration of Saint-Nazaire), Charier, Architectes Ingénieurs Associés, Vinci construction, the Regional Federation of Buildings, and the Regional Federation of Public Works. The authors wish to thank these partners for their patronage. We are also grateful to the Associate Editor and two referees for their careful reading and comments which have greatly contributed to improve the paper.

Appendix

For the bias of $\widehat{f_i}$ we have successively:

$$\begin{aligned}
\text{Bias}\{\widehat{f_i}(x_i;h_{ii})\} &= \mathbb{E}\{\widehat{f_i}(x_i;h_{ii})\} - f_i(x_i) \\
&= \mathbb{E}\{\mathbb{K}_{x_i,h_{ii}}(X_{i1})Y_1\} - f_i(x_i) \\
&= \sum_{z_i\in\mathbb{T}} m(z_i)g(z_i)K_{x_i;h_{ii}}(z_i) - \sum_{z_i\in\mathbb{T}} m(z_i)g(z_i) - f_i(x_i) \\
&= \mathbb{E}\{(mg)(\mathcal{K}_{x_i,h_{ii}})\} - \mathbb{E}\{m(\mathbf{X}^i)\} - \{\mathbb{E}(\mathbf{Y}^i|\mathbf{X}^i = x_i) - \mathbb{E}(Y)\}. \qquad (9)
\end{aligned}$$

Then, using in Equation (9) the following Taylor expansion

$$\begin{aligned}
\mathbb{E}\{(mg)(\mathcal{K}_{x_i,h_{ii}})\} &= \mathbb{E}\Big[(mg)\{\mathbb{E}(\mathcal{K}_{x_i,h_{ii}})\} + \{\mathcal{K}_{x_i,h_{ii}} - \mathbb{E}(\mathcal{K}_{x_i,h_{ii}})\}(mg)^{(1)}\{\mathbb{E}(\mathcal{K}_{x_i,h_{ii}})\} \\
&\quad + o\{\mathcal{K}_{x_i,h_{ii}} - \mathbb{E}(\mathcal{K}_{x_i,h_{ii}})\}^2\Big] \\
&= (mg)\{\mathbb{E}(\mathcal{K}_{x_i,h_{ii}})\} + o(h_{ii}) \\
&= (mg)(x_i) + u_i(x_i,h_{ii})(mg)^{(1)}(x_i) + \frac{1}{2}\text{Var}(\mathcal{K}_{x_i,h_{ii}})(mg)^{(2)}(x_i) + o(h_{ii})
\end{aligned}$$

results in $\text{Bias}\{\widehat{f_i}(x_i;h_{ii})\} \to 0$ as $h_{ii} \to 0$, under the condition (A2) and (A3) of discrete associated kernel. For the variance of $\widehat{f_i}$, we get

$$\begin{aligned}
\text{Var}\{\widehat{f_i}(x_i;h_{ii})\} &= \frac{1}{n^2}\sum_{l=1}^{n}\text{Var}\{\mathbb{K}_{x_i,h_{ii}}(X_{il})Y_l\} \\
&= \frac{1}{n}\Big[\mathbb{E}\Big\{\mathbb{K}^2_{x_i,h_{ii}}(X_{il})Y_l^2\Big\} - \mathbb{E}^2\Big\{\mathbb{K}_{x_i,h_{ii}}(X_{il})Y_l\Big\}\Big] \\
&= \frac{1}{n}\Big[\sum_{z_i\in\mathbb{T}}\mathbb{K}^2_{x_i,h_{ii}}(z_i)m^2(z_i)g(z_i) - \Big\{\sum_{z_i\in\mathbb{T}}\mathbb{K}_{x_i,h_{ii}}(z)m(z_i)g(z_i)\Big\}^2\Big]. \qquad (10)
\end{aligned}$$

The first term of the previous equation can be expressed as

$$\begin{aligned}
\sum_{z_i\in\mathbb{T}}\mathbb{K}^2_{x_i,h_{ii}}(z_i)m^2(z_i)g(z_i) &= \mathbb{K}^2_{x_i,h_{ii}}(x_i)m^2(x_i)g(x_i) + \sum_{z_i\neq x_i\in\mathbb{T}}\mathbb{K}^2_{x_i,h_{ii}}(z_i)m^2(z_i)g(z_i) \\
&= \sum_{z_i\neq x_i\in\mathbb{T}} m^2(z_i)g(z_i) + o(h_{ii}) \\
&= -m^2(x_i)g(x_i) + \sum_{z_i\in\mathbb{T}} m^2(z_i)g(z_i) + o(h_{ii}) \\
&= -m^2(x_i)g(x_i) + \mathbb{E}\{m^2(\mathbf{X}^i)\} + o(h_{ii});
\end{aligned}$$

indeed, as $h_{ii} \to 0$, $\mathbb{K}_{x_i,h_{ii}}(x_i) \to 1$ and $\mathbb{K}_{x_i,h_{ii}}(z_i) \to 0$ for $z_i \neq x_i$. Then, for the second term of the equation in (10), we get

$$\begin{aligned}
& \Big\{ \sum_{z_i \in \mathbb{T}} \mathbb{K}_{x_i,h_{ii}}(z) m(z_i) g(z_i) \Big\}^2 \\
= & \sum_{z_i \in \mathbb{T}} \sum_{t_i \in \mathbb{T}} \mathbb{K}_{x_i,h_{ii}}(z_i) \mathbb{K}_{x_i,h_{ii}}(t_i) m(z_i) m(t_i) g(z_i) g(t_i) \\
= & \mathbb{K}^2_{x_i,h_{ii}}(x_i) m^2(x_i) g^2(x_i) + \sum_{z_i \neq x_i \in \mathbb{T}} \sum_{t_i \neq x_i \in \mathbb{T}} \mathbb{K}_{x_i,h_{ii}}(z_i) \mathbb{K}_{x_i,h_{ii}}(t_i) m(z_i) m(t_i) g(z_i) g(t_i) \\
= & \sum_{z_i \neq x_i \in \mathbb{T}} \sum_{t_i \neq x_i \in \mathbb{T}} m(z_i) m(t_i) g(z_i) g(t_i) + o(h_{ii}) \\
= & -m^2(x_i) g^2(x_i) + \mathbb{E}^2\{m(\mathbf{X}^i)\} + o(h_{ii})
\end{aligned}$$

Thus, the variance of $\widehat{f_i}$ results finally in

$$\begin{aligned}
\mathrm{Var}\{\widehat{f_i}(x_i; h_{ii})\} &= -\frac{1}{n} m^2(x_i) g(x_i) + \frac{1}{n} m^2(x_i) g^2(x_i) + \frac{1}{n} [\mathbb{E}\{m^2(\mathbf{X}^i)\} - \mathbb{E}^2\{m(\mathbf{X}^i)\}] + o(h_{ii}) \\
&= \frac{1}{n} m^2(x_i) g(x_i) \{-1 + g(x_i)\} + \frac{1}{n} \mathrm{Var}\{m(\mathbf{X}^i)\} + o(h_{ii}).
\end{aligned}$$

The proof of Proposition 1 requires to use the Hoeffeding lemma on a probability inequality for sums of bounded random variables [5].

Lemma 1 Let $Z_1, Z_2, ..., Z_n$ be i.i.d. random variables with finite second moments. If there exist constants a and b such that $\Pr(Z_i \in [a, b]) = 1$, then given $\epsilon > 0$ we have

$$\Pr\left(| \frac{1}{n} \sum_{i=1}^{n} Z_i | \geq \epsilon \right) \leq 2 \exp\left\{ \frac{-n\epsilon^2}{2\epsilon(b-a) + 2\mathrm{V}ar(Z_i)} \right\}$$

To prove the Proposition 1 we first observe that we have the following decomposition:

$$\widehat{f_i}(x_i; h_{ii}) - f_i(x_i) = [\widehat{f_i}(x_i; h_{ii}) - \mathbb{E}\{\widehat{f_i}(x_i; h_{ii})\}] + \mathrm{Bias}\{\widehat{f_i}(x_i; h_{ii})\}.$$

We already shown that $\mathrm{Bias}\{\widehat{f_i}(x_i; h_{ii})\} \to 0$ as $h_{ii} \to \infty$. Now, our main concern is the term

$$\widehat{f_i}(x_i; h_{ii}) - \mathbb{E}\{\widehat{f_i}(x_i; h_{ii})\} = \frac{1}{n} \sum_{l=1}^{n} Z_l \text{ with } Z_l = \mathbb{K}_{x_i,h_{ii}}(z_{il}) Y_i - \mathbb{E}\{\mathbb{K}_{x_i,h_{ii}}(z_{il}) Y_i\}.$$

For any $x \in \mathbb{T}$, there exists $0 < M_1 < \infty$ and $0 < M_2 < \infty$ such that we have $\mid Z_l \mid \leq M_1$ and $\mathrm{Var}(Z_l) \leq \mathbb{E}\{(K_{x_i;h_{ii}}(X_{il}) - 1)Y_l\}^2 < M_2$ since $K_{x_i;h_{ii}}(\cdot)$ is a probability mass function and Y_l is a bounded random variable. Therefore, according to the Hoeffding lemma, one has

$$\Pr\left(\mid \widehat{f_i}(x_i) - \mathbb{E}\{\widehat{f_i}(x_i)\} \geq \epsilon \mid\right) = \Pr\left(\mid \frac{1}{n}\sum_{l=1}^{n} Z_l \mid \geq \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2\epsilon M_1 + 2M_2}\right),$$

for any $\epsilon > 0$. Consequently, the Borel-Cantelli lemma leads to get $\widehat{f_i}(x_i; h_{ii}) - \mathbb{E}\{\widehat{f_i}(x_i; h_{ii})\} \xrightarrow{a.s.} 0$ since $\sum_{\geq n} \Pr\left(\mid \widehat{f_i}(x_i; h_{ii}) - \mathbb{E}\{\widehat{f_i}(x_i; h_{ii})\} \geq \epsilon \mid\right) < \infty$.

References

[1] Andrianandraina, A. Ventura, T. Senga Kiessé, B. Cazacliu, I. Rachida, H.M.G. van der Werf, (2015). Sensitivity Analysis of Environmental Process Modeling in a Life Cycle Context: A Case Study of Hemp Crop Production. Journal of Industrial Ecology, 19 (6), 978-993.

[2] E. de Rocquigny, (2006). La maîtrise des incertitudes dans un contexte industriel. 1ère partie: une approche méthodologique globale basée sur des exemples. Journal de la Société Française de Statistique, 147(3), 33-71.

[3] J. Geweke, (2009). Complete and Incomplete Econometric Models. Princeton University Press, New Jersey.

[4] W. Hastings, (1970). Monte carlo sampling methods using markov chains and their applications, Biometrica, 57, 97-109.

[5] W. Hoeffding, (1963). Probability inequalities for sums of bounded random variables. Journal of American Statistical Association, 58, 13-30.

[6] T. Homma and A. Saltelli, (1996). Importance measures in global sensitivity analysis of nonlinear models, Reliability Engineering & System Safety, 52(1), 1-17, Elsevier.

[7] C. C. Kokonendji, T. Senga Kiessé and S. S. Zocchi, (2007). Discrete triangular distributions and non-parametric estimation for probability mass function, Journal of Non-parametric Statistics, 8(6), 241-254.

[8] B. Iooss and P. Lemaître, (2015). A review on global sensitivity analysis methods. In Uncertainty Management in Simulation-Optimization of Complex Systems, Springer US, 101-122.

[9] B. Iooss, (2011). Revue sur l'analyse de sensibilité globale de modèles numériques. Journal de la Société Française de Statistique, 152(1), 3-25.

[10] T. Ishigami and T. Homma, (1990). An importance qualification technique in uncertainty analysis for computer models. In:Proceedings of the ISUMA90, first international symposium on uncertainty modelling and analysis, University of Maryland, p.398–403.

[11] Q. Li and J. Zhou, (2005). The uniqueness of cross-validation selected smoothing parameters in kernel estimation of non-parametric models. Econometric Theory, 21, 1017-1025.

[12] X. Luo, Z. Lu and X. Xu, (2014). Non-parametric kernel estimation for the ANOVA decomposition and sensitivity analysis. Reliability Engineering and System Safety 130, 140–148.

[13] R. Martin, I. Lazakis, S. Barbouchi and L. Johanning, (2016). Sensitivity analysis of offshore wind farm operation and maintenance cost and availability. Renewable Energy, 85, 1226-1236.

[14] N. Metropolis, A. W. Rosenbluth, M N. Rosenbluth, A H. Teller and E. Teller, (1953). Equations of state calculations by fast computing machine, J. Chem. Phys, 21, 1087-1093.

[15] M. Ratto, A. Pagano and P. Young, (2007). State Dependent Parameter metamodelling and sensitivity analysis. Computer Physics Communications, 177, 863-876.

[16] M. Rosenblatt, (1956). Remarks on some non-parametric estimates of a density function. Annals of Mathematicals Statistics 27, 832–837.

[17] M. Rosenblatt, (1969). Conditional probability density and regression estimates. In:Krishnaiah PR, editor. Multivariate analysis, 2nd ed. p. 25–31.

[18] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana and S. Tarantola, (2008). Global Sensitivity Analysis: The Primer, Wiley-Interscience.

[19] T. Senga Kiessé and A. Ventura, (2016). Discrete non-parametric kernel estimation for global sensitivity analysis.Reliability Engineering & System Safety, 46, 47-54.

[20] I. M. Sobol, (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation. 55(1), 271-280.

[21] M. S. Sobom and C. C. Kokonendji, (2016). Effects of associated kernels in non-parametric multiple regressions, Journal of Statistical Theory and Practice, 10(2), 456-471.

[22] P. Trucco, E. Cagno, F. Ruggeri and O. Grandea, (2008). A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. Reliability Engineering and System Safety 93, 845-856

[23] N. Zougab, S. Adjabi and C. C. Kokonendji, (2013). Bayesian Approach in Non-parametric Count Regression with Binomial Kernel, Communications in Statistics, 43(5), 1052-1063.

[24] X. Zhang, M. L. King and H. L. Shang, (2016). Bayesian bandwidth selection for a non-parametric regession model with mixed types of regressors, Econometrics, 4(2), 24.

[25] X. Zhang, M. L. King and H. L. Shang, (2014). A sampling algorithm for bandwidth estimation in a non-parametric regression model with a flexible error density. Computational Statistics & Data Analysis, 78, 218-234.

[26] X. Zhang, R. D. Brooks and M. L. King, (2009). A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation, Journal of Econometrics, 153, 21-32, April.

[27] X. Zhang, M. L. King and R. J. Hyndman, (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation, Computational Statistics and Data Analysis, 50(11), 3009-3031.

Sobol indices			
	Continuous case $X_i \in [-\pi, \pi]$	Discrete case $X_i \in \{-3, -2, \ldots, 2, 3\}$	Mixed case $X_1 \in \{-3, -2, \ldots, 2, 3\}$ $X_{i=2,3} \in [-\pi, \pi]$
S_1	0.40	0.42	0.39
S_2	0.29	0.19	0.31
S_{13}	0.31	0.39	0.30

Table 1: Sobol sensitivity indices for continuous [12], discrete [19] and mixed inputs (this study) of the Ishigima test function

	Continuous case $X_i \in [-\pi, \pi]$	Discrete case $X_i \in \{-3, -2, \ldots, 2, 3\}$	Mixed case $X_1 \in \{-3, -2, \ldots, 2, 3\}$ $X_{i=2,3} \in [-\pi, \pi]$
Decomposition of model			
f_0	2.5	2.2	2.5
f_1	$(1 + \frac{\pi^4}{50})\sin(x_1)$	$\left(1 + \frac{14}{5}\right)\sin(x_1)$	$\left(1 + \frac{\pi^4}{50}\right)\sin(x_1)$
f_2	$5\sin^2(x_2) - f_0$	$5\sin^2(x_2) - f_0$	$5\sin^2(x_2) - f_0$
f_{13}	$0.1(x_3^4 - \frac{\pi^4}{5})\sin(x_1)$	$0.1\left(x_3^4 - 14\right)\sin(x_1)$	$0.1(x_3^4 - \frac{\pi^4}{5})\sin(x_1)$
Decomposition of variance			
V	$\frac{29}{8} + \frac{\pi^4}{50} + \frac{\pi^8}{1800}$	$\frac{2}{7}\left(\frac{252}{25}\sum_{x_1=1}^{3}\sin^2(x_1) + 25\sum_{x_2=1}^{3}\sin^4(x_2)\right) - f_0^2$	$\frac{25}{8} + \frac{2}{7}\left(1 + \frac{\pi^4}{25} + \frac{\pi^8}{900}\right)\sum_{x_1=1}^{3}\sin^2(x_1)$
V_1	$\frac{(50+\pi^4)^2}{5000}$	$\frac{722}{175}\sum_{x_1=1}^{3}\sin^2(x_1)$	$\frac{2}{7}\left(1 + \frac{\pi^4}{50}\right)^2\sum_{x_1=1}^{3}\sin^2(x_1)$
V_2	$\frac{25}{8}$	$\frac{2}{7}\sum_{x_2=1}^{3}\{5\sin^2(x_2) - f_0\}^2$	$\frac{25}{8}$
V_{13}	$\frac{2\pi^8}{5625}$	$\frac{56}{1225}\sum_{x_1=1}^{3}\sin^2(x_1)$	$\frac{2}{7}\left(\frac{4\pi^8}{5625}\right)\sum_{x_1=1}^{3}\sin^2(x_1)$

Table 2: Elementary terms of the model and variance decompositions for continuous [12], discrete [19] and mixed inputs (this study) of the Ishigima test function

	Continuous case $X_i \in [-\pi, \pi]$	Discrete case $X_i \in \{-3, -2, \ldots, 2, 3\}$	Mixed case $X_1 \in \{-3, -2, \ldots, 2, 3\}$ $X_{i=2,3} \in [-\pi, \pi]$
Variance of conditional expectation			
$\mathrm{Var}\{\mathbb{E}(Y\|X_{-1})\}$	$\frac{25}{8}$	$\frac{2}{7}\sum_{x_2=1}^{3}\{5\sin^2(x_2) - 2.2\}^2$	$\frac{25}{8}$
$\mathrm{Var}\{\mathbb{E}(Y\|X_{-2})\}$	$\frac{1}{2}\left(1 + \frac{\pi^8}{900} + \frac{\pi^4}{25}\right)$	$\frac{7744}{1225}\sum_{x_2=1}^{3}\sin^2(x_2)$	$\frac{2}{7}\left(1 + \frac{\pi^4}{25} + \frac{\pi^8}{900}\right)\sum_{x_1=1}^{3}\sin^2(x_1)$
$\mathrm{Var}\{\mathbb{E}(Y\|X_{-3})\}$	$\frac{1}{2}\left(1 + \frac{\pi^4}{50}\right)^2 + \frac{25}{8}$	$\frac{722}{175}\sum_{x_1=1}^{3}\sin^2(x_1) + \frac{2}{7}\sum_{x_2=1}^{3}\{5\sin^2(x_2) - 2.2\}^2$	$\frac{2}{7}\left(1 + \frac{\pi^4}{50}\right)^2\sum_{x_1=1}^{3}\sin^2(x_1) + \frac{25}{8}$
Total sensitivity index			
ST_1	0.71	0.81	0.69
ST_2	0.29	0.19	0.31
ST_3	0.31	0.39	0.30

Table 3: Variance of conditional expectations and total sensitivity indices for the Ishigima test function in continuous case [12], discrete case [19] and mixed case (this study)

n	Error model	parameter	Estimate	BMSE	SIF
100	N(0,0.2^2)	h_{11}	0.038	0.001	23.35
		h_{22}	0.045	0.003	24.55
		h_{33}	0.690	0.005	26.42
		σ_m^2	0.498^2	0.054	2.756
100	N(0,0.5^2)	h_{11}	0.041	0.002	29.27
		h_{22}	0.196	0.005	26.77
		h_{33}	0.526	0.015	29.50
		σ_m^2	0.686^2	0.051	9.857

Table 4: The BMSE and SIF indicators for the convergence of MCMC.

	N_{sim}	n	$\bar{\tilde{S}}_1$	$\bar{\tilde{S}}_2$	$\bar{\tilde{S}}_3$	$\bar{\tilde{S}}_{12}$	$\bar{\tilde{S}}_{13}$	$\bar{\tilde{S}}_{23}$	$\bar{\tilde{S}}_{123}$
Analytical values for mixed inputs			0.39	0.31	0	0	0.30	0	0
Mixed kernel estimator	100	100	0.364	0.219	0.024	0.049	0.128	0.109	-0.039
		250	0.363	0.230	0.016	0.026	0.236	0.078	0.006
		500	0.423	0.206	0.002	0.014	0.269	0.072	0.015
		1000	0.394	0.235	0.009	0.012	0.238	0.051	0.006
Continuous kernel estimator	100	100	0.366	0.186	0.029	0.038	0.091	0.076	-0.003
		250	0.359	0.201	0.019	0.024	0.168	0.083	0.006
		500	0.386	0.197	0.011	0.017	0.215	0.057	0.006
		1000	0.391	0.226	0.008	0.012	0.248	0.042	-0.007
Mixed kernel estimator	200	100	0.357	0.216	0.028	0.045	0.101	0.098	-0.029
		250	0.370	0.195	0.016	0.025	0.171	0.080	0.008
		500	0.365	0.182	0.018	0.014	0.227	0.079	0.009
		1000	0.389	0.224	0.008	0.011	0.235	0.049	0.004
Continuous kernel estimator	200	100	0.388	0.311	0.055	0.058	0.355	0.156	-0.325
		250	0.367	0.188	0.016	0.029	0.176	0.080	0.007
		500	0.383	0.209	0.011	0.015	0.214	0.059	0.004
		1000	0.395	0.228	0.008	0.009	0.246	0.045	-0.007

Table 5: Average values of estimated sensitivity indices for mixed input parameters of the Ishigami function.

	N_{sim}	n	$\overline{MAE}_1$	$\overline{MAE}_2$	$\overline{MAE}_3$	$\overline{MAE}_{12}$	$\overline{MAE}_{13}$	$\overline{MAE}_{23}$	$\overline{MAE}_{123}$
		100	**0.036**	**0.100**	**0.024**	0.059	**0.171**	0.109	0.097
Mixed	100	250	0.070	**0.112**	**0.016**	0.029	**0.063**	**0.078**	**0.036**
kernel estimator		500	0.031	**0.106**	**0.008**	**0.014**	**0.058**	0.072	**0.015**
		1000	**0.007**	**0.077**	**0.002**	**0.012**	**0.030**	0.051	**0.006**
		100	0.063	0.144	0.029	**0.047**	0.208	**0.076**	**0.064**
Continuous	100	250	**0.035**	**0.112**	0.019	**0.026**	0.131	0.083	0.037
kernel estimator		500	**0.019**	0.115	0.011	0.020	0.084	**0.057**	0.021
		1000	0.015	0.086	0.009	0.015	0.051	**0.042**	0.016
		100	**0.062**	**0.117**	**0.028**	**0.052**	0.204	**0.102**	0.086
Mixed	200	250	**0.037**	**0.116**	**0.016**	**0.031**	0.128	**0.080**	0.041
kernel estimator		500	**0.020**	**0.103**	**0.009**	**0.014**	**0.072**	0.079	**0.014**
		1000	**0.009**	**0.072**	**0.002**	**0.011**	**0.035**	0.049	**0.004**
		100	0.064	0.129	0.030	**0.052**	**0.194**	0.104	**0.072**
Continuous	200	250	0.037	0.124	**0.016**	**0.031**	**0.124**	**0.080**	**0.039**
kernel estimator		500	0.021	**0.103**	0.011	0.020	0.086	**0.059**	0.020
		1000	0.016	0.085	0.008	0.015	0.053	**0.045**	0.015

Table 6: Average values of the MAE criterion for estimated sensitivity indices for mixed input parameters of the Ishigami function.

		n	$\widehat{ST}_1$	$\widehat{ST}_2$	$\widehat{ST}_3$
Analytical values for mixed inputs			0.69	0.31	0.30
Mixed kernel estimator	100	100	0.635	0.462	0.461
		250	0.608	0.382	0.411
		500	0.697	0.369	0358
		1000	0.692	0.345	0.352
Continuous kernel estimator	100	100	0.607	0.520	0.410
		250	0.702	0.467	0.414
		500	0.702	0.386	0.392
		1000	0.706	0.353	0.371
Mixed kernel estimator	200	100	0.635	0.462	0.461
		250	0.652	0.422	0.317
		500	0.697	0.369	0358
		1000	0.685	0.339	0.345
Continuous kernel estimator	200	100	0.657	0.420	0.410
		250	0.602	0.401	0.414
		500	0.702	0.366	0.342
		1000	0.696	0.353	0.371

Table 7: Average values of estimated total sensitivity indices for mixed input parameters of the Ishigami function.

Input	Description	Unit	Probability distribution
			Continuous
Q	Maximal annual flowrate	m^3/s	truncated Gumbel distribution $\mathcal{G}(1013;558)$ on $[500,3000]$
K_s	Strickler coefficient	-	truncated normal distribution $\mathcal{N}(30;8)$ on $[15,\infty[$
			Discrete
Z_v	River downstream level	m	Triangular $\mathcal{T}(49;50;51)$
Z_m	River upstream levels	m	Triangular $\mathcal{T}(54;55;56)$
H_d	Dyke height	m	Uniform $\mathcal{U}(7;9)$
C_b	Bank level	m	Triangular $\mathcal{T}(55;55:5;56)$
L	Length of the river stretch	m	Triangular $\mathcal{T}(4990;5000;5010)$
B	River width	m	Triangular $\mathcal{T}(295;300;305)$

Table 8: Input parameters of the flood model and their probability distributions

Indices (%)	Z_v	K_s	Q	H_d	C_b
	Mixed inputs				
S_i	**19.4**	10.1	48.3	**14.7**	0.8
ST_i	**17.1**	14.3	59.2	**14.0**	3.7
	Continuous inputs				
S_i	18.3	15.9	35.5	12.5	3.8
ST_i	22.9	25.3	45.5	18.1	3.8

Table 9: Estimated Sobol indices for mixed inputs (this study) and continuous inputs [8] for the flood model. Results of discrete inputs are in bold style.