Discrete non-parametric kernel estimation for global sensitivity analysis

Tristan Senga Kiessé, Anne Ventura

Discrete non-parametric kernel estimation for global sensitivity analysis

Tristan Senga Kiessé *, Anne Ventura

L'Université Nantes Angers Le Mans (LUNAM), Chaire Génie Civil Eco-construction, Institut de Recherche en Génie Civil et Mécanique GeM UMR - CNRS 6183, 58 rue Michel Ange, 44600 Saint-Nazaire, France

This work investigates the discrete kernel approach for evaluating the contribution of the variance of discrete input variables to the variance of model output, via analysis of variance (ANOVA) decomposition. Until recently only the continuous kernel approach has been applied as a metamodeling approach within sensitivity analysis framework, for both discrete and continuous input variables. Now the discrete kernel estimation is known to be suitable for smoothing discrete functions. We present a discrete non-parametric kernel estimator of ANOVA decomposition of a given model. An estimator of sensitivity indices is also presented with its asymtotic convergence rate. Some simulations on a test function analysis and a real case study from agricultural have shown that the discrete kernel approach outper-forms the continuous kernel one for evaluating the contribution of moderate or most influential discrete parameters to the model output.

1. Introduction

In the literature many works about reliability analysis approaches in general, and sensitivity analysis (SA) methods more specially, are related to different problems such as the important case of non-independent random inputs [6] and have various application domains such as maritime industry [19] or environment [14]. In most cases, a mathematical modeling of the studied system is frequently revealed to be useful when the variations of input parameters in a model imply a large variability of the results with some impacts on their accuracy. In this context, the probabilistic way is of interest to encompass the variation in the input parameters of the model. SA methods are then useful to conduct such a study since they aim to evaluate how the variation of input parameters contributes to the variation of the output of a model. Particularly, works in SA have highlighted the encountered interesting aspect concerning the evaluation of the influence of discrete (categorical or ordinal) inputs. Indeed, in system reliability studies, several models involving in various engineering contexts have input discrete variables. And, one of the reliability engineering issues is to accurately evaluate the influence of such parameters.

Amongst various SA approaches, let us consider a well-known method based on the analysis of variance (ANOVA) decomposition of model f for quantifying the influence of input $X_{i,i=1,2,\ldots,k} \in \mathbb{T}$ on the output $Y \in \mathbb{R}$. That method consists of the calculation of sensitivity indices given by [18] such that

$$S_i = \frac{\mathbb{V}\{\mathbb{E}(Y|X_i)\}}{\mathbb{V}(Y)}, \quad S_{ij} = \frac{\mathbb{V}\{\mathbb{E}(Y|X_i,X_j)\}}{\mathbb{V}(Y)}, \ldots \tag{1}$$

The measure of first order S_i evaluates the contribution of the variation of X_i to the total variance of Y, the measure of second order S_{ij} evaluates the contribution of the interaction of X_i and X_j on the output, and so on. Various statistical tools as splines, generalized linear or additive model, polynomial are useful in a metamodeling approach for providing an estimation of conditional expectation $\mathbb{E}(Y|X_i)$ and, consequently, of the main effect sensitivity measure S_i [4]. In the framework of the non-parametric smoothing, some methods as the continuous kernel-based estimation [16] or the State-Dependent Parameter estimation [13] are good choices for estimating $\mathbb{E}(Y|X_i)$. About the two estimation methods, [15,20] are respectively one of the original references of nonparametric and state-dependent parameter estimates. Nowadays [11] have shown that continuous kernel estimation is equal or better than the SDP estimation in terms of performance. However until recently in the literature the continuous kernel estimation is evenly applied on continuous input variables as on discrete ones while discrete kernel estimation suitable for discrete functions is now known [7].

The discrete associated kernel method was developed for smoothing discrete functions as probability mass functions (pmf) or count regression functions on a discrete support $\mathbb{T}$ such as $\mathbb{T} = \mathbb{N}$, the set of positive integers, or $\mathbb{T} = \mathbb{Z}$, the set of integers. For

* Corresponding author. Tel.: +33 2 72 64 87 40.
E-mail addresses: tristan.sengakiesse@univ-nantes.fr (T. Senga Kiessé), anne.ventura@univ-nantes.fr (A. Ventura).

a fixed target x on discrete support $\mathbb{T}$ and a smoothing parameter $h>0$, this method is based on the definition of the *discrete associated kernel* $K_{x,h}(\cdot)$ which is a pmf of random variable (rv) $\mathcal{K}_{x,h}$ with support $\mathbb{S}_x$ satisfying

$x \in \mathbb{S}_x \quad (A1)$,

$\lim_{h\to 0}\mathbb{E}(\mathcal{K}_{x,h}) = x \quad (A2)$,

$\lim_{h\to 0}\mathbb{V}(\mathcal{K}_{x,h}) = 0 \quad (A3)$.

These three assumptions, fulfilled by both continuous and discrete kernels, insure good asymptotic properties for the corresponding kernel estimator [10]. Thus, for $(a,x)\in\mathbb{N}\times\mathbb{T}$ and $h>0$, an example of discrete associated kernel is the discrete symmetric triangular one with rv $\mathcal{K}_{a;x,h}$ on support $\mathbb{S}_x=\{x-a,\ldots,x-1,x,x+1,\ldots,x+a\}$ with a pmf given by

$$\Pr(\mathcal{K}_{a;x,h}=z)=\frac{(a+1)^h-|y-x|^h}{P(a,h)}, z\in\mathbb{S}_x,$$

with $P(a,h)=(2a+1)(a+1)^h-2\sum_{k=1}^{a}k^h$ a normalizing constant. From the discrete kernel methodology, a discrete non-parametric estimator of $\mathbb{E}(Y|X_i)$ was proposed by [1] adapted from the continuous version of [12,22] as follows:

$$\widehat{m}_n(x;h)=\sum_{i=1}^{n}\frac{Y_iK_{x,h}(X_i)}{\sum_{j=1}^{n}K_{x,h}(X_j)},$$

with the arbitrary sequence of smoothing parameters $h=h(n)>0$ fulfilling $\lim_{n\to\infty}h(n)=0$ and $K_{x,h}(\cdot)$ a discrete associated kernel as defined previously.

In this paper the non-parametric regression estimator $\widehat{m}_n$ using a discrete symmeric triangular kernel is investigated as a novel approach in SA methods for providing estimated sensitivity indices for discrete input variables X_i. Thus, the discrete kernel estimation approach is studied as a contribution to reliability analysis for model with discrete input parameters. To illustrate the performance of discrete kernel approach in comparison to continuous kernel approach, some simulations are realized using Ishigami test function and an application is proposed on a real case from agricultural. That latter concerns the evaluation of the influence of some parameters on the environmental impacts generated during the Hemp Crop production by farmers [2].

2. Non-parametric discrete triangular regression

This section presents first a review of the non-parametric univariate regression estimator using symmetric discrete triangular kernel with the asymptotic expansion of its global squared error as presented by [3]. Herein, the optimal convergence rate of the discrete triangular regression estimator is added.

Assume that $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ are n independent copies of (X,Y) defined on $\mathbb{T}(\subseteq\mathbb{Z})\times\mathbb{R}$. We are interested in the non-parametric regression model

$$Y=m(X)+\epsilon,$$

where $m(\cdot)=\mathrm{E}(Y|X=\cdot)$ is an unknown regression function and the random covariate X is independent of the unobservable error variable ϵ's assumed to have zero mean and finite variance. For $a\in\mathbb{N}$, a fixed point $x\in\mathbb{T}$ and a smoothing parameter $h>0$, let us consider the discrete non-parametric estimator $\widehat{m}_n$ of m defined in (2) using a discrete triangular symmetric kernel such that

$$\widehat{m}_n(a;x,h)=\sum_{i=1}^{n}\frac{Y_iK_{a;x,h}(X_i)}{\sum_{j=1}^{n}K_{a;x,h}(X_j)}. \tag{2}$$

First, about some asymptotic properties of estimator $\widehat{m}_n(a;x,h)$ in (2), the asymptotic part of its mean integrated squared error MISE [21] defined by

$$\mathrm{MISE}\{\widehat{m}_n(x;a,h)\}=\sum_{x\in\mathbb{T}}\mathrm{Var}\{\widehat{m}_n(x;a,h)\}+\sum_{x\in\mathbb{T}}\mathrm{Bias}^2\{\widehat{m}_n(x;a,h)\}.$$

is given by

$$\mathrm{AMISE}\{\widehat{m}_n(x;a,h)\}=\frac{h^2}{4}\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)+\{1-h\mathrm{A}(a)\}^2\sum_{x\in\mathbb{T}}\frac{\mathrm{Var}(Y|X=x)}{nf(x)}.$$

This last expression is obtained by calculating asymptotic bias and variance of $\widehat{m}_n(x;a,h)$ in (2) using the following expansions of the modal probability and variance of the discrete symmetric triangular kernel:

$$\Pr(\mathcal{K}_{a;x,h}=x)=1-2h\mathrm{A}(a)+O(h^2) \text{ and } \mathrm{Var}(\mathcal{K}_{a;x,h})=2h\mathrm{V}(a)+O(h^2),$$

with $\mathrm{A}(a)=a\log(a+1)-\sum_{k=1}^{a}\log(k)$ and $\mathrm{V}(a)=\{a(2a^2+3a+1)/6\}\log(a+1)-\sum_{k=1}^{a}k^2\log(k)$ (refer to [3] for more details). Then, an asymptotical optimal bandwidth h_{opt} is obtained by minimizing the asymptotic part AMISE of $\widehat{m}_n(a;x,h)$ in (2) such that

$$\widehat{h}_{opt}(a,n)=\frac{A(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{A}^2(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)+n\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)}\sim C_0n^{-1}$$

with

$$C_0=\frac{\mathrm{A}(a)\sum_{x\in\mathbb{T}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)}.$$

Finally, we get the following inequality:

$$\begin{aligned}\mathrm{AMISE}\{\widehat{m}_n(x;a,h_{opt})\}&\sim n^{-1}\left[\frac{C_0^2}{n}\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)\right.\\&\left.+\left\{1-\frac{C_0}{n}A(a)\right\}^2\sum_{x\in\mathbb{T}}\frac{\mathrm{Var}(Y|X=x)}{f(x)}\right]\le n^{-1}\quad C_0^2\mathrm{V}^2(a)\sum_{x\in\mathbb{T}}W^2(x)\\&\left.+\left[1+\{C_0A(a)\}^2\right]\sum_{x\in\mathbb{T}}\frac{\mathrm{Var}(Y|X=x)}{f(x)}\right)\end{aligned}$$

where AMISE $\{\widehat{m}_n(x;a,h_{opt})\}$ tends to 0 as $n\to\infty$. Thus, for $a\in\mathbb{N}$, the optimal asymptotic root MISE of estimator $\widehat{m}_n$ with kernel $K_{a;x,h}$ is $O(n^{-1/2})$ resulting in

$$m(x)=\widehat{m}_n(x;a,h_{opt})+O(n^{-1/2}),\quad x\in\mathbb{T}.$$

Note that the discrete kernel estimation and the resulting asymptotic expansions of estimator's bias and variance depend on two preconditions: discrete random variable and smooth hypothesis. For $x\in\mathbb{T}$, a discrete associated kernel satisfying assumptions (A1)–(A3) has asymptotically the same behavior that a Dirac type kernel $D_x(y), y\in\mathbb{S}_x$, such that $D_x(y)=1$ at $y=x$ and 0 for any $y\neq x$. That explains also the good asymptotic properties of the corresponding estimator.

3. Non-parametric kernel estimator for sensitivity analysis

This section aims at building the estimator of ANOVA decomposition of the model $Y=f(X_1,X_2,\ldots,X_k)$ given by

$$Y=f_0+\sum_{i=1}^{k}f_i(X_i)+\sum_{i<j}f_{ij}(X_i,X_j)+\cdots+f_{12\ldots k}(X_1,X_2,\ldots,X_k), \tag{3}$$

where each term is defined by

$$f_0=\mathbb{E}(Y), f_i=\mathbb{E}(Y|X_i)-f_0,\quad f_{ij}=\mathbb{E}(Y|X_i,X_j)-f_i-f_j-f_0,\ldots \tag{4}$$

Non-parametric kernel estimation of such model originates in the work of [11] for continuous case. The multidimensional version of non-parametric regression estimator $\widehat{m}_n$ is presented for the calculation of Sobol indices when measuring the contribution of two or more variables to the variance of Y.

3.1. Multivariate non-parametric regression

Let us consider $\mathbf{x}=(x_1,x_2,\ldots,x_d)^\top \in \mathbb{T}^d \subseteq \mathbb{N}^d$ a target vector and $\mathbf{H}=\mathbf{Diag}(h_{11},\ldots,h_{dd})$ a bandwidth matrix with $h_{ii}>0$ such that $\mathbf{H}\equiv\mathbf{H}_n$ goes to the null matrix $\mathbf{0_d}$ as $n\to\infty$. Assume $(\mathbf{X}^k,Y^k), k=1,2,\ldots,n$, be a sequence of iid random vectors defined on $\mathbb{T}^d\times\mathbb{R}$ with $m(\cdot)=\mathbb{E}(Y^k|\mathbf{X}^k=\cdot)$. The multivariate non-parametric regression estimator $\widehat{m}_n^d$ of m can be defined by

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

where the multivariate associated kernel $K_{x,\mathbf{H}}(\cdot)=\prod_{i=1}^{d}K^{[i]}_{x_i,h_{ii}}(\cdot)$ is defined as a product of univariate associated kernel $K^{[i]}_{x_i,h_{ii}}$ with its corresponding rv $\mathcal{K}^{[i]}_{x_i,h_{ii}}$ on support $\mathbb{S}_{x_i,h_{ii}}$, for all $i=1,2,\ldots,d$. Therefore, according to assumptions (A1), (A2) and (A3) for univariate associated kernel, the multivariate associated kernel of support $\mathbb{S}_{\mathbf{x},\mathbf{H}}=\times_{j=1}^{d}\mathbb{S}_{x_i,h_{ii}}$ is a pmf satisfying

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

where $\mathcal{K}_{\mathbf{x},\mathbf{H}}$ denotes the rv with pmf $K_{\mathbf{x},\mathbf{H}}$ and both $\mathbf{a}(\mathbf{x},\mathbf{H})=(a_1(\mathbf{x},\mathbf{H}),\ldots,a_d(\mathbf{x},\mathbf{H}))^\top$ and $\mathbf{B}(\mathbf{x},\mathbf{H})=(b_{ij}(\mathbf{x},\mathbf{H}))_{i,j=1,\ldots,d}$ tend, respectively, to null vector $\mathbf{0}$ and null matrix $\mathbf{0_d}$ as $\mathbf{H}\to\mathbf{0_d}$ [17].

For $\mathbf{a}=(a_1,a_2,\ldots,a_d)^{\mathrm{T}}\in\mathbb{N}^d$, the multivariate estimator $\widehat{m}_n^d$ using discrete symmetric triangular kernel $K_{\mathbf{a},\mathbf{x},\mathbf{H}}(\cdot)=\prod_{i=1}^{d}K^{[i]}_{a_i,x_i,h_{ii}}(\cdot)$ with an optimal bandwidth matrix $\mathbf{H}_{opt}$ such as $h_{opt,ii}\sim C_1 n^{-1/d}$ (with constant C_1) satisfies

$$m(\mathbf{x})=\widehat{m}_n^d(\mathbf{x};\mathbf{a},\mathbf{H}_{opt})+O(n^{-1/(d+1)}),\quad \mathbf{x}\in\mathbb{T}^d.$$

Remark 2. The asymptotic convergence rates $O(n^{-1/2})$ in univariate case and $O(n^{-1/(d+1)})$ in multivariate case only hold for discrete random variables which satisfied proper smooth hypothesis. The previous convergence rates do not hold for continuous random variables where the asymptotic root MISE of non-parametric regression estimator is $O(n^{-2/5})$ in univariate case and $O(n^{-2/(4+d)})$ in multivariate case [11].

The data-driven bandwidth matrix selection procedure is an extension of univariate cross-validation criterion to multivariate case called least squared cross-validation criterion (LSCV). Thus, the optimal bandwidth matrix is obtained by $\mathbf{H}_{cv}=\arg\min_{\mathbf{H}\in\mathcal{H}}$ LSCV $(\mathbf{H})$ such that

$$\mathrm{LSCV}(\mathbf{H})=\frac{1}{n}\sum_{k=1}^{n}\{Y^k-\widehat{m}^d_{n,-k}(\mathbf{X}^k,\mathbf{H})\}^2,$$

with $\widehat{m}^d_{n,-k}$ an estimate computed as $\widehat{m}_n^d$ in Eq. (5) by excluding $\mathbf{X}^k$ and $\mathcal{H}$ is a set of bandwidth matrices $\mathbf{H}$.

3.2. Kernel estimator of ANOVA decomposition

From Eq. (4), the estimator of f_0 can be obtained by

$$\widehat{f}_0=\mathbb{E}\left\{\widehat{m}_n^d(\mathbf{x};\mathbf{H}_{opt})\right\}=\frac{\mathbb{E}_{X^k}\{K_{\mathbf{x},\mathbf{H}}(\mathbf{X}^k)\}}{\sum_{l=1}^{n}K_{\mathbf{x},\mathbf{H}}(\mathbf{X}^l)}\sum_{k=1}^{n}Y^k=\frac{1}{n}\sum_{k=1}^{n}Y^k$$

which is the arithmetic average of $Y^k, k=1,2,\ldots,n$. The terms of first order f_i in the decomposition of the response model equation in (4) are estimated by

$$\widehat{f}_i(x_i;h_{ii})=\frac{1}{n}\sum_{k=1}^{n}K_{x_i,h_{ii}}(X_i^k)Y^k-\frac{1}{n}\sum_{k=1}^{n}Y^k=\frac{1}{n}\sum_{k=1}^{n}\mathbb{K}_{x_i,h_{ii}}(X_i^k)Y^k$$

with $\mathbb{K}_{x_i,h_{ii}}(X_i^k)=K_{x_i,h_{ii}}(X_i^k)-1$. In the same way, the terms of second order f_{ij} in (4) are estimated as follows:

$$\widehat{f}_{ij}(x_i,x_j;\mathbf{H})=\widehat{\mathbb{E}}(Y^k|X_i^k,X_j^k)-\widehat{f}_i-\widehat{f}_j-\widehat{f}_0=\frac{1}{n}\sum_{i=1}^{n}\mathbb{K}_{\mathbf{x},\mathbf{H}}(\mathbf{X}^k)Y^k$$

with $\mathbb{K}_{\mathbf{x},\mathbf{H}}(\cdot)=K_{\mathbf{x},\mathbf{H}}(\cdot)-K^{[i]}_{x_i,h_{ii}}(\cdot)-K^{[j]}_{x_j,h_{jj}}(\cdot)-1$, $\mathbf{x}=(x_i,x_j)$ and $\mathbf{H}=\mathbf{Diag}(h_{ii},h_{jj})$; where $K_{\mathbf{x},\mathbf{H}}(\cdot)$ is the multivariate associated kernel defined in Eq. (5) as a product of univariate associated kernel $K^{[i]}_{x_i,h_{ii}}(\cdot)$ and $K^{[j]}_{x_j,h_{jj}}(\cdot)$.

Now we can obtain the estimated terms of the variance decomposition. From Eq. (3) we first express the decomposition of the total variance of model output Y such as

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

where each variance term is given by

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

We then get the estimated variance terms by

$$\widehat{\mathbb{V}}(Y)=\mathbb{E}_{\mathbf{X}^k}\{\widehat{m}_n^d(x;h)\}^2-\widehat{f}_0^2,\ \widehat{\mathbb{V}_i}=\mathbb{E}_{\mathbf{X}^k}\{\widehat{f}_i(x_i;h_{ii})\}^2,$$
$$\widehat{\mathbb{V}_{ij}}=\mathbb{E}_{\mathbf{X}^k}\{\widehat{f}_{ij}(x_i,x_j;h_{ii},h_{jj})\}^2,\ldots$$

It finally ensues the calculation of the estimated Sobol indices such as for main effect sensisitivity indices in Eq. (1) we get

$$\widehat{S}_i=\frac{\widehat{\mathbb{V}_i}}{\widehat{\mathbb{V}}(Y)}=\frac{S_i+O(n^{-1/(d+1)})}{1+O(n^{-1/(d+1)})}\to S_i \text{ as } n \text{ goes to } \infty.$$

Indeed, we have $f_0=\widehat{f}_0+O(n^{-1/2})$ and $f_i=\widehat{f}_i(x_i;h_{opt})+O(n^{-1/2})$ then $\widehat{\mathbb{V}}(Y)=\mathbb{V}(Y)+O(n^{-1/(d+1)})$ and $\widehat{\mathbb{V}_i}=\mathbb{V}_i+O(n^{-1/(d+1)})$.

4. Simulations on Ishigami test function

In this section we propose to evaluate the application of both discrete and continuous kernel estimation procedures to a test function. We used the terms in the ANOVA decomposition calculated as follows. In the discrete case

$$f_0=\sum_{x_1,x_2,\ldots x_k}f(x_1,x_2,\ldots,x_k)\prod_{i=1}^{k}\Pr(X_i=x_i)$$

$$f_i(x_i)=\sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k}f(x_1,x_2,\ldots,x_k)\prod_{j=1,j\neq i}^{k}\Pr(X_j=x_j)-f_0$$

$$f_{ij}(x_i,x_j)=\sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{j-1},x_{j+1},\ldots,x_k}f(x_1,x_2,\ldots,x_k)\times\prod_{l\neq i,l\neq j}^{k}$$
$$\Pr(X_l=x_l)-f_i-f_j-f_0$$
$$\vdots \tag{6}$$

where
$\Pr(X_i=x_i)=1/(q-p+1)\mathbb{I}_{X_i=x_i},\ x_i\in\mathbb{T}=\{p,p-1,\ldots,q-1,q\}\subseteq\mathbb{Z},$ is the discrete uniform distribution with $\mathbb{I}_A$ the indicator function of any given event A that takes the value 1 if the event A occurs and 0 otherwise. Then, the variance terms in decomposition result in

$$\mathbb{V}(Y)=\mathrm{Var}\{f(X)\}=\sum_{x_1,x_2,\ldots,x_k}\{f(x_1,x_2,\ldots,x_k)\}^2\prod_{i=1}^{k}\Pr(X_i=x_i)-f_0^2$$
$$\mathbb{V}_i=\mathrm{Var}\{f_i(X_i)\}=\sum_{x_i}\{f_i(x_i)\}^2\Pr(X_i=x_i)$$
$$\vdots \tag{7}$$

4.1. Ishigami function

The test function considered is the Ishigami one [5] given by

$$Y=f(X_1,X_2,X_3)=\sin(X_1)+I\sin^2(X_2)+JX_3^4\sin(X_1)$$

where $X_i, i=1,2,3$, are iid variables on $[-\pi,\pi]$ assumed to be discrete and uniformly distributed such as $\mathbb{T}=\{-3,-2,-1,0,1,2,3\}$. The kernel estimator $\widehat{m}_n$ in (2) using discrete symmetric triangular

kernel is applied in comparison to the continuous version of $\widehat{m}_n$ using the Gaussian kernel on support $\mathbb{S}_x=\mathbb{R}$ given by

$$K_{x,h}(t)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-t}{h}\right)^2\right\},\quad t\in\mathbb{S}_x.$$

In practice we will use the parameter value $a=1$ for discrete symmetric triangular kernel since it was proved to be the best in terms of performance [8].

The ANOVA decomposition of Y for Ishigami function is given by

$$Y=f_0+f_1(X_1)+f_2(X_2)+f_3(X_3)+f_{12}(X_1,X_2)+f_{13}(X_1,X_3)+f_{23}(X_1,X_3)+f_{123}(X_1,X_2,X_3).$$

Based on the decomposition of model output Y in (3) and (6), we have

$$f_0=\sum_{x_1,x_2,x_3\in\mathbb{T}} f(x_1,x_2,x_3)\prod_{i=1}^{3}\Pr(X_i=x_i)=\sum_{x_2,x_3\in\mathbb{T}}\left\{\sum_{x_1\in\mathbb{T}}\sin(x_1)\Pr(X_1=x_1)+a\sin^2(x_2)\sum_{x_1\in\mathbb{T}}\Pr(X_1=x_1)+bx_3^4\sum_{x_1\in\mathbb{T}}\sin(x_1)\Pr(X_1=x_1)\right\}=a\sum_{x_2=-3}^{3}\sin^2(x_2)\Pr(X_2=x_2)=\frac{2a}{7}\sum_{x_2=1}^{3}\sin^2(x_2)$$

with $\Pr(X_i=x_i)=(1/7)\mathbb{I}_{X_i=x_i}$, $x_i\in\mathbb{T}$, the discrete uniform distribution. We successively obtain

$$f_1(x_1)=\sum_{x_2,x_3\in\mathbb{T}} f(x_1,x_2,x_3)\prod_{i=2}^{3}\Pr(X_i=x_i)=\sum_{x_2,x_3\in\mathbb{T}}\left\{\sin(x_1)+a\sin^2(x_2)+b\sin(x_1)\sum_{x_3=1}^{3}x_3^4\Pr(X_3=x_3)\right\}\Pr(X_2=x_2)$$

$$-f_0=\left(1+\frac{2b}{7}\sum_{x_3=1}^{3}x_3^4\right)\sin(x_1)$$

then

$$f_2(x_2)=\sum_{x_1,x_3\in\mathbb{T}} a\sin^2(x_2)\Pr(X_1=x_1)\Pr(X_2=x_2)-f_0=a\sin^2(x_2)-f_0$$

and $f_3(x_3)=0$.

For interaction term between different parameters, we have

$$f_{13}(x_1,x_3)=\sum_{x_2\in\mathbb{T}} f(x_1,x_2,x_3)\Pr(X_2=x_2)-f_0-f_1-f_3=b\left(x_3^4-\frac{2}{7}\sum_{x_3=1}^{3}x_3^4\right)\sin(x_1)$$

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

It results in the following decomposition of the variance of Y by using the expressions in (7),

$$\mathbb{V}(Y)=\frac{2}{7}\left\{1+a^2+\frac{2b^2}{7}\sum_{x_3=1}^{3}x_3^8+\frac{4b}{7}\sum_{x_3=1}^{3}x_3^4\right\}\sum_{x_2=1}^{3}\sin^2(x_2)-f_0^2$$

$$\mathbb{V}_1=\frac{2}{7}\left(1+\frac{2b}{7}\sum_{x_3=1}^{3}x_3^4\right)^2\sum_{x_1=1}^{3}\sin^2(x_1)$$

$$\mathbb{V}_2=\frac{2}{7}\sum_{x_2=1}^{3}\{a\sin^2(x_2)-f_0\}^2$$

$$\mathbb{V}_{13}=\frac{4b^2}{49}\left\{\sum_{x_3=1}^{3}\left(x_3^4-\frac{2}{7}\sum_{x_3=1}^{3}x_3^4\right)^2\right\}\sum_{x_1=1}^{3}\sin^2(x_1)$$

and $\mathbb{V}_3=\mathbb{V}_{12}=\mathbb{V}_{23}=0$.

Finally, the main effect sensitivity indices in (1) are $S_1=0.42$, $S_2=0.19$, $S_3=0$ and $S_{13}=0.26$ when considering $I=5$ and $J=0.1$. Note that these values are obviously some approximations of the main effect sensitivity measures of the continuous versions of the same variables defined on $[-\pi,\pi]$. Thus, in the continuous case, we have $S_1=0.40$, $S_2=0.29$, $S_3=0$ and $S_{13}=0.31$.

Fig. 1 illustrates the discrete kernel regression applied for estimating the univariate conditional moment $\mathbb{E}(Y|X_i)$ with discrete inputs X_i. One can see that the inputs X_1 and X_2 have a main effect on Y while X_3 has a null main effect with a globally flat pattern of $\mathbb{E}(Y|X_3)$.

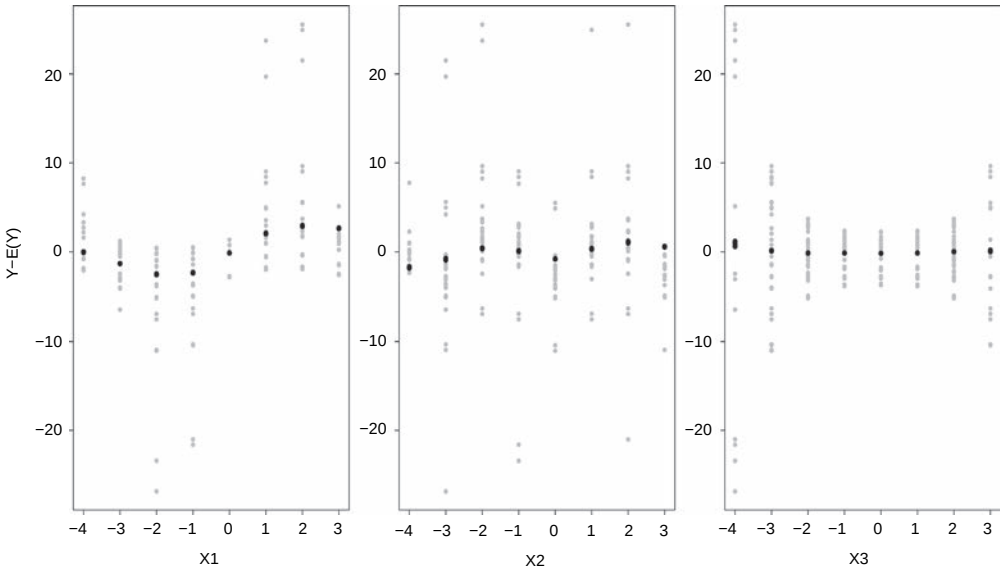

Fig. 1. Plot of centered Ishigami test function (grey dots) and non-parametric discrete kernel estimation (black dots).

4.2. Results

The discrete triangular symmetric regression estimator and the continuous Gaussian kernel regression estimator are applied. In addition for first order indices, a comparison is realized by using symmetric discrete triangular kernel with modified parameter $a_0 \in \mathbb{N}$ such that for $x \in \mathbb{T} = \{-3, -2, -1, 0, 1, 2, 3\}$

$$a_0 = 1 \equiv \begin{cases} a = 0, & \text{if } x = -3 \\ a = 1, & \text{if } x \in \mathbb{T} \setminus \{-3\}. \end{cases}$$

This modification was proposed by [9] as a solution to boundary bias effect.

To evaluate the performance of estimators, we use a MC strategy:

(i) the random generation of a number $N = 100$ samples (X_1, X_2, X_3) of size $n \in \{250, 500, 750, 1000\}$,
(ii) for each sample, the software calculation of the average first order Sobol indices $\overline{S_i} = (1/N)\sum_{j=1}^{N} S_i^{(j)}, i = 1, 2, 3$, and their confidence interval by considering the 5% and 95% percentiles.

The error criterion used is the mean absolute error (MAE) defined as

$$\overline{\mathrm{MAE}}_j(S_i) = \frac{1}{N}\sum_{j=1}^{N} |\, S_i - \hat{S}_i^{(j)} \,|,$$

where $\hat{S}_i^{(j)}$ is the j-th adjustment of main effect sensitivity indice S_i. At last, note that for both discrete and continuous kernel estimators, the bandwidth choice is realized using cross-validation (CV) procedure defined as follows. For a given discrete kernel $\mathrm{K}_{x,h}$ with $x \in \mathbb{T}$ and $h > 0$, the CV procedure is useful for finding an optimal bandwidth $h_{cv} = \arg\min_{h>0} \mathrm{CV}(h)$ minimizing the function $h \mapsto CV(h)$ such that

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \{Y_i - \widehat{m}_{n,-i}(X_i)\}^2.$$

The leave-one-out kernel estimator $\widehat{m}_{n,-i}(X_i; h)$ of $\widehat{m}_n(x; h)$ is calculated by excluding X_i such as

$$\widehat{m}_{n,-i}(X_i; h) = \sum_{j \neq i}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j \neq i}^{n} K_{x,h}(X_j)}.$$

The score function CV is an estimator asymptotically unbiased of the term depending on parameter $h > 0$ in the mean integrated squared error of estimator $\widehat{m}_n(x; h)$.

Looking at the results presented in Table 1, the discrete triangular symmetric kernel estimator is competing to the continuous Gaussian kernel estimator in terms of MAE for evaluating the main effect of parameters X_1 and X_2. In contrast the main effect of parameter X_3 close to 0 is better estimated by the continuous kernel estimator than the discrete kernel estimator. Further, Table 2 shows some results regarding the sensitivity indices for interaction terms between different parameters. Similar to Table 1 the discrete triangular kernel estimator outperforms the continuous Gaussian one and both estimators detect the strong interaction between parameters X_1 and X_3. Finally, as the sample size n increases, the errors provided by the two estimators converge monotonically towards 0 except for the interaction term $X_1 X_3$ due to the number of simulations $N = 100$.

4.2.1. Potential boundary bias for estimated indices S_i

The boundary bias of kernel estimator occurs when there is large probability mass close to the boundary. The accuracy of the estimated sensitivity indices did not seem to be improved by applying discrete kernel estimator with symmetric triangular kernel $a_0 = 1$ used for solving boundary bias. Looking at Table 1, the values of criterion $\overline{\mathrm{MAE}}_i$ were globally closed by using discrete symmetric triangular kernels with $a = 1$ and $a_0 = 1$. However, the sensitivity indices values were over-estimated using modified parameter $a_0 = 1$ while they were under-estimated using parameter $a = 1$ in comparison with the analytical values, in particular for the main effect S_1. Finally, the potential impact of boundary bias was not clearly perceptible but this may be worth exploring, especially regarding at a largest sample size $n \geq 1000$.

4.2.2. Effects of simulation number N

For estimating first order Sobol indices, in what follows we are interested in the influence of the number of simulation by increasing N from 100 to 200. Figs. 2 and 3 reports the comparison of $\overline{\mathrm{MAE}}(S_i), i = 1, 2, 3$, for $N = 100$ and 200, respectively, and sample sizes $n \in \{100, 200, \ldots, 2000\}$. In Fig. 2 corresponding to $N = 100$, the behavior of the curves of criterion $\overline{\mathrm{MAE}}_i$ did not show clearly which estimator provided the better performance for approximating S_1 while discrete estimator outperformed the continuous for S_2 and continuous estimator is better for S_3. Further, the error criterion $\overline{\mathrm{MAE}}_i$ for the three parameters was not monotonically decreasing as sample size n was increasing.

Table 1
Average first order Sobol indices $\overline{S_i}$ calculated by discrete and continuous kernel estimations applied to Ishigami test function.

Input parameters	n	Discrete triangular kernal estimator				Continuous Gaussian kernal estimator	
		with $a=1$		with $a_0=1$			
		$\overline{S_i}$	$\overline{\mathrm{MAE}_i}$	$\overline{S_i}$	$\overline{\mathrm{MAE}_i}$	$\overline{S_i}$	$\overline{\mathrm{MAE}_i}$
X_1	250	0.411	**0.035**	0.424	0.037	0.399	**0.035**
	500	0.408	**0.021**	0.424	0.022	0.397	0.032
	750	0.416	**0.017**	0.424	0.018	0.427	0.022
	1000	0.417	0.016	0.423	**0.015**	0.424	**0.015**
X_2	250	0.219	**0.045**	0.244	0.043	0.219	0.046
	500	0.223	**0.031**	0.234	0.033	0.235	0.034
	750	0.229	0.027	0.232	0.028	0.230	**0.026**
	1000	0.230	**0.024**	0.231	0.025	0.236	0.029
X_3	250	0.009	0.009	0.025	0.025	0.003	**0.003**
	500	0.005	0.005	0.012	0.012	0.001	**0.001**
	750	0.004	0.004	0.008	0.008	0.002	**0.002**
	1000	0.003	0.003	0.006	0.006	0.001	**0.001**

Table 2
Average second order Sobol indices $\overline{S_{ij}}$ calculated by discrete and continuous kernel estimations applied to Ishigami test function.

Interaction terms	n	Discrete triangular kernel estimator with $a=1$		Continuous Gaussian kernel estimator	
		$\overline{\hat{S}_{ij}}$	$\overline{\mathrm{MAE}_{ij}}$	$\overline{\hat{S}_{ij}}$	$\overline{\mathrm{MAE}_{ij}}$
X_1X_2	250	0.014	**0.023**	0.034	0.045
	500	0.005	**0.024**	0.010	**0.024**
	750	0.008	**0.013**	0.006	0.019
	1000	0.001	**0.009**	0.006	0.016
X_1X_3	250	0.285	**0.037**	0.357	0.097
	500	0.286	**0.014**	0.351	0.091
	750	0.289	**0.033**	0.350	0.090
	1000	0.286	**0.027**	0.351	0.091
X_2X_3	250	0.045	**0.045**	0.062	0.062
	500	0.022	**0.030**	0.031	0.031
	750	0.015	**0.015**	0.022	0.022
	1000	0.010	**0.010**	0.015	0.015

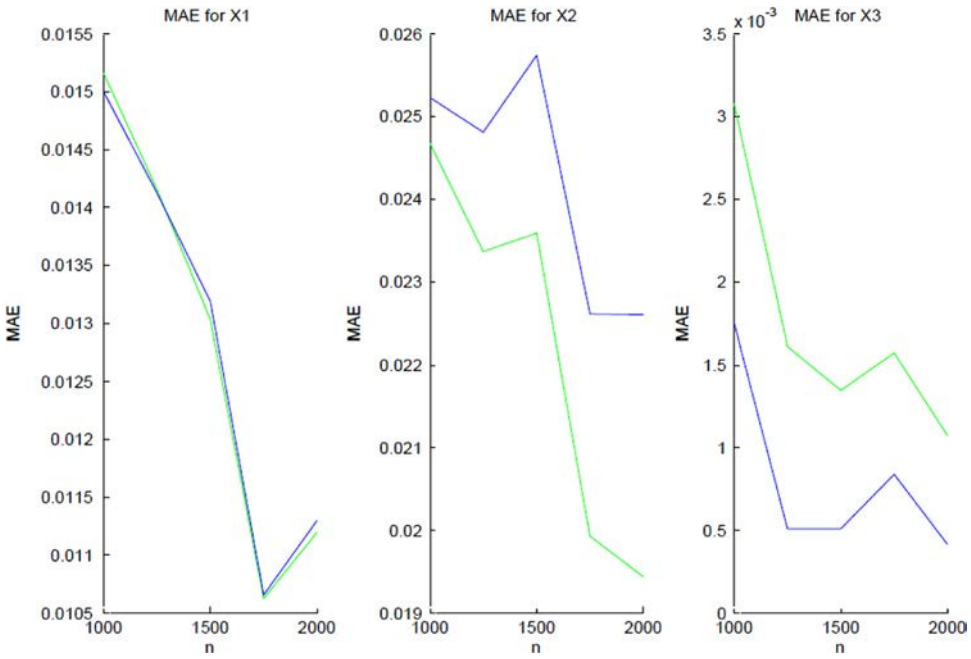

Fig. 2. Comparison of $\overline{\text{MAE}}$ for $N=100$ repetitions on Sobol indices for input parameters (X_1, X_2, X_3) of Ishigami test function by using non-parametric discrete kernel estimations (green line) in comparison with non-parametric continuous kernel estimation (blue line). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

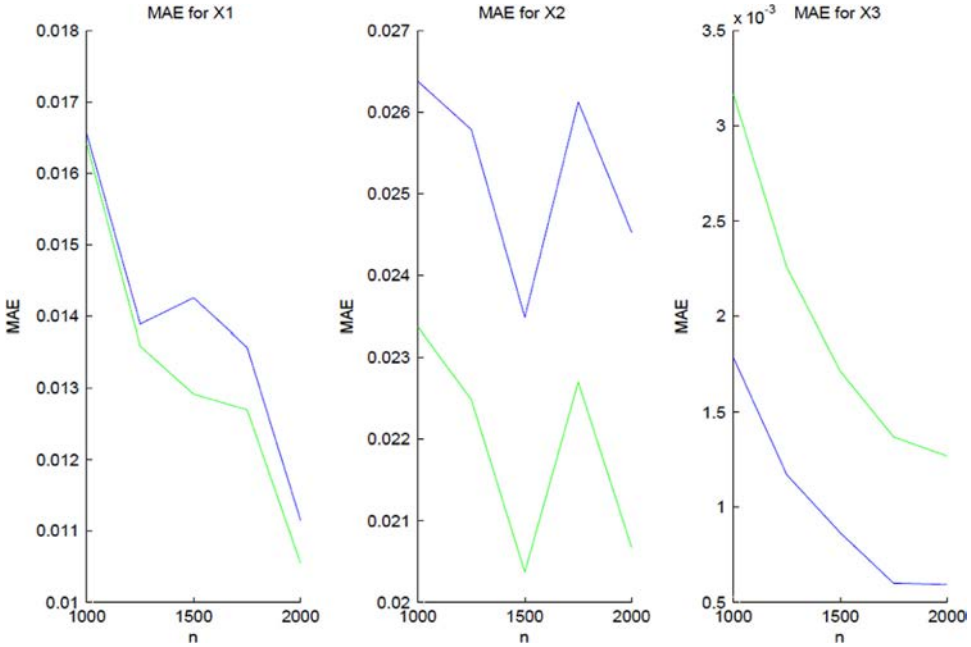

Fig. 3. Comparison of $\overline{\text{MAE}}$ for $N=200$ repetitions on Sobol indices for input parameters (X_1, X_2, X_3) of Ishigami test function by using non-parametric discrete kernel estimations (green line) in comparison with non-parametric continuous kernel estimation (blue line). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

By increasing the number of simulations N from 100 to 200 in Fig. 3, one can see more clearly than in the previous figure that the convergence rates of the discrete kernel estimator were globally faster than that of the continuous kernel estimator for both S_1 and S_2 but not for S_3. Further, the error criterion $\overline{\text{MAE}}_i$ was now monotonically decreasing regarding at parameter X_3 and only for discrete kernel estimator regarding at parameter X_1 but not at all regarding at X_2. Thus the results would be better by increasing $N > 200$. However, an optimal number of simulation N which would provide a monotonically decreasing error criterion for discrete and continuous kernel estimators must be found. For this study we will keep $N=100$ in the following section for the illustration on the real case.

Note that about the computational effort, it is very little (around 17 s for one simulation with sample size $n = 1000$) for both continuous and discrete kernels. Further, there is not any big difference between the two kernel approaches if the same number of samples are used.

Table 3
Average first order Sobol indices $\overline{S_i}$ calculated by kernel estimations and Monte Carlo simulation applied Hemp Crop production system, parameters having integer positive values are written in bold.

Impact categories	Parameters	Kernel estimator with $a=1$		Monte Carlo
		$\overline{S_i}$		$\overline{S_i}$
		Cont.	Discr.	
	Clay content of the soil	0.182	0.201	0.244
Eutrophication	Quantity of nitrogen fertilizer	0.108	–	0.118
	Allocation method	0.057	0.057	0.083
Ecological toxicity	**Engine release year**	0.368	0.375	0.363
	Engine rated power	0.108	0.150	0.098
	Working speed	0.076	–	0.072
Human toxicity	**Engine release year**	0.402	0.404	0.397
	Engine rated power	0.111	0.153	0.103
	Working speed	0.082	–	0.082

5. Simulations on a real case study

The real case study aims at evaluating the input parameters of processes involving in the Hemp Crop production which mostly influence the outcomes of a Life cycle assessment (LCA). LCA is a tool which aims at assessing environmental impacts over all whole *life cycle* of a product, i.e. from the extraction of raw materials to end of life as well as recycling by including fabricating and utilization. The processes involving at each step of a life cycle system are described by some numerical or analytical models. The outcomes considered herein are environmental impacts on human health (*human toxicity*) and marine or terrestrial ecosystem (*eutrophication, ecological toxicity*). Three models are used in this study (fuel consumption model for agricultural operations, exhaust emissions models from engines used for agricultural operations, and direct field emissions models from the crop) for the Hemp Crop production corresponding to a total of 52 input parameters, amongst them some discrete parameters uniformly distributed on support $\mathbb{N}$ such as *the release year of agricultural engine*, *the engine rated power* or *the clay content of the soil*. Note that a parameter name *allocation method* is presented but it is a qualitative parameter with coded values 1 or 2 . For more details, the complete study is presented in [2]. In this part, the first order Sobol indices of some discrete input parameters are only studied. For the second order Sobol indices, a complete work is in progress on this real case study which requires a multivariate estimation mixing continuous and discrete kernels.

Some direct MC simulations are used in comparison to kernel methods for estimating the conditional expectation $\mathbb{E}(Y|X)$ where the response Y is an environmental impact indicator. Similarly to the previous part, a MC strategy is applied: to obtain similar results of Sobol indices, we use $N=100$ replications of sample size $n=1000$ for kernel methods while we use $N=500$ of sample size $n=5000$ for direct MC simulations. Table 3 presents the results of main effect sensitivity Sobol indices of the three most influential (discrete and continuous) input parameters obtained. Note that the discrete kernel method is applied only on discrete parameters.

In this case study, the estimated values of Sobol first order indices cannot be compared to theoretical values, we can just compare between them the Sobol indice values provided by the three methods. It appears that the three methods provide some results of same order. In particular, the discrete kernel estimation provides some values greater than continuous kernel estimation, except for the allocation method parameter which is not influential. Thus, the results provided by discrete kernel method are confirmed.

6. Concluding remarks

This work is interested in discrete kernel estimation approach as a novel approach in reliability analysis suitable when discrete input parameters involved in a model. It pursues various works in sensibility analysis framework on the application of (continuous) non-parametric kernel method for estimating sensitivity indices calculated from ANOVA decomposition. The studied discrete non-parametric kernel method is appropriate only for discrete or ordinal variables, not for real continuous cases. The discrete approach is proposed as a competing approach more suitable for discrete input parameters than continuous non-parametric kernel method. First, the theoretical asymptotical convergence rate of the discrete kernel estimator is better than that of the continuous kernel estimator. Then, the realized simulations point out that the discrete approach is faster than continuous one in the sense of average MAE for moderate or most influential input parameters. However, the discrete kernel approach seems to be limited when estimating the main influence of discrete input parameters having a weak contribution to the variance of the model output, in comparison to continuous approach which provides better estimation in this case. The boundary bias was treated in this work but may be something worth exploring in the future as well, to further verify and improve the estimation accuracy of the proposed approach. In addition, a minimum number of simulations depending on sample sizes needs to be found to insure that the error criterion used is monotonically decreasing.

The aspect of curse of dimensionality is not included in this work. Thus, some future prospects will be to investigate multivariate estimation mixing continuous and discrete kernels when evaluating the influence of both discrete (count and categorical) and continuous input variables.

Acknowledgements

The research and education chair of civil engineering and eco-construction is financed 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.

References

[1] Abdous B, Kokonendji CC, Senga Kiessé T. On semiparametric regression for count explanatory variables. J Stat Plan Inference 2012;142(6):1537–48.

[2] Andrianandraina, Ventura A, Senga Kiessé T, Cazacliu B, Rachida I, van der Werf HMG. Sensitivity analysis of environmental process modeling in a life cycle context: a case study of Hemp crop production. J Ind Ecol 2014. http://dx.doi.org/10.1111/jiec.12228.

[3] Cuny HE, Senga Kiessé T. On modeling wood formation using parametric and semiparametric regressions for count data. Commun Stat—Simul Comput 2014. http://dx.doi.org/10.1080/03610918.2013.875570.

[4] Iooss B. Review of global sensitivity analysis of numerical models. J Soc Française Stat 2011;152(1):3–25.

[5] Ishigami T, Homma T. 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; 1990. p. 398–403.

[6] Rosenblatt M. Conditional probability density and regression estimates. In: Krishnaiah PR, editor. Multivariate analysis, 2nd ed. New York: Academic Press; 1969. p. 25–31.
[7] Kokonendji CC, Senga Kiessé T. Discrete associated kernel method for smoothing discrete function and extensions. Stat Methodol 2011;8(6):497–516.
[8] Kokonendji CC, Zocchi SS. Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions. Stat Probab Lett 2010;80:1655–62.
[9] Kokonendji CC, Senga Kiessé T, Zocchi SS. Discrete triangular distributions and non-parametric estimation for probability mass function. J Nonparametr Stat 2007;19:241–54.
[10] Libengué FG. Méthode non-paramétrique des noyaux asociés mixtes et applications [Ph.D. thesis]. Universities of Franche-Comté (France) and Ouagadougou (Burkinafaso); 2013.
[11] Luo X, Lu Z, Xu X. Non-parametric kernel estimation for the ANOVA decomposition and sensitivity analysis. Reliab Eng Syst Saf 2014;130:140–8.
[12] Nadaraya EA. On estimating regression. Theory Probab Appl 1964;9:141–2.
[13] Ratto M, Pagano A, Young P. State dependent parameter metamodelling and sensitivity analysis. Comput Phys Commun 2007;177:863–76.
[14] Rechard RP, Liu H-H, Tsang YW, Finsterle S. Site characterization of the Yucca Mountain disposal system for spent nuclear fuel and high-level radioactive waste. Reliab Eng Syst Saf 2014;122:32–52.
[15] Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat 1956;27:832–7.
[16] Rosenblatt M. Conditional probability density and regression estimates. In: Krishnaiah PR, editor. Multivariate analysis, 2nd ed. New York: Academic Press; 1969. p. 25–31.
[17] Sobom MS, Kokonendji CC. Effects of associated kernels in non-parametric multiple regressions; 2015. arxiv.org/abs/1502.01488v1.
[18] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 2001;55:271–80.
[19] Trucco P, Cagno E, Ruggeri F, Grandea O. A Bayesian belief network modelling of organisational factors in risk analysis: a case study in maritime transportation. Reliab Eng Syst Saf 2008;93:845–56.
[20] Young PC. Time variable and state dependent modelling of nonstationary and nonlinear time series. In: Rao TS, editor. Developments in time series analysis. London: Chapman and Hall; 1993. p. 374–413.
[21] Wand MP, Jones MC. Kernel smoothing. London, New York: Chapman and Hall; 1995.
[22] Watson GS. Smooth regression analysis. Sankhyā Ser A 1964;26:359–72.