Non-parametric methods for global sensitivity analysis of model output with dependent inputs

Thierry A. Mara, Stefano Tarantola, Paola Annoni

Non-parametric methods for global sensitivity analysis of model output with dependent inputs*

Thierry A. Mara[a,*], Stefano Tarantola[b], Paola Annoni[c]

[a]*PIMENT, Department of Physics, UFR ST, University of Reunion Island, 15 Avenue Rene Cassin 97715 Saint-Denis, Reunion*
[b]*Econometrics and Applied Statistics Group, Joint Research Centre, European Commission, Ispra (VA), Italy*
[c]*Economic Analysis Unit, Directorate General for Regional and Urban Policy, European Commission, Brussels, Belgium*

Abstract

This paper addresses the issue of performing global sensitivity analysis of model output with dependent inputs. First, we define variance-based sensitivity indices that allow for distinguishing the independent contributions of the inputs to the response variance from their mutual dependent contributions. Then, two sampling strategies are proposed for their non-parametric, numerical estimation. This approach allows us to estimate the sensitivity indices not only for individual inputs but also for groups of inputs. After testing the accuracy of the non-parametric method on some analytical test functions, the approach is employed to assess the importance of dependent inputs on a computer model for the migration of radioactive substances in the geosphere.

Keywords: dependent inputs, Rosenblatt transformation, variance-based sensitivity indices, dependent contributions, independent contributions, Iman & Conover sampling procedure, radionuclide migration

*Corresponding author: mara@univ-reunion.fr
*Published in Environmental Modelling & Software, Vol. 72, pp. 173–183, 2015, doi: 10.1016/j.envsoft.2015.07.010

Contents

1. Introduction

Uncertainty and sensitivity analyses (UASA) of model output have become popular during the last decade. Their success comes from their ability in providing relevant information into complex processes via their numerical simulation modelling. To perform UASA, the modeller explores the input space and evaluates the impact of the inputs on the numerical model responses (see Norton, 2015, for a recent review). The choice of the model responses to analyze depends on the objective of the survey (see Saltelli and Tarantola, 2002, for some possible sensitivity analysis settings). To quantitatively assess the importance of the model inputs for a given response, two global sensitivity measures can be computed: the variance-based sensitivity measures (Sobol', 1993; Homma and Saltelli, 1996) and the moment-independent measures (Borgonovo, 2006; Plischke et al., 2013; Pianosi and Wagener, 2015). Variance-based sensitivity measures are most often computed because of their ability to provide a picture of the model structure (Oakley and O'Hagan, 2004).

In the recent literature, two types of global sensitivity analysis (GSA) can be distinguished: the case of independent inputs (when the joint pdf can be expressed as the product of its marginals) and the case of dependent inputs (when the previous does not hold). Dependency may be caused by the presence of constraints across inputs (*e.g.* inputs defined on a non-rectangular domain) or by the fact that experimental data and expert judgement are used. Linear correlation between inputs, treated for example in Kucherenko et al. (2012) and Mara and Tarantola (2012), is a particular case of dependency.

The case of independent input is simpler to tackle because: i) many computational and efficient methods exist to compute the sensitivity indices, ii) samples are easy to generate, iii) variance-based sensitivity indices allow

to rank the inputs by order of importance (Sobol', 2001), iv) the ANOVA (ANalysis Of VAriance)-decomposition is unique and shows the model structure and v) analytical benchmarks are easy to derive.

On the contrary, UASA of model output with dependent inputs is more challenging. Indeed, none of the points above is valid any longer. In particular, the ANOVA decomposition cannot provide a description of the model structure (Oakley and O'Hagan, 2004). Even the definition of helpful and easy-to-compute global sensitivity indices is an issue. One of the most popular ones is the so-called first-order sensitivity index also called correlation ratio (McKay, 1996), which allows to address the issue of factor prioritization (Saltelli and Tarantola, 2002). Sensitivity indices with dependent inputs can be computed by either parametric (*i.e.* interpolation, regression, see Oakley and O'Hagan, 2004; Da Veiga et al., 2009; Li et al., 2010) or non-parametric, non-model based methods (McKay, 1996; Xu and Gertner, 2008a; Xu, 2013).

Kucherenko et al. (2012) extend the definition of first-order and total sensitivity indices, initially defined in Sobol' (1993) and Homma and Saltelli (1996), to the case of dependent inputs. The authors propose a non-parametric method to estimate the new sensitivity indices. The method requires the knowledge of the conditional probability densities and the capability of sampling from those. Gaussian copulas are employed as a basis for the generation of the conditional samples.

Mara and Tarantola (2012) introduce a set of sensitivity indices to analyze models for the specific case of correlated inputs, distinguishing between correlated and uncorrelated contributions of inputs on model responses. The computation of those indices is undertaken with a parametric method, specifically the polynomial chaos expansion.

In the present article, we establish the link between the indices proposed by Kucherenko et al. (2012) and those defined in Mara and Tarantola (2012) and we show that they can be defined for the more general case of dependent input by considering the Rosenblatt transformation (Rosenblatt, 1952). Rosenblatt transformation requires the knowledge of the conditional densities and as such, is comparable to the approach of Kucherenko et al. (2012). In the particular case of correlated input, we proposed a simpler method that estimates the sensitivity indices without requiring the knowledge of conditional probability densities. By contrast, this method is computationally more expensive than Kucherenko et al. (2012). The proposed approach can be easily extended to estimate sensitivity indices for groups of inputs.

The paper is organized as follows: in Section 2 we define the variance-based sensitivity indices for model with dependent input variables. In Section 3, we provide the sampling strategy to estimate the sensitivity indices. Sections 4 and 5 are devoted to numerical examples, namely testing the method on analytical test functions and on a more complex computer model for radionuclide transport in the geosphere. Section 6 concludes.

2. Definition of the sensitivity indices

2.1. From dependent variables to independent variables

Let $f(\mathbf{x})$ be a square integrable function over an n-dimensional space and $\mathbf{x} = \{x_1, \cdots, x_n\} \sim p(\mathbf{x})$ a continuous random vector defined by a joint probability density function $p(\mathbf{x})$. Thanks to the Rosenblatt transformation (RT), described in Appendix A, it is always possible to transform $\mathbf{x}$ into a random vector $\mathbf{u} = (u_1, \cdots, u_n)$ uniformly and independently distributed over the unit hypercube $\mathbb{K}^n = [0,1]^n$. RT is not unique in general; there are actually $n!$ possibilities corresponding to all possible permutations of

the elements of $\mathbf{x}$. In our case, we consider only the RT obtained after circularly reordering the set $(x_1, \ldots, x_n)$, resulting in n RT transformations. We denote by $\mathbf{u}^i \ \forall i = 1, \ldots, n$ the Rosenblatt transformation of the set $(x_i, x_{i+1}, \ldots, x_n, x_1, \ldots, x_{i-1})$. We write:

$$(x_i, x_{i+1}, \cdots, x_n, x_1, \cdots, x_{i-1}) \sim p(\mathbf{x}) \xrightarrow{RT} (u_1^i, \cdots, u_n^i) \sim \mathrm{U}^n(0,1) \tag{1}$$

Such a mapping is bijective, and we have $f(x_1, \ldots, x_n) = g_i(\mathbf{u}^i)$. Because the u_k^i's are independent, instead of performing the UASA of $f(\mathbf{x})$, we perform the UASA of $g_i(\mathbf{u}^i)$. Indeed, global sensitivity analysis is well-established for functions with independent input variables.

2.2. Variance-based sensitivity measures

For a set of independent variables$\mathbf{u}^i = (u_1^i, \cdots, u_n^i)$, uniformly distributed over the unit hypercube $\mathbb{K}^n = [0,1]^n$, the following ANOVA-decomposition is proven unique by Sobol' (1993)

$$g_i(\mathbf{u}^i) = g_0 + \sum_{j_1=1}^{n} g_{j_1}(u_{j_1}^i) + \sum_{j_2>j_1}^{n} g_{j_1 j_2}(u_{j_1}^i, u_{j_2}^i) + \cdots + g_{1\cdots n}(u_1^i, \ldots, u_n^i) \tag{2}$$

where, $g_0 = E[g_i(\mathbf{u}^i)] = \int_{\mathbb{K}^n} g_i(\mathbf{u}^i) \mathrm{d}\mathbf{u}^i$ and the summands in (2) are such that,

$$\int_0^1 g_{j_1 \cdots j_s} \mathrm{d}u_{j_k}^i = 0 \text{ if } k \in \{1, \cdots, s\}. \tag{3}$$

As a consequence, the summands in (2) are orthogonal and the following variance decomposition can be derived,

$$V = \sum_{j_1=1}^{n} V_{j_1} + \sum_{j_2>j_1}^{n} V_{j_1 j_2} + \cdots + V_{1\cdots n} \tag{4}$$

where, $V_{j_1 \cdots j_s} = \int_{\mathbb{K}^s} g_{j_1 \cdots j_s}^2 du_{j_1}^i du_{j_2}^i \ldots du_{j_s}^i$. The variance-based sensitivity measures (also called Sobol' indices) are defined by dividing (4) by the total variance V. The following variance-based sensitivity indices can then be defined:

- the first-order sensitivity index that measures the contribution of u_k^i to the variance of f,

$$S_{u_k^i} = \frac{V\Big[E\big[g_i(\mathbf{u}^i)|u_k^i\big]\Big]}{V\big[g_i(\mathbf{u}^i)\big]} = \frac{V_k}{V}, \tag{5}$$

- the total sensitivity index that measures the overall contribution of u_k^i to the variance of f (including its marginal and cooperative effects with the other inputs),

$$ST_{u_k^i} = \frac{E\Big[V\big[g_i(\mathbf{u}^i)|\mathbf{u}_{\sim k}^i\big]\Big]}{V\big[g_i(\mathbf{u}^i)\big]} = \frac{\sum_{s=1}^{n}\sum_{\{j_1,\cdots,j_s\}\ni k} V_{j_1\cdots j_s}}{V}. \tag{6}$$

The individual variance-based sensitivity indices have the following properties:

1. $0 \leq S_{u_k^i} \leq ST_{u_k^i} \leq 1$, the higher $ST_{u_k^i}$ the more u_k^i is a relevant input while if $ST_{u_k^i} = 0$, u_k^i is irrelevant and can be fixed at an arbitrary value in its uncertainty range without changing the variance of f.
2. $\sum_{k=1}^{n} S_{u_k^i} \leq 1$ and $\left(1 - \sum_{k=1}^{n} S_{u_k^i}\right)$ represents the amount of variance explained by the interactions. An additive function is such that $\sum_{k=1}^{n} S_{u_k^i} = 1$ and consequently $S_{u_k^i} = ST_{u_k^i}$, $\forall k \in [\![1, n]\!]$.

Links can be established between the sensitivity indices of u_k^i and those of x_k, as shown in the next Section.

2.3. Interpretation of the individual sensitivity indices

The joint pdf of $\mathbf{x}$ can be written in terms of conditional distributions as:

$$p(\mathbf{x}) = p(x_i)p(x_{i+1}|x_i)\dots p(x_n|x_i, x_{i+1}, \dots x_{n-1})p(x_1|x_i, \dots, x_n)\dots p(x_{i-1}|\mathbf{x}_{\sim(i-1)}) \tag{7}$$

with $\mathbf{x}_{\sim(i-1)} = (x_1, x_2, \ldots, x_{i-2}, x_i, \ldots, x_n)$. The Rosenblatt transformation in Equation (1) establishes a one-to-one mapping (*i.e.* bijection) between $\mathbf{x}$ and $\mathbf{u}^i$,

$$[(x_i), (x_{i+1}|x_i), \ldots, (x_1|(x_i, x_{i+1} \ldots x_n)), \cdots, (x_{i-1}|\mathbf{x}_{\sim(i-1)})] \leftrightarrow (u_1^i, u_2^i, \ldots, u_n^i). \tag{8}$$

The sensitivity indices of u_1^i are those of x_i because $u_1^i = F_1(x_i)$, where F_1 is the unconditional cumulative distribution function of x_i. Hence, denoting by S_i and ST_i the sensitivity indices of x_i, we have $S_i = S_{u_1^i}$ and $ST_i = ST_{u_1^i}$. The indices S_i and ST_i include the effects of the dependence of x_i with other inputs. For this reason Mara and Tarantola (2012) call them the *full* sensitivity indices of x_i.

The sensitivity indices of u_2^i are those of $(x_{i+1}|x_i)$ and represent the sensitivity indices of x_{i+1} without its mutual dependent contribution with x_i. Similarly for the other sensitivity indices.

The sensitivity indices of u_n^i are of particular interest. Indeed, they represent the effects of x_{i-1} that are not due to its dependence with the other variables $\mathbf{x}_{\sim(i-1)}$. In Mara and Tarantola (2012), the authors call them the uncorrelated effects of x_{i-1}. In the present paper, we call these sensitivity indices the independent contributions of x_{i-1} and we denote them by $S_{i-1}^{ind} = S_{u_n^i}$ and $ST_{i-1}^{ind} = ST_{u_n^i}$. Note that, because of the inequalities $0 \leq S_{u_k^i} \leq ST_{u_k^i} \leq 1$ previously discussed, it is straightforward to infer that $0 \leq S_i \leq ST_i \leq 1$ (case $k = 1$) and $0 \leq S_i^{ind} \leq ST_i^{ind} \leq 1$ ($k = n$). However, there are no such relationships between (S_i, ST_i^{ind}), (S_i, S_i^{ind}), (S_i^{ind}, ST_i) and (ST_i^{ind}, ST_i).

An input whose importance is only due to its dependence with other inputs has full total effect ($ST_i > 0$) but a null total independent contribution ($ST_i^{ind} = 0$). However, in this case the input cannot be fixed because it brings

a contribution through its dependency with one or more other inputs. An input can be fixed only when both ST_i and ST_i^{ind} are null.

2.4. Formal definitions of the sensitivity indices

The following new sensitivity measures come as a consequence of the previous discussion,

$$S_i = \frac{V\left[E\left[g_i(\mathbf{u}^i)|u_1^i\right]\right]}{V\left[g_i(\mathbf{u}^i)\right]} = \frac{V\left[E\left[f(\mathbf{x})|x_i\right]\right]}{V\left[f(\mathbf{x})\right]}, \tag{9}$$

$$ST_i^{ind} = \frac{E\left[V\left[g_{i+1}(\mathbf{u}^{i+1})|\mathbf{u}_{\sim n}^{i+1}\right]\right]}{V[g_{i+1}(\mathbf{u}^{i+1})]} = \frac{E\left[V\left[f(\mathbf{x})|\mathbf{x}_{\sim i}\right]\right]}{V\left[f(\mathbf{x})\right]} \tag{10}$$

$$S_i^{ind} = \frac{V\left[E\left[g_{i+1}(\mathbf{u}^{i+1})|u_n^{i+1}\right]\right]}{V\left[g_{i+1}(\mathbf{u}^{i+1})\right]} = \frac{V\left[E\left[f(\mathbf{x})|(\bar{x}_i|\mathbf{x}_{\sim i})\right]\right]}{V\left[f(\mathbf{x})\right]}, \tag{11}$$

$$ST_i = \frac{E\left[V\left[g_i(\mathbf{u}^i)|\mathbf{u}_{\sim 1}^i\right]\right]}{V\left[g_i(\mathbf{u}^i)\right]} = \frac{E\left[V\left[[f(\mathbf{x})|(\bar{\mathbf{x}}_{\sim i}|x_i)\right]\right]}{V\left[f(\mathbf{x})\right]} \tag{12}$$

$\forall i = 1, \ldots, n$, with the convention that $\mathbf{u}^1 = \mathbf{u}^{n+1}$, in formulas (10) and (11). The variables with an overbar are conditionally distributed.

The previous definitions can be extended to the definition of the sensitivity indices for groups of inputs. For instance, let us set $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ where $\mathbf{y}$ is a subset of s inputs ($s < n$). Then, we have

$$S_{\mathbf{y}} = \frac{V\left[E\left[f(\mathbf{x})|\mathbf{y}\right]\right]}{V\left[f(\mathbf{x})\right]}, \tag{13}$$

$$ST_{\mathbf{y}}^{ind} = 1 - S_{\mathbf{z}} = \frac{E\left[V\left[f(\mathbf{x})|\mathbf{z}\right]\right]}{V\left[f(\mathbf{x})\right]} \tag{14}$$

$$S_{\mathbf{y}}^{ind} = \frac{V\left[E\left[f(\mathbf{x})|(\bar{\mathbf{y}}|\mathbf{z})\right]\right]}{V\left[f(\mathbf{x})\right]}, \tag{15}$$

$$ST_{\mathbf{y}} = 1 - S_{\mathbf{z}}^{ind} = \frac{E\left[V\left[f(\mathbf{x})|(\bar{\mathbf{z}}|\mathbf{y})\right]\right]}{V[f(\mathbf{x})]} \tag{16}$$

Formulas (9-10) and (13-14) were first defined in Kucherenko et al. (2012). These authors also derived the integral definitions that we recall in the next subsection. The integral definitions of the sensitivity indices are reported in the next section. The proofs are given in Appendix B.

2.5. Integral definitions of the individual sensitivity indices

By setting $\mathbf{y} = x_i$ and $\mathbf{z} = \mathbf{x}_{\sim i}$ in the equations (39), (40), (43) and (44) in Appendix B, the following four integral definitions of the individual sensitivity indices of x_i are derived,

$$S_i = \frac{1}{V}\left[\int_{\mathbb{R}^n} f(x_i, \mathbf{x}_{\sim i}) p(x_i, \mathbf{x}_{\sim i}) \mathrm{d}x_i \mathrm{d}\mathbf{x}_{\sim i} \left(\int_{\mathbb{R}^{n-1}} f(x_i, \bar{\mathbf{x}}'_{\sim i}) p(\bar{\mathbf{x}}'_{\sim i}|x_i) \mathrm{d}\bar{\mathbf{x}}'_{\sim i} - \int_{\mathbb{R}^n} f(x'_i, \mathbf{x}'_{\sim i}) p(x'_i, \mathbf{x}'_{\sim i}) \mathrm{d}\mathbf{x}'_{\sim i} \mathrm{d}x'_i\right)\right] \tag{17}$$

$$ST_i^{ind} = \frac{1}{2V} \int_{\mathbb{R}^{n+1}} \left(f(x'_i, \mathbf{x}'_{\sim i}) - f(\bar{x}_i, \mathbf{x}'_{\sim i})\right)^2 p(x'_i, \mathbf{x}'_{\sim i}) p(\bar{x}_i|\mathbf{x}'_{\sim i}) \mathrm{d}x'_i \mathrm{d}\bar{x}_i \mathrm{d}\mathbf{x}'_{\sim i} \tag{18}$$

$$S_i^{ind} = \frac{1}{V}\left[\int_{\mathbb{R}^n} f(\bar{x}_i, \mathbf{x}_{\sim i}) p(\bar{x}_i|\mathbf{x}_{\sim i}) p(\mathbf{x}_{\sim i}) \mathrm{d}\bar{x}_i \mathrm{d}\mathbf{x}_{\sim i} \left(\int_{\mathbb{R}^{n-1}} f(\bar{x}_i, \mathbf{x}'_{\sim i}) p(\mathbf{x}'_{\sim i}) \mathrm{d}\mathbf{x}'_{\sim i} - \int_{\mathbb{R}^n} f(x'_i, \mathbf{x}'_{\sim i}) p(x'_i, \mathbf{x}'_{\sim i}) \mathrm{d}x'_i \mathrm{d}\mathbf{x}'_{\sim i}\right)\right] \tag{19}$$

$$ST_i = \frac{1}{2V} \int_{\mathbb{R}^{n+1}} \left(f(x'_i, \bar{\mathbf{x}}'_{\sim i}) - f(x_i, \bar{\mathbf{x}}'_{\sim i})\right)^2 p(\mathbf{x}'_{\sim i}|x'_i) p(\bar{\mathbf{x}}'_{\sim i}|x'_i) p(x'_i) p(x_i) \mathrm{d}\bar{\mathbf{x}}'_{\sim i} \mathrm{d}x_i \mathrm{d}x_i \tag{20}$$

Six samples of size N are necessary to evaluate $f(x_i, \mathbf{x}_{\sim i})$, $f(x'_i, \mathbf{x}'_{\sim i})$, $f(x_i, \bar{\mathbf{x}}'_{\sim i})$, $f(\bar{x}_i, \mathbf{x}'_{\sim i})$ $f(\bar{x}_i, \mathbf{x}_{\sim i})$, $f(x'_i, \bar{\mathbf{x}}'_{\sim i})$ and compute the sensitivity indices. They are generated with the inverse Rosenblatt transformation (see in Appendix A Equation (37)). In Section 3, we show that $4n$ samples are necessary to compute all the set of sensitivity indices.

2.6. The case of correlated input

The Rosenblatt transformation requires the knowledge of conditional probability densities. Such information is unknown in some applications (see the example in Section 5). However, when the dependency structure is defined by a rank correlation matrix $\mathbf{R}$, the procedure of Iman and Conover (1982) (IC), described hereafter, can be used to generate the input sample. Let $\mathbf{z}^{nc}$ be a vector of independent standard normal variables and $\{F_1, \cdots, F_n\}$ the marginal cumulative distributions of the set of correlated inputs $\mathbf{x}$. Although the z_i^{nc}'s are independent, a sample of $\mathbf{z}^{nc}$ has a correlation matrix $\mathbf{C}_z$ that is not a perfect identity matrix. The procedure to produce $\mathbf{x}$ is based on the following four-step algorithm,

1. Compute the lower Cholesky factorization of $\mathbf{R}$, $\mathbf{R} = \mathbf{L}\mathbf{L}^T$, with:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}.$$

   and denote $\mathbf{\Lambda}$, the inverse matrix of $\mathbf{L}$,

$$\mathbf{\Lambda} = \mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nn} \end{bmatrix}.$$

2. Find $\mathbf{Q}$ such that, $\mathbf{C}_z = \mathbf{Q}\mathbf{Q}^T$
3. Generate the normally distributed correlated variables,

$$\mathbf{z^c} = \mathbf{z}^{nc}(\mathbf{Q}^{-1})^T\mathbf{L}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \tag{21}$$

4. Perform the following transformation: $x_j = F_j^{-1}\left(\phi(z_j^c)\right)$ where ϕ is the cumulative standard normal distribution.

From the latter relationship, it can be guessed that the sensitivity indices of x_j are those of z_j^c since there is an one-to-one mapping between $\mathbf{x}$ and $\mathbf{z}^c$. Indeed, we note that $f(x_1, \ldots, x_n) = f(F_1^{-1}(\phi(z_1^c)), \ldots, F_n^{-1}(\phi(z_n^c))) = g(\mathbf{z}^{nc})$. It has to be noted that the Pearson correlation matrix $\mathbf{C}$ is not equal to the Spearman rank correlation matrix $\mathbf{R}$. If $\mathbf{C}$ is desired, then $\mathbf{R}$ must be modified in order to get $\mathbf{C}$ with the IC procedure. The empirical formulas derived in Liu and Kiureghian (1986) or the algorithm proposed in Li et al. (2008) can be used to achieve this goal.

Besides, from Equation (21), it can be deduced that,

- $z_1^{nc} = z_1^c \sim p(z_1^c)$,
- $z_2^{nc} = \lambda_{21} z_1^c + \lambda_{22} z_2^c \sim p(z_2^c | z_1^c)$,
- $\cdots$,
- $z_n^{nc} = \lambda_{n1} z_1^c + \cdots + \lambda_{nn} z_n^c \sim p(z_n^c | \mathbf{z}_{\sim n}^c)$

and $(x_1, x_2, \ldots, x_n) \sim p(x_1)p(x_2|x_1)\ldots p(x_n|\mathbf{x}_{\sim n})$. Hence, the vector $\mathbf{z}^{nc}$ plays the same role as the Rosenblatt transform $\mathbf{u}^1$ except that $\mathbf{z}^{nc}$ is a vector of independent standard normal variables while $\mathbf{u}^1$ is a vector of independent variables uniformly distributed over the unit hypercube. Once again, instead of performing the UASA of $f(\mathbf{x})$ we perform the UASA of $g(\mathbf{z}^{nc})$ with independent variables.

3. Monte Carlo methods

3.1. Sampling strategy with RT

To compute $(S_i, S_{i-1}^{ind}, ST_i, ST_{i-1}^{ind})$, four samples of a given size N are necessary if one refers to the non-parametric method of Saltelli (2002). They are generated with the inverse Rosenblatt transformation (see in Appendix

A Equation (37)). First, an uniformly distributed sample $\mathbf{u}^i$ is created to produce $\mathbf{x} \sim p(\mathbf{x})$. Then, a second independent uniformly distributed sample $\mathbf{u}^{i'}$ is created to produce $\mathbf{x}' \sim p(\mathbf{x})$. The two previous samples are combined as follows $(u_1^i, \mathbf{u}_{\sim 1}^{i'})$ to obtain $(x_i, \bar{\mathbf{x}}'_{\sim i}) \sim p(x_i)p(\bar{\mathbf{x}}'_{\sim i}|x_i)$. From these three samples one can compute S_i and ST_i. Finally, a fourth sample $(\bar{x}_{i-1}, \mathbf{x}'_{\sim i-1}) \sim p(\bar{x}_{i-1}|\mathbf{x}_{\sim i-1})p(\mathbf{x}_{\sim i-1})$ is created from $(u_n^i, \mathbf{u}_{\sim n}^{i'})$ and allows for evaluating $(ST_{i-1}^{ind}, S_{i-1}^{ind})$.

$$(u_1^i, \cdots, u_n^i) \xrightarrow{IRT} (x_i, \cdots, x_n, x_1, \cdots x_{i-1}) \sim p(\mathbf{x}) \tag{22}$$

$$(u_1^{i'}, \cdots, u_n^{i'}) \xrightarrow{IRT} (x'_i, \cdots, x'_n, x'_1, \cdots x'_{i-1}) \sim p(\mathbf{x}') \tag{23}$$

$$(u_1^i, u_2^{i'}, \cdots, u_n^{i'}) \xrightarrow{IRT} (x_i, \bar{x}'_{i+1}, \cdots, \bar{x}'_n, \cdots \bar{x}'_{i-1}) \sim p(x_i)p(\bar{\mathbf{x}}'_{\sim i}|x_i) \tag{24}$$

$$(u_2^{i'}, \cdots, u_{n-1}^{i'}, u_n) \xrightarrow{IRT} (x'_i, x'_{i+1}, \cdots, x'_{i-2}, \bar{x}_{i-1}) \sim p(\mathbf{x}'_{\sim i-1})p(\bar{x}_{i-1}|\mathbf{x}'_{\sim i-1}) \tag{25}$$

Three samples are necessary to assess the full sensitivity indices of the group of factors $\mathbf{y} = (x_1, \ldots, x_s)$, respectively $\mathbf{u}^1$, $\mathbf{u}^{1'}$ and $(u_1^{1'}, u_2^{1'}, \ldots, u_s^{1'}, u_{s+1}^1, \ldots, u_n^1)$. In order to estimate $(S_i, S_i^{ind}, ST_i, ST_i^{ind})$, $\forall i = 1, \ldots, n$, $4n$ samples are required, obtained with the four previous samples by varying $i \in [\![1, n]\!]$.

3.2. Sampling strategy with IC procedure

As discussed in the previous section, the sensitivity indices of z_1^{nc} are the full indices of x_1 while those of z_n^{nc} are the independent indices of x_n. To compute $(S_i, S_{i-1}^{ind}, ST_i, ST_{i-1}^{ind})$, four samples of a given size N are necessary.

The four samples are of the form, $\mathbf{z}^{nc} = (z_1^{nc}, \ldots, z_n^{nc})$, $\mathbf{z}^{nc'} = (z_1^{nc'}, \ldots, z_n^{nc'})$, $(z_1^{nc'}, z_2^{nc'}, \ldots, z_n^{nc})$ and $(z_1^{nc}, \ldots, z_{n-1}^{nc}, z_n^{nc'})$, with $\mathbf{z}^{nc}$ and $\mathbf{z}^{nc'}$ two independent standard normal samples such that,

$$(z_1^{nc}, \cdots, z_n^{nc}) \xrightarrow{IC} (x_i, \cdots, x_n, x_1, \cdots x_{i-1}) \sim p(\mathbf{x}) \tag{26}$$

$$(z_1^{nc'}, \cdots, z_n^{nc'}) \xrightarrow{IC} (x'_i, \cdots, x'_n, x'_1, \cdots x'_{i-1}) \sim p(\mathbf{x}') \tag{27}$$

$$(z_1^{nc}, z_2^{nc'}, \cdots, z_n^{nc'}) \xrightarrow{IC} (x_i, \bar{x}'_{i+1}, \cdots, \bar{x}'_n, \cdots \bar{x}'_{i-1}) \sim p(x_i)p(\bar{\mathbf{x}}'_{\sim i}|x_i) \tag{28}$$

$$(z_1^{nc'}, \cdots, z_{n-1}^{nc'}, z_n^{nc}) \xrightarrow{IC} (x'_i, x'_{i+1}, \cdots, x'_{i-2}, \bar{x}_{i-1}) \sim p(\mathbf{x}'_{\sim i-1})p(\bar{x}_{i-1}|\mathbf{x}'_{\sim i-1}) \tag{29}$$

Three samples are necessary to assess the full sensitivity indices of the group of factors $\mathbf{y}$, respectively $\mathbf{z}^{nc}$, $\mathbf{z}^{nc'}$ and $(z_1^{nc'}, z_2^{nc'}, \ldots, z_s^{nc'}, z_{s+1}^{nc}, \ldots, z_n^{nc})$. In order to estimate $(S_i, S_i^{ind}, ST_i, ST_i^{ind})$, $\forall i = 1, \ldots, n$, $4n$ samples are required. For this purpose, the Iman and Conover's (IC) sampling procedure is repeated n times by circularly reordering the vector $\mathbf{x}$ and changing the rank correlation matrix accordingly.

3.3. Monte-Carlo estimators

Let us denote by $\mathbf{x}$ and $\mathbf{x}'$ two independent samples of size N obtained from either (22-23) or (26-27), depending on the strategy employed (RT or IC). We denote by $\mathbf{x}^i$ and $\mathbf{x}^{i-1}$ the sample obtained with (24-25) respectively (or (28-29)). The Monte-Carlo estimates of the sensitivity indices are given

by,

$$\hat{S}_i = \frac{\frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k) \times (f(\mathbf{x}_k^i) - f(\mathbf{x}'_k))}{\hat{V}} \tag{30}$$

$$\hat{ST}_i^{ind} = \frac{\frac{1}{N}\sum_{k=1}^{N} \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}} \tag{31}$$

$$\hat{S}_{i-1}^{ind} = \frac{\frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)}{\hat{V}} \tag{32}$$

$$\hat{ST}_i = \frac{\frac{1}{N}\sum_{k=1}^{N} \left(f(\mathbf{x}_k^{i}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}} \tag{33}$$

where, $\mathbf{x}_k^* = (x_{k1}^*, \cdots, x_{kn}^*)$ is the k-th MC trial in the sample $\mathbf{x}^*$, $k \in [\![1, N]\!]$ and $\hat{V}$ is the total variance estimate that can be computed as the average of the total variances computed with each sample $\mathbf{x}^*$.

4. Numerical test cases

4.1. A linear model

Let us consider the simple linear model $f(x_1, x_2, x_3) = x_1 + x_2 + x_3$, where the x_i's are standard normal random variables with correlation matrix:

$$\mathbf{C} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}.$$

Analytical sensitivity indices (S_i,ST_i^{ind}) for this linear model are derived in Mara and Tarantola (2012). The accuracy of the non-parametric approach can then be assessed. We considered two different sets of correlation coefficients, $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.5, 0.8, 0)$ and $(-0.5, 0.2, -0.7)$ respectively. For both cases, the computation of the 3×2 indices requires the 3×4 samples

generated as follows,

$$
\begin{aligned}
(z_1^{nc}, z_2^{nc}, z_3^{nc}) &\xrightarrow{IC} (x_1, x_2, x_3), (x_2, x_3, x_1), (x_3, x_1, x_2) \\
(z_1^{nc'}, z_2^{nc}, z_3^{nc}) &\xrightarrow{IC} (x'_1, \bar{x}_2, \bar{x}_3), (x'_2, \bar{x}_3, \bar{x}_1), (x'_3, \bar{x}_1, \bar{x}_2) \\
(z_1^{nc}, z_2^{nc}, z_3^{nc'}) &\xrightarrow{IC} (x_1, x_2, \bar{x}'_3), (x_2, x_3, \bar{x}'_1), (x_3, x_1, \bar{x}'_2) \\
(z_1^{nc'}, z_2^{nc'}, z_3^{nc'}) &\xrightarrow{IC} (x'_1, x'_2, x'_3), (x'_2, x'_3, x'_1), (x'_3, x'_1, x'_2)
\end{aligned}
$$

where, $x'_i \sim p(x_i)$, $x_i \sim p(x_i)$ and $(\bar{x}_i, \bar{x}_j) \sim p(x_i, x_j|x_k)$ with $i \neq j \neq k$.

Note that in IC procedure, we used only four samples of $\mathbf{z}^{nc}$ to generate the twelve samples of $\mathbf{x}$. For each sample of $\mathbf{z}^{nc}$, four samples of $\mathbf{x}$ is obtained by circularly permuting the variables in the set. Of course, the circular permutations imply a modification of the correlation matrix $\mathbf{C}$. For instance, for the set (x_2, x_3, x_1) we have

$$
\mathbf{C} = \begin{bmatrix} 1 & \rho_{23} & \rho_{12} \\ \rho_{23} & 1 & \rho_{31} \\ \rho_{12} & \rho_{31} & 1 \end{bmatrix}.
$$

We used pseudo-random samples of size $N = 1\,000$ each. Therefore, the total computational cost is 12 000. Our discussion focuses on the computation of (S_i, ST_i^{ind}), $\forall i = 1, 2, 3$. The bootstrap estimates of size 10 000 have been performed for each couple of indices. The mean bootstrap estimates of S_i and ST_i^{ind} are shown in Table 1 for two different correlation structures. We can note that the estimates are rather accurate.

For the case $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.5, 0.8, 0)$, we find that $S_1 = 0.94$ which means that the overall - correlated and independent - contribution of x_1 to the output variance is 94% (Table 1). The remaining amount of variance (6%) is then explained by x_2 and x_3 without their correlated contributions with x_1. Consequently, for this correlation structure, the knowledge of x_1 only, suffices to predict the model output accurately. Figure 1, on the left,

[Insert Table 1 about here]

depicts the original three-dimensional scatterplots of the sample (x_1, x_2, x_3) (the circles) and the sample generated from x_1 (crosses along a straight line). On the right, the scatterplots show that the responses are very close. The determination coefficient R^2 is equal to 0.94 which coincides with the first-order effect of x_1. Alternatively, by noting that $ST_1^{ind} = 0.02$, one can infer that the independent contribution of x_1 is only 2%. This means that 98% of the variance is explained by the pair (x_2, x_3) also via their correlation with x_1.

In the case of negative correlations $(\rho_{12}, \rho_{13}, \rho_{23}) = (-0.5, 0.2, -0.7)$ the independent contributions are larger than the full marginal contribution (see also Xu and Gertner, 2008b). Also in this case, S_1 is the largest first-order index. On the one hand, if the modeller wants to decrease the variance of the output s/he should reduce the uncertainty on x_1. On the other hand, the modeller should avoid to focus on x_2 as s/he would not be able to achieve a consistent reduction in the output variance ($S_2 = 0.04$). Should it be possible to exclude x_2 from the model? The answer is no, because the contribution of x_1 and x_3 is only 63% (*i.e.* $1 - ST_2^{ind} = 0.63$).

[Insert Figure 1 about here]

4.2. A non-linear model with non-linear dependences

The function analyzed in this example is : $f(\mathbf{x}) = x_1x_2 + x_3x_4$ where $(x_1, x_2) \in [0, 1[^2$ is uniformly distributed within the triangle $x_1 + x_2 \leq 1$ and $(x_3, x_4) \in]0, 1]^2$ is uniformly distributed within the triangle $x_3 + x_4 \geq 1$. In this case, the inputs are strictly dependent and the procedure of Iman & Conover is not appropriate to generate the samples because the dependency

across inputs is not described by a rank correlation matrix. The Rosenblatt transformation is therefore necessary.

The Rosenblatt transformation of (x_1, x_2) yields the following mapping (see details in C),

$$\begin{cases} x_1 &= 1 - \sqrt{1 - u_1^1} \\ x_2 &= u_2^1 \sqrt{1 - u_1^1} \end{cases} \tag{34}$$

in which $(u_1^1 \neq 1, u_2^1) \in \mathbb{K}^2$. Because of the symmetry, the RT transformation of (x_2, x_1) is obtained by simply inverting x_1 and x_2 in Equation (34). In the same way, RT of (x_3, x_4) writes,

$$\begin{cases} x_3 &= \sqrt{u_3^1} \\ x_4 &= (u_4^1 - 1)\sqrt{u_3^1} + 1 \end{cases} \tag{35}$$

Therefore, performing the sensitivity analysis of $f(\mathbf{x}) = x_1 x_2 + x_3 x_4$ is equivalent to performing the sensitivity analysis of $g_1(\mathbf{u}^1) = \left(1 - \sqrt{1 - u_1^1}\right) u_2^1 \sqrt{1 - u_1^1} + \sqrt{u_3^1}\left((u_4^1 - 1)\sqrt{u_3^1} + 1\right)$ with the independent variables $\mathbf{u}^1 \in \mathbb{K}^4 - (1, \cdot, 0, \cdot)$. We recall that the sensitivity indices of u_1^1 are those of x_1, those of u_2^1 are those of x_2 that are not due to its dependence with x_1, and so on. In this numerical exercise, we are interested by the variance-based sensitivity indices of groups of variables, namely: S_1 the full first-order effect of x_1 (*i.e.* u_1^1), S_{12}^{closed} the full closed-order effect of (x_1, x_2) (*i.e.* (u_1^1, u_2^1)), S_{123}^{closed} the full closed-order effect of (x_1, x_2, x_3) (*i.e.* (u_1^1, u_2^1, u_3^1)), knowing that $S_{1234}^{closed} = 1$.

By using the *pick and freeze* method (Saltelli, 2002), five samples of $\mathbf{u}^1$ are necessary: $(u_1^1, u_2^1, u_3^1, u_4^1)$, $(u_1^{1'}, u_2^1, u_3^1, u_4^1)$, $(u_1^{1'}, u_2^{1'}, u_3^1, u_4^1)$, $(u_1^{1'}, u_2^{1'}, u_3^{1'}, u_4^1)$ and $(u_1^{1'}, u_2^{1'}, u_3^{1'}, u_4^{1'})$. The two independent reference samples $\mathbf{u}^1$ and $\mathbf{u}^{1'}$ are uniformly distributed over the unit hypercube. Samples of size $N = 1024$ are generated and the bootstrap technique is employed to assess the variability of the sensitivity estimates.

[Insert Figure 2 about here]

The analytical values of the sensitivity indices are: $S_1 = \frac{1}{30}$, $S_{12}^{closed} = \frac{1}{10}$ and $S_{123}^{closed} = \frac{1}{3}$. They are plotted in Figure 2 with the estimated sensitivity indices. The results are very accurate and the mean bootstrap estimates are very closed to the true values. The bias of the estimator is very small. For the sake of completeness, the other sensitivity indices are $S_1 = S_2 = \frac{1}{30}$, $ST_1^{ind} = ST_2^{ind} = \frac{1}{15}$, $S_3 = S_4 = \frac{7}{30}$ and $ST_3^{ind} = ST_4^{ind} = \frac{2}{3}$. These results indicate that (x_3, x_4) are the most preponderant variables. Because the pair (x_1, x_2) does not interact with (x_3, x_4), a reduction of $S_{12}^{closed} = 10\%$ of the variance of $f(\mathbf{x})$ would be achieved by fixing (x_1, x_2).

5. Application to radionuclides transport in the geosphere

5.1. The Level E model

We now discuss the application to a model developed by the Nuclear Energy Agency of the OECD for predicting the radiologic release to humans due to the underground migration of radionuclides from a nuclear waste disposal site. The model is known as Level E (OECD/NEA PSAC User Group, 1989; OECD/NEA PSAG User Group, 1993) and, with time, has become a benchmark model in global sensitivity analysis studies (Saltelli and Marivoet, 1990; Saltelli and Tarantola, 2002; Ratto et al., 2007; Borgonovo et al., 2012).

Level E simulates the radiological dose released from a nuclear waste disposal site to humans. The dose is due to the underground migration of radionuclides. Level E has been widely utilized in the literature. We recall its utilization as a benchmark for Monte Carlo calculations in OECD/NEA PSAC User Group (1989), OECD/NEA PSAG User Group (1993), for variance-based techniques in

[Insert Table 2 about here]

Saltelli and Tarantola (2002), for emulators in Ratto et al. (2007) and, recently, for moment-independent methods, in Castaings et al. (2012) . While we refer to OECD/NEA PSAC User Group (1989) for a detailed description of the model, a succinct illustration is proposed here. The repository is represented as a point source and the one-dimensional dispersion is tracked over geological time scales (up to 10^7 years). The model describes the transport of iodine (^{129}I), neptunium, uranium and thorium ($^{237}Np \rightarrow {}^{233}U \rightarrow {}^{229}Th$) through two geosphere layers characterized by specific hydro-geological properties. The governing equations account for radioactive decay, dispersion, advection and chemical reaction between the migrating nuclides and the porous medium. Model output uncertainty is caused by twelve uncertain model inputs whose probability distributions were assigned on the basis of expert judgement (see Table 2 and OECD/NEA PSAG User Group, 1993). Two output of this model are analyzed in the literature. The maximum radiological dose simulated over the time period up to 10^7 years and the radiological dose at given times.

5.2. Results and discussion

A sensitivity analysis of the level E model was performed by accounting for the correlations among the twelve input parameters shown in Table 3. To simplify the analysis the initial set of 12 parameters is reduced to six factors by grouping all the parameters related to a specific layer i, $i = 1, 2$:

$$Gr_{(1)} = (v^{(1)}, l^{(1)}, R^{(1)}, R_C^{(1)})$$

$$Gr_{(2)} = (v^{(2)}, l^{(2)}, R^{(2)}, R_C^{(2)})$$

[Insert Table 3 about here]

Results are shown in Figure 3 which presents the set of sensitivity indices $(S_i, S_i^{ind}, ST_i, ST_i^{ind})$ at given times. The time interval simulated by the model is from 20,000 to 9,000,000 years in the future. The parameters describing layer 1, grouped in group 1 (Gr_1), together with the stream flow rate W are found to be the most important ones in terms of model output sensitivity at almost any times. Still the two factors Gr_1 and W behave in a different way. As regards Gr_1, its full first-order index S_{Gr_1} and full total index ST_{Gr_1} assume high values at almost any time points (see graphs on the left column of Figure 3), suggesting that some variables in Gr_1 are important, both in terms of direct influence on the model output and through interactions. Gr_1 is also important in terms of independent contribution to the output uncertainty given that the values of $S_{Gr_1}^{ind}$ and $ST_{Gr_1}^{ind}$ are pretty high for most time points (see graphs on the right column of Figure 3).

Parameter W contributes to the output sensitivity through the full total index ST_W and its independent component ST_W^{ind} (see graphs in the bottom row of Figure 3). Parameters related to layer 2, grouped in Gr_2, have in general a lower influence on the output variability. They show high values of the full total index ST_{Gr_2}, correlated and independent (bottom-left graph of Figure 3), for almost all points in time. The uncorrelated component of ST_{Gr_2} is non-irrelevant as shown by the bottom-right graph of Figure 3. The other input parameters are less important. Among them, the containment time T is influencing only through correlation and interactive effects at certain time points (high values of full total index with almost null values of full first-order effect and independent components, both for the first-order and total effect). Clearly, T is a spurious parameter only contributing to the model response variance because of its correlation with $v^{(1)}$.

[**Insert Figure 3 about here**]

6. Conclusion

We propose a non-parametric strategy to compute sensitivity indices of model outputs with dependent inputs. These indices were initially introduced in Kucherenko et al. (2012) and Mara and Tarantola (2012). The procedure allows for detecting those inputs that contribute to the variation of the model response per se and through their dependency with the other inputs. We introduce and use the inverse Rosenblatt transformation that is particularly suited to compute the sensitivity indices when the dependency structure across the inputs is not described by a (rank) correlation matrix. Its implementation is delicate because it requires the knowledge of the conditional densities. When this latter is not known, but the (rank) correlation structure is, a simpler procedure based on the technique of Iman and Conover can be adopted.

The implementation of the proposed procedure for groups of inputs is conceptually easier than in Kucherenko et al. (2012), whereby sampling from probability densities conditional upon two or more inputs can be challenging. Comparatively to the emulation-based approach derived in Mara and Tarantola (2012) and to the procedure proposed by Kucherenko et al. (2012), the proposed non-parametric method is easier to implement, yet computationally more expensive.

The proposed method, as well as that by Kucherenko et al. (2012), allows for computing bootstrap confidence intervals for the sensitivity indices. On the contrary, this is not possible with Mara and Tarantola (2012) approach because the emulation-based step cannot be bootstrapped.

The application to a benchmark radionuclide model, the so-called Level E,

allows us to show the usefulness of the proposed approach which distinguishes inputs that are important through a direct effect on the output from those that are relevant only indirectly, i.e. through the dependency structure.

Acknowledgments: The authors are grateful to the anonymous reviewers for their insightful comments that helped improving the manuscript. The corresponding author would like to thank the French National Research Agency for its financial support (Research project RESAIN n°ANR-12-BS06-0010-02).

References

Borgonovo, E., 2006. Measuring uncertainty importance: Investigation and comparison of alternative approaches. Risk Analysis 26 (5), 1349–1361.

Borgonovo, E., Castaings, W., Tarantola, S., 2012. Model emulation and moment-independent sensitivity analysis: An application to envi- ronmental modelling. Environmental Modelling and Software 34, 105–115.

Castaings, W., Borgonovo, E., Morris, M. D., Tarantola, S., 2012. Sampling strategies in density-based sensitivity analysis. Environmental Modelling and Software 38, 13–26.

Da Veiga, S., Wahl, F., Gamboa, F., 2009. Local polynomial estimation for sensitvity analysis of models with correlated inputs. Technometrics 51 (4), 452–463.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis. Reliability Engineering and System Safety 52, 1–17.

Iman, R. I., Conover, W. J., 1982. A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics: Simulation and Computation 11, 311–334.

Kucherenko, S., Tarantola, S., Annoni, P., 2012. Estimation of global sensitivity indices for models with dependent variables. Computer Physics Communications 183, 937–946.

Li, G., Rabitz, H., Yelvington, P. E., ad F. Bacon, O. O. O., Kolb, C. E., Schoendorf, J., 2010. Global sensitivity analysis for systems with independent and/or correlated inputs. J. Chemical Physics 114, 6022–6032.

Li, H., Lu, Z., Yuan, X., 2008. Nataf transformation based point estimate method. Chinese Science Bulletin 53 (17), 2586–2592.

Liu, P. L., Kiureghian, A. D., 1986. Multivariate distribution models with prescribed marginals and covariances. Probabilistic Engineering Mechanics 1 (2), 105–112.

Mara, T. A., Tarantola, S., 2012. Variance-based sensitivity indices for models with dependent inputs. Reliability Engineering and System Safety 107, 115–121.

McKay, M. D., 1996. Variance-based methods for assessing uncertainty importance. Tech. rep., Technical Report NUREG-1150,UR-1996-2695, Los Alamos National Laboratory.

Norton, J., 2015. An introduction to sensitivity assessment of simulation models. Environmental Modelling and Software 69, 166–174.

Oakley, J. E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a bayesian approach. J. Royal Stat. Soc. B 66, 751–769.

OECD/NEA PSAC User Group, 1989. PSACOIN Level E intercomparison. An international code intercomparison exercise on a hypothetical safety assessment: Case study for radioactive waste disposal systems. Technical report, OECD-NEA, Paris.

OECD/NEA PSAG User Group, 1993. PSACOIN Level S intercomparison. An international code intercomparison exercise on a hypothetical safety assessment: Case study for radiowaste disposal. Technical report, OECD-NEA, Paris.

Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling and Software 67, 1–11.

Plischke, E., Borgonovo, E., Smith, C. L., 2013. Global sensitivity measures from given data. European Journal of Operational Research 226 (3), 536–550.

Ratto, M., Pagano, A., Young, P., 2007. State dependent parameter metamodelling and sensitivity analysis. Computer Physics Communications 117 (11), 863–876.

Rosenblatt, M., 1952. Remarks on the multivariate transformation. Annals of Mathematics and Statistics 43, 470–472.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Computational Physics Communications 145, 280–297.

Saltelli, A., Marivoet, J., 1990. Non-parametric statistics in sensitivity analysis for model output: a comparison of selected techniques. Reliability Engineering and System Safety 28, 229–253.

Saltelli, A., Tarantola, S., 2002. On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. Journal of the American Statistical Association 97, 702–709.

Sobol', I. M., 1993. Sensitivity estimates for nonlinear mathematical models. Math. Mod. and Comput. Exp. 1, 407–414.

Sobol', I. M., 2001. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation 55, 271–280.

Xu, C., 2013. Decoupling correlated and uncorrelated uncertainty contributions for nonlinear models. Applied Mathematical Modelling 37, 9950–9969.

Xu, C., Gertner, G. Z., 2008a. A general first-order global sensitivity analysis method. Reliability Engineering and System Safety 93, 1060–1071.

Xu, C., Gertner, G. Z., 2008b. Uncertainty and sensitivity analysis for models with correlated parameters. Reliability Engineering and System Safety 93, 1563–1573.

Appendix A Rosenblatt transformation

Let $\mathbf{x} = (x_1, \ldots, x_n) \sim p(\mathbf{x})$ be a set of continuous dependent random variables, with joint probability density function $p(\mathbf{x})$ that can be re-written as $p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_n|\mathbf{x}_{\sim n})$ where $\mathbf{x}_{\sim i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Let $F_i(x_i|\mathbf{v})$ be the cumulative distribution function of $p(x_i|\mathbf{v})$, with $\mathbf{v} \subseteq \mathbf{x}_{\sim i}$. The Rosenblatt transformation (Rosenblatt, 1952) of $\mathbf{x}$ provides with a set of independent random variables $\mathbf{u}^1$ uniformly distributed over the unit hypercube $\mathbb{K}^n = [0, 1]^n$. That is,

$$\begin{cases} u_1^1 = F_1(x_1) \\ u_2^1 = F_2(x_2|x_1) \\ \vdots \\ u_n^1 = F_n(x_n|\mathbf{x}_{\sim n}) \end{cases} \tag{36}$$

The Rosenblatt transformation is unique if and only if $\mathbf{x}$ is a set of independent variables, that is, $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)\ldots p(x_n)$. In this case, the ANOVA decomposition shown in Equation (2) (see the main text of the paper) is unique. In general, the Rosenblatt transformation is not unique and there are $n!$ possibilities depending on how the random variables are ordered in the set $\mathbf{x}$. We denote by $\mathbf{u}^i$ the Rosenblatt transform of the set $(x_i, \ldots, x_n, x_1, \ldots, x_{i-1})$ obtained after the $(i-1)^{th}$ circular permutation of the canonical set. Such transformations require the knowledge of the conditional cumulative distribution functions $F_i(x_i|\mathbf{v})$.

Rosenblatt transformations are usually employed to generate a set of dependent inputs distributed with respect to a given probability density function $p(\mathbf{x})$ from a set of independently and uniformly distributed variables $\mathbf{u}^1$. For

this purpose, the inverse Rosenblatt transform is employed,

$$\begin{cases} x_1 = F_1^{-1}(u_1^1) \\ x_2 = F_2^{-1}(u_2^1|x_1) \\ \vdots \\ x_n = F_n^{-1}(u_n|\mathbf{x}_{\sim n}) \end{cases} \tag{37}$$

Appendix B The integral definitions of the sensitivity indices

B.1 For the first-order sensitivity index

Let us denote $\mathbf{u} = (\mathbf{v}, \mathbf{w})$ one of the Rosenblatt transforms of $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. It comes that,

$$\begin{aligned} V\left[E\left[g(\mathbf{v},\mathbf{w})|\mathbf{v}\right]\right] &= \int_{\mathbb{K}^s} \mathrm{d}\mathbf{v} \left(\int_{\mathbb{K}^{n-s}} g(\mathbf{v},\bar{\mathbf{w}})\mathrm{d}\bar{\mathbf{w}}\right)^2 - \left(\int_{\mathbb{K}^n} g(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{u}\right)^2 \\ &= \int_{\mathbb{K}^s} \mathrm{d}\mathbf{v} \int_{\mathbb{K}^{n-s}} g(\mathbf{v},\bar{\mathbf{w}})\mathrm{d}\bar{\mathbf{w}} \int_{\mathbb{K}^{n-s}} g(\mathbf{v},\bar{\mathbf{w}}')\mathrm{d}\bar{\mathbf{w}'} \\ &\quad - \int_{\mathbb{K}^n} g(\mathbf{u})\mathrm{d}\mathbf{u} \int_{\mathbb{K}^n} g(\mathbf{u}')\mathrm{d}\mathbf{u}' \end{aligned}$$

which, by using Bayes rule writes,

$$V\left[E\left[g(\mathbf{v},\mathbf{w})|\mathbf{v}\right]\right] = \int_{\mathbb{K}^n} g(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w} \left(\int_{\mathbb{K}^{n-s}} g(\mathbf{v},\bar{\mathbf{w}}')\mathrm{d}\bar{\mathbf{w}}' - \int_{\mathbb{K}^n} g(\mathbf{v}',\mathbf{w}')\mathrm{d}\mathbf{v}'\mathrm{d}\mathbf{w}'\right). \tag{38}$$

Now, if the RT is such that,

$$\begin{aligned} \mathrm{d}\mathbf{v} &= p(\mathbf{y})\mathrm{d}\mathbf{y} \\ \mathrm{d}\mathbf{w} &= p(\bar{\mathbf{z}}|\mathbf{y})\mathrm{d}\bar{\mathbf{z}} \end{aligned}$$

we get $E\left[f(\mathbf{x})\right] = \int_{\mathbb{R}^n} f(\mathbf{y},\mathbf{z})p(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z} = \int_{\mathbb{K}^n} g(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}$

Changing the variables in (38) yields the integral definition of the numerator in Equation (13),

$$\begin{aligned} V\left[E\left[f(\mathbf{y},\mathbf{z})|\mathbf{y}\right]\right] = \int_{\mathbb{R}^n} & f(\mathbf{y},\mathbf{z})p(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z} \\ & \left(\int_{\mathbb{R}^{n-s}} f(\mathbf{y},\bar{\mathbf{z}}')p(\bar{\mathbf{z}}'|\mathbf{y})\mathrm{d}\bar{\mathbf{z}}' - \int_{\mathbb{R}^n} f(\mathbf{y}',\mathbf{z}')p(\mathbf{y}',\mathbf{z}')\mathrm{d}\mathbf{y}'\mathrm{d}\mathbf{z}'\right) \end{aligned}$$

(39)

But, if the RT is such that,

$$\begin{aligned} \mathrm{d}\mathbf{v} &= p(\bar{\mathbf{y}}|\mathbf{z})\mathrm{d}\bar{\mathbf{y}} \\ \mathrm{d}\mathbf{w} &= p(\mathbf{z})\mathrm{d}\mathbf{z} \end{aligned}$$

then, changing the variables in (38) yields the integral definition of the numerator in Equation (15),

$$V\left[E\left[f(\mathbf{y},\mathbf{z})|(\mathbf{y}|\mathbf{z})\right]\right] = \int_{\mathbb{R}^n} f(\bar{\mathbf{y}},\mathbf{z})p(\bar{\mathbf{y}}|\mathbf{z})p(\mathbf{z})\mathrm{d}\bar{\mathbf{y}}\mathrm{d}\mathbf{z} \left(\int_{\mathbb{R}^{n-s}} f(\bar{\mathbf{y}},\mathbf{z}')p(\mathbf{z}')\mathrm{d}\mathbf{z}' - \int_{\mathbb{R}^n} f(\mathbf{y}',\mathbf{z}')p(\mathbf{y}',\mathbf{z}')\mathrm{d}\mathbf{y}'\mathrm{d}\mathbf{z}'\right). \tag{40}$$

B.2 For the total sensitivity index

We start with the law of total variance,

$$E\left[V\left[g(\mathbf{u})|\mathbf{w}\right]\right] = V\left[g(\mathbf{u})\right] - V\left[E\left[g(\mathbf{u})|\mathbf{w}\right]\right]. \tag{41}$$

We can write,

$$V\left[g(\mathbf{u})\right] = \frac{1}{2}\int_{\mathbb{K}^n} g^2(\bar{\mathbf{v}},\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v} + \frac{1}{2}\int_{\mathbb{K}^n} g^2(\mathbf{v}',\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v}' - \left(E\left[g(\mathbf{u})\right]\right)^2$$

Besides, from (38), it can be inferred that,

$$V\left[E\left[g(\mathbf{v},\mathbf{w})|\mathbf{w}\right]\right] = \int_{\mathbb{K}^{n-s}} \mathrm{d}\mathbf{w}\int_{\mathbb{K}^s} g(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{v}\int_{\mathbb{K}^s} g(\mathbf{v}',\mathbf{w})\mathrm{d}\bar{\mathbf{v}}' - \left(E\left[g(\mathbf{u})\right]\right)^2$$

By replacing the two previous relations in (41) yields,

$$\begin{aligned} E\left[V\left[g(\mathbf{u})|\mathbf{w}\right]\right] &= \frac{1}{2}\int_{\mathbb{K}^n} g^2(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v} + \frac{1}{2}\int_{\mathbb{K}^n} g^2(\mathbf{v}',\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{v}' \\ &\quad - \int_{\mathbb{K}^{n-s}} \mathrm{d}\mathbf{w}\int_{\mathbb{K}^s} g(\mathbf{v},\mathbf{w})\mathrm{d}\mathbf{v}\int_{\mathbb{K}^s} g(\mathbf{v}',\mathbf{w})\mathrm{d}\mathbf{v}' \end{aligned}$$

which is equivalent to,

$$E\left[V\left[g(\mathbf{u})|\mathbf{w}\right]\right] = \frac{1}{2}\int_{\mathbb{K}^{n+s}} \left(g(\mathbf{v}',\mathbf{w}') - g(\mathbf{v},\mathbf{w}')\right)^2 \mathrm{d}\mathbf{v}'\mathrm{d}\mathbf{w}'\mathrm{d}\mathbf{v}. \tag{42}$$

As previously, if the RT is,

$$\begin{aligned} \mathrm{d}\mathbf{v} &= p(\mathbf{y})\mathrm{d}\mathbf{y} \\ \mathrm{d}\mathbf{w} &= p(\bar{\mathbf{z}}|\mathbf{y})\mathrm{d}\bar{\mathbf{z}} \end{aligned}$$

then, changing the variables in (42) yields the integral definition of the numerator in Equation (16),

$$E\left[V\left[f(\mathbf{z},\mathbf{y})|(\mathbf{z}|\mathbf{y})\right]\right] = \frac{1}{2}\int_{\mathbb{R}^{n+s}} \left(f(\mathbf{y}',\bar{\mathbf{z}}') - f(\mathbf{y},\bar{\mathbf{z}}')\right)^2 p(\bar{\mathbf{z}}'|\mathbf{y}')p(\mathbf{y}')p(\mathbf{y})\mathrm{d}\mathbf{y}'\mathrm{d}\bar{\mathbf{z}}'\mathrm{d}\mathbf{y} \tag{43}$$

But, if the RT is such that,

$$\begin{aligned} \mathrm{d}\mathbf{v} &= p(\bar{\mathbf{y}}|\mathbf{z})\mathrm{d}\bar{\mathbf{y}} \\ \mathrm{d}\mathbf{w} &= p(\mathbf{z})\mathrm{d}\mathbf{z} \end{aligned}$$

then, changing the variables in (42) yields the integral definition of the numerator in Equation (14),

$$E\left[V\left[f(\mathbf{y},\mathbf{z})|\mathbf{z}\right]\right] = \frac{1}{2}\int_{\mathbb{R}^{n+s}} \left(f(\mathbf{y}',\mathbf{z}') - f(\bar{\mathbf{y}},\mathbf{z}')\right)^2 p(\mathbf{y}',\mathbf{z}')p(\bar{\mathbf{y}}|\mathbf{z}')\mathrm{d}\mathbf{y}'\mathrm{d}\mathbf{z}'\mathrm{d}\bar{\mathbf{y}} \tag{44}$$

Appendix C RT of the variables in section 4.2

Let $(x_1, x_2) \in [0,1[^2$ be uniformly distributed over the triangle $x_1 + x_2 \leq 1$. The joint pdf is $p(x_1, x_2) = 2$ and the following pdfs can be obtained,

$$p_1(x_1) = \int_0^{1-x_1} p(x_1, x_2)\mathrm{d}x_2 = 2(1 - x_1) \tag{45}$$

$$p_{2|1}(x_2|x_1) = \frac{p(x_1, x_2)}{p_1(x_1)} = \frac{1}{(1 - x_1)} \tag{46}$$

The associated cumulative distribution functions are

$$f_1(x_1) = \int_0^{x_1} p_1(x)\mathrm{d}x = x_1(2-x_1) \tag{47}$$

$$f_{2|1}(x_2,x_1) = \int_0^{x_2} p_{2|1}(x|x_1)\mathrm{d}x = \frac{x_2}{1-x_1} \tag{48}$$

and the Rosenblatt transforms of (x_1, x_2) are,

$$\begin{cases} u_1^1 = x_1(2-x_1) \\ u_2^1 = \dfrac{x_2}{1-x_1} \end{cases} \tag{49}$$

This transformation being bijective from $[0,1[\times[0,1]$ to $[0,1[\times[0,1]$, we can invert the previous equations and find,

$$\begin{cases} x_1 = 1-\sqrt{1-u_1^1} \\ x_2 = u_2^1\sqrt{1-u_1^1} \end{cases} \tag{50}$$

These relationships allow to generate samples uniformly distributed over the triangle $x_1+x_2 \leq 1$ from samples uniformly distributed over the unit hypercube $(u_1^1, u_2^1) \in \mathbb{K}^2$ excluding $(u_1^1, u_2^1) = (1, \cdot)$.

In the same way, we show that the Rosenblatt transform of $(x_3, x_4) \in]0,1]^2$ uniformly distributed over the triangle $x_3 + x_4 \geq 1$, yields,

$$p_3(x_3) = \int_{1-x_3}^{1} p(x_3,x_4)\mathrm{d}x_4 = 2x_3 \tag{51}$$

$$p_{4|3}(x_4|x_3) = \frac{1}{x_3} \tag{52}$$

$$\begin{cases} u_3^1 = f_3(x_3) = x_3^2 \\ u_4^1 = f_{4|3}(x_4,x_3) = \dfrac{x_3+x_4-1}{x_3} \end{cases} \tag{53}$$

$$\begin{cases} x_3 = \sqrt{u_3^1} \\ x_4 = (u_4^1-1)\sqrt{u_3^1}+1 \end{cases} \tag{54}$$

Tables

List of Tables

Table 1: Analytical first-order sensitivity indices (S_i, ST_i^{ind}) for different correlation structures and their mean bootstrap estimates ($\hat{S}_i, \hat{ST}_i^{ind}$).

$(\rho_{12}, \rho_{13}, \rho_{23})$	Input	$S_i = ST_i$	$\hat{S}_i$	$S_i^{ind} = ST_i^{ind}$	$\hat{ST}_i^{ind}$
$(0.5, 0.8, 0)$	x_1	0.94	0.95	0.02	0.02
	x_2	0.40	0.38	0.05	0.05
	x_3	0.58	0.60	0.03	0.03
$(-0.5, 0.2, -0.7)$	x_1	0.49	0.49	0.72	0.70
	x_2	0.04	0.05	0.37	0.37
	x_3	0.25	0.26	0.48	0.50

Table 2: Inputs list for the Level E model

Notation	**Definition**	**Distribution**	**Range**	**Units**
T	Containment time	Uniform	$[100, 1000]$	yr
k_I	Leach rate for Iodine	Log-uniform	$[10^{-3}, 10^{-2}]$	$mols/yr$
k_c	Leach rate for Np chain	Log-uniform	$[10^{-6}, 10^{-5}]$	$mols/yr$
$v^{(1)}$	Water speed in geosphere's layer 1	Log-uniform	$[10^{-3}, 10^{-1}]$	m/yr
$l^{(1)}$	Length of geosphere's layer 1	Uniform	$[100, 500]$	m
$R^{(1)}$	Retention factor for I (first layer)	Uniform	$[1, 5]$	$-$
$R_C^{(1)}$	Retention coeff. for Np chain layer 1	Uniform	$[3, 30]$	$-$
$v^{(2)}$	Water speed in geosphere's layer 2	Log-uniform	$[10^{-2}, 10^{-1}]$	m/yr
$l^{(2)}$	Length of geosphere's layer 2	Uniform	$[50, 200]$	m
$R^{(2)}$	Retention factor for I (layer 2)	Uniform	$[1, 5]$	$-$
$R_C^{(2)}$	Retention coeff. for Np chain layer 2	Uniform	$[3, 30]$	$-$
W	Stream flow rate	Log-uniform	$[10^5, 10^7]$	m^2/yr

Table 3: Configuration for correlated input of the Level E model

pairs of correlated factors	correlation
k_I, k_C	0.5
$R^{(1)}, R_C^{(1)}$	0.3
$R^{(2)}, R_C^{(2)}$	0.3
$T, v^{(1)}$	- 0.7
$v^{(1)}, v^{(2)}$	0.5
$R^{(1)}, R^{(2)}$	0.5
$R_C^{(1)}, R_C^{(2)}$	0.5

Figures

List of Figures

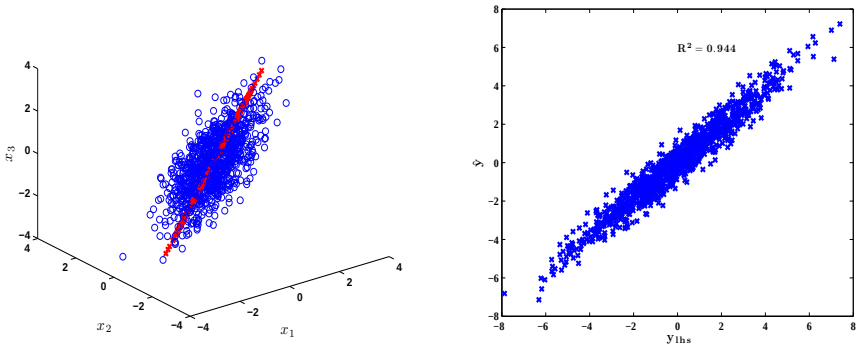

Figure 1: On the left, two samples of the random variables are depicted. The circles represent the pseudo-random sample an the red (line) crosses, sample generated from x_1 alone (see text for explanation). On the right, comparison of the model responses respectively evaluated with the original sample (y_{lhs}) and the sample generated from x_1 ($\hat{y}$). A good adequacy is observed between the responses, meaning that, given the correlation structure, the knowledge of x_1 alone is sufficient to assess the model response uncertainty.

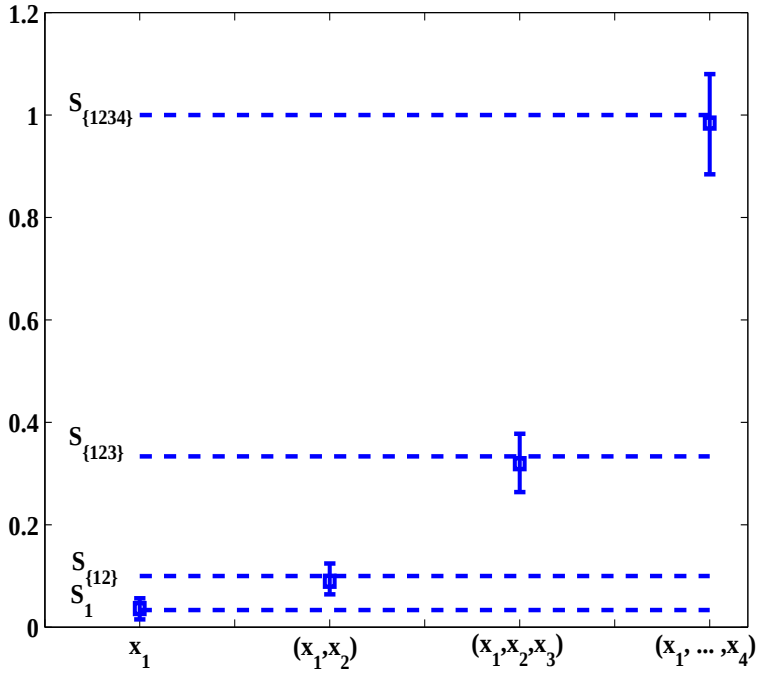

Figure 2: Bootstrap estimates of the sensitivity indices $S_{\mathbf{y}}$ for different groups of inputs $\mathbf{y}$. The dashed-lines are the analytical values. The squares are the mean bootstrap estimates while errorbars represent the intervals of variation.

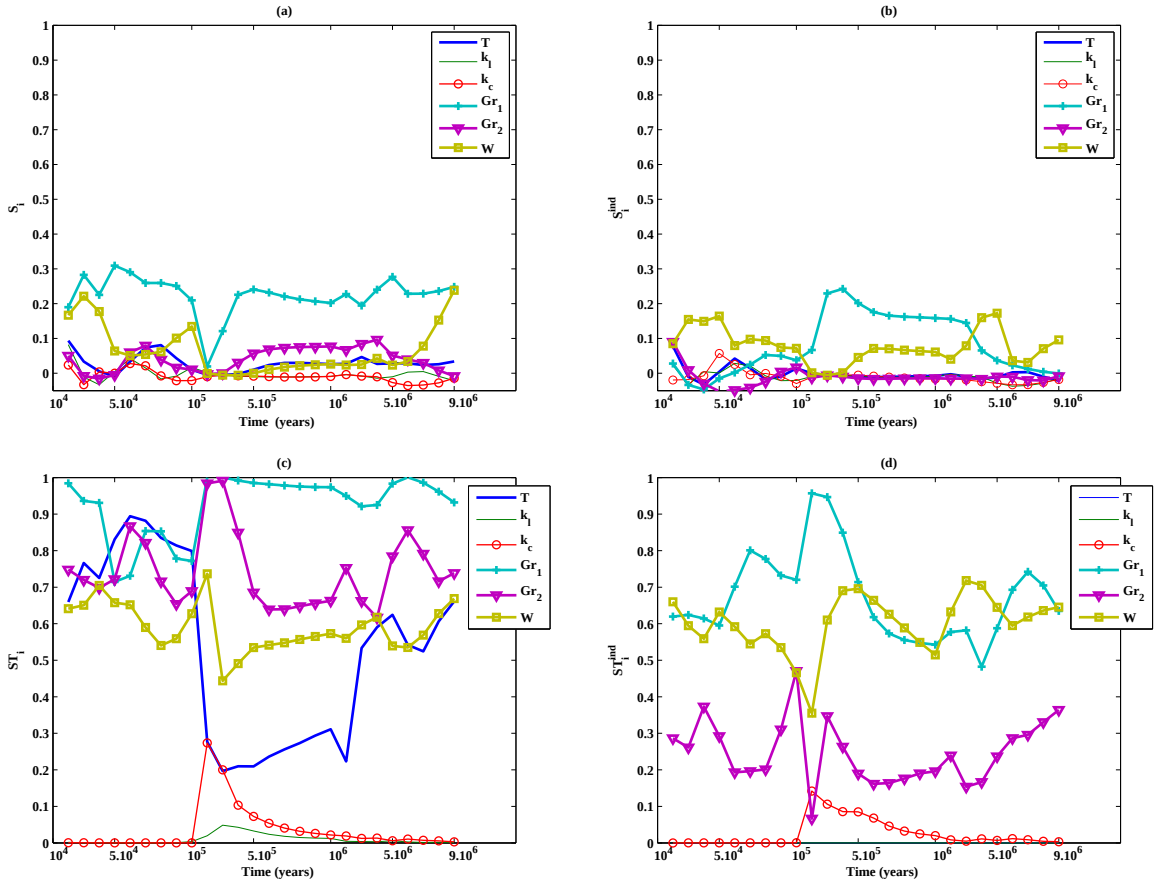

Figure 3: Level E estimated variance-based sensitivity indices: (a) full first-order indices, (b) independent first-order indices, (c) full total indices and (d) independent total indices. See text for explanations.