Environmental Modelling & Software 92 (2017) 40–56

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Position Paper

Metamodeling and global sensitivity analysis for computer models with correlated inputs: A practical approach tested with a 3D light interception computer model

J.-P. Gauchi [a, *], A. Bensadoun [a], F. Colas [b], N. Colbach [b]

[a] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[b] INRA, UMR1347 Agroécologie, 21065 Dijon, France

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form
29 December 2016
Accepted 30 December 2016

Keywords:
Agroecology
FlorSys
Correlated inputs
Metamodeling
Partial least squares regression
Polynomial chaos expansion
Sensitivity indices

ABSTRACT

Models of biophysical processes are often time-consuming and their inputs are frequently correlated. This situation of non-independence between the inputs is always a challenge in view of simultaneously achieving a global sensitivity analysis of the model output and a metamodeling of this output. In this paper, a novel practical method is proposed for reaching this two-fold goal. It is based on a truncated Polynomial Chaos Expansion of the output whose coefficients are estimated by Partial Least Squares Regression. The method is applied to a computer model for heterogeneous canopies in arable crops, aimed to predict crop:weed competition for light. We now have fast-running metamodels that simultaneously provide good approximations of the outputs of this computer model and a clear overview of its input influences thanks to new sensitivity indices.

1. Introduction

Many techniques exist today for metamodeling a computer model output (Gasca and Sauer, 2000; Bates et al., 2003; Rasmussen and Williams, 2006; Wang and Shan, 2007; Stanfill et al., 2015). On the other hand, several methods exist for defining and estimating the Sensitivity Indices (SI) of computer (nonlinear) model inputs on the computer model output, based on the variance of this output. According to the technique used, even an effective metamodel can lead to very wrong estimates of the SI because the main goal of a metamodel is generally not to provide estimates of the SI but, instead, for prediction purposes. Similarly, correct estimates of the SI can be obtained by Monte Carlo techniques but they cannot lead to an effective metamodel.

The mathematical definition of this SI type, based on the variance of the output, was given by Sobol' (Sobol', 1993; Lemieux, 2009), and is referred to as the Sobol' Sensitivity Indices (SSI). It is based on the Hoeffding-Sobol decomposition of the total functional variance of an output (Hoeffding, 1948; Sobol', 1993), i.e., a generalization for nonlinear models of the usual decomposition of the total variance for linear models. The estimation of these SSI leads to a Global Sensitivity Analysis (GSA) (Saltelli et al., 2000, 2004; Saltelli, 2002) of the output.

It therefore remains a difficult two-fold challenge to *simultaneously* obtain an effective metamodel and correct SI estimates of this type. In order to meet this two-fold challenge, methods based on a truncated Polynomial Chaos Expansion (PCE) of the response (Sudret, 2008; Crestaux et al., 2009; Blatman and Sudret, 2011) where the coefficients are estimated by Ordinary Least Squares Regression (OLSR), were proposed. However, these methods are relevant only if the random inputs of the computer model are continuous and independent because the SSI are rigorously defined only in this situation (Sobol', 1993). They are not mathematically founded in the case of correlated inputs because the Hoeffding-Sobol decomposition no longer holds in this case. We therefore propose new SI in this paper that are not based on the Hoeffding-Sobol decomposition, which are different from the SSI.

The need for metamodels is crucial today in many applications in several scientific areas, including the agronomical and ecological sciences (Colbach, 2010; Marie and Simioni, 2014) because

* Corresponding author.
E-mail address: jean-pierre.gauchi@inra.fr (J.-P. Gauchi).

http://dx.doi.org/10.1016/j.envsoft.2016.12.005
1364-8152/

computer models very often take too much computing time to run, whereas an adapted metamodel takes a short time to run. It is true that several methods already exist where the inputs are correlated (Jacques et al., 2006; Li and Rabitz, 2012; Mara and Tarantola, 2012; Kucherenko et al., 2012), but these methods are not always convenient to use or applicable to our agronomical concerns. Below, we give a single disadvantage of each of these methods in order to understand why they are not really convenient and easy to use.

The method proposed by Jacques et al. (2006) leads to SI that are not decomposable to first-order effects and interaction effects. The method proposed by Li and Rabitz (2012) is based on very heavy mathematical tools (tensor product spline bases) and is consequently poorly adapted to many inputs (more than five or six). Moreover, their method does not lead to a single functional decomposition because the latter particularly depends on the number and choice of some approximating functions. The method proposed by Mara and Tarantola (2012) is based on a first step of decorrelation of the correlated inputs by means of a classical Gram-Schmidt orthogonalization (that lead to orthogonalized inputs). In a second step, relevant SI can then be obtained but these SI are interpretable only via the orthogonalized inputs and not via the natural inputs, which represents an obvious disadvantage. The method proposed by Kucherenko et al. (2012) is very heavy because it is based on the generation of conditional densities of Gaussian inputs (the case of uniformly distributed inputs is not mentioned) via the sophisticated copula techniques. This leads to a considerable number of samplings and, furthermore, no clear meaningful SI are obtained for separating first-order and total effects.

We propose a simpler and more practical alternative method here where the continuous inputs are correlated. This method simultaneously provides sensitivity indices of a new kind, as well as a metamodel. It is based on a truncated PCE of the response whose coefficients are estimated by Partial Least Squares Regression (PLSR), whereas Sudret (2008) used OLSR. This method is particularly well-adapted when the continuous inputs - in moderate number (typically ≤ 15) - are stochastically linked (correlated) or even deterministically (functionally) linked, on the one hand, and when a single computer run is moderately time-consuming (typically less than one minute), on the other. These input numbers and time-consuming values obviously depend on the type of computer used. They are given for a Pentium IV desk computer (with a clock speed of about 3 GHz) equipped with a 12-giga RAM. More details are given on this subject in the Discussion section.

In this paper, this method is applied to a biophysical computer model in the field of agroecology. Models that describe and predict biophysical processes that occur in the field are needed for agroecological crop management, but often require a significant number of inputs and are time-consuming (Lô-Pelzer et al., 2010; Vos et al., 2010; Colbach et al., 2014). If the inputs are considered as independent, a first approximation is used to make simulations faster by replacing these models with emulators or parsimonious metamodels that depend only on the most important inputs (Colbach, 2010; Marie and Simioni, 2014). This method was applied to a 3D individual-based light interception model (Munier-Jolain et al., 2013) whose aim was to predict crop:weed competition for light in heterogeneous canopies. This model is a central component of the multi-annual weed dynamics model, FlorSys, aimed at testing agroecological cropping systems (Colbach et al., 2014). A crucial difference in the present paper is that it considers non-negligible and even strong correlations between the inputs.

The rest of the article is organized as follows. Section 2 is devoted to a persuasive illustration of the influence of the input correlation on the sensitivity indices obtained by PCE (and OLSR) of the response, for two well-known academic models used as test functions in GSA (the so-called Ishigami and Sobol' functions). Section 3 presents our new method. Section 4 is devoted to an application to the two preceding academic models. Section 5 is devoted to the application to a case study with a process-based light interception model, revealing the effectiveness of this new approach. Section 6 contains the discussion and conclusions. Section 7 gives details about software/data availability. An appendix provides a list of the numerous abbreviations used in our paper.

2. Influence of the input correlation

This study on the influence of input correlation was a motivation for proposing new sensitivity indices adapted to the management of correlated inputs present in computer models (e.g., biophysical models), as well as to innovation using metamodeling techniques. Sobol' defined the First Order Sobol' Sensitivity Indices, referred to in this paper as the FOSSI, and the Total Sobol' Sensitivity Indices, referred to as the TSSI (Sobol', 1993; Lemieux, 2009). These FOSSI and TSSI are estimated by a classical method based on a truncated PCE whose coefficients are computed by OLSR (Sudret, 2008). Note that the inputs must be independent for the mathematical validity of the FOSSI and TSSI, as well as that of their estimations: the PC_d-PESI(OLS) and PC_d-TSI(OLS), defined at the end of Subsection 3.1, where d is the degree of the truncated PCE.

In this section, we only provide a simple illustration, obtained by a simulation study, of the influence of the correlations between the inputs on the value of these PC_d-PESI(OLS) and PC_d-TSI(OLS), for two very well-known academic models in the GSA domain: the Ishigami function (Saltelli et al., 2000; Chap. 2) and the Sobol' function (Sobol', 2003). The advantages of using these two functions are two-fold: (a) They are strongly nonlinear (this is the reason why the FOSSI and the TSSI are so different from each other; their analytical values are compared below), and it is therefore a challenge to obtain good respective estimations; and (b) The quality of any estimation method can always be evaluated because the FOSSI and TSSI *analytical values are known* (Saltelli et al., 2000) for these two functions.

The Ishigami function has three inputs $\mathbf{X} = (X_1, X_2, X_3)$ that are linked to the output Y according to:

$$Y = \sin(X_1) + \theta_1[\sin(X_2)]^2 + \theta_2 X_3^4 \sin(X_1) \tag{1}$$

where $\theta_1 = 7$, and $\theta_2 = 0.1$, given in Ishigami and Homma (1990). Each X_j is a uniform random variable on the interval $[-\pi; \pi]$. The analytical values of the FOSSI for the independent X_1, X_2 and X_3 are 0.3138, 0.4424 and 0, respectively, and the analytical values of the TSSI for X_1, X_2 and X_3 are 0.5574, 0.4424 and 0.2436, respectively.

The Sobol' function has eight inputs that are linked to the output Y according to:

$$Y = \prod_{j=1}^{8} \frac{|4X_j - 2| + a_j}{1 + a_j} \tag{2}$$

where $\mathbf{a} = (1, 2, 5, 10, 20, 50, 100, 500)$, given in Sudret (2008). Each X_j is a uniform random variable on the interval $[0; 1]$. Since the last four $FOSSI_j$ and $TSSI_j$, $j = 5, \ldots, 8$, are close to zero, we consider only the first four inputs, $X_j, j = 1, \ldots, 4$, in this paper, whereas $X_j, j = 5, \ldots, 8$ were set to the value of 1/2 (i.e., their mean value). The analytical values of the FOSSI for the independent X_1, X_2, X_3 and X_4 are then 0.6037, 0.2683, 0.0671 and 0.0200, respectively, and the analytical values of the TSSI for X_1, X_2, X_3 and X_4 are then 0.6342, 0.2945, 0.0756 and 0.0227, respectively.

For both functions, a simulation study made it possible to perceive the influence of correlation between the inputs according