Computers & Geosciences 37 (2011) 917–927

Contents lists available at ScienceDirect

Computers & Geosciences

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

Global sensitivity analysis of large-scale numerical landslide models based on Gaussian-Process meta-modeling

Jeremy Rohmer *, Evelyne Foerster

BRGM, Natural Risks and Safety of CO_2 Geological Storage, 3, av. Claude Guillemin BP 36009, 45060 Orléans Cedex 2, France

ARTICLE INFO

Article history:
Received 15 December 2009
Received in revised form
3 December 2010
Accepted 26 February 2011
Available online 16 March 2011

Keywords:
Landslide model
Computationally demanding code
Global sensitivity analysis
Sobol' indices
Meta-model
Gaussian process

ABSTRACT

Large-scale landslide prediction is typically based on numerical modeling, with computer codes generally involving a large number of input parameters. Addressing the influence of each of them on the final result and providing a ranking procedure may be useful for risk management purposes. This can be performed by a variance-based global sensitivity analysis. Nevertheless, such an analysis requires a large number of computer code simulations, which appears impracticable for computationally demanding simulations, with computation times ranging from several hours to several days. To overcome this difficulty, we propose a "meta-model"-based strategy consisting in replacing the complex simulator by a "statistical approximation" provided by a Gaussian-process (GP) model. This allows computation of sensitivity measures from a limited number of simulations. For illustrative purposes, the proposed methodology is used to rank in terms of importance the properties of the elastoplastic model describing the complex behavior of the slip surface in the La Frasse landslide (Switzerland). One limitation of the GP-based methodology is that the computation of sensitivity measures is associated with uncertainty as the simulator is approximated using a training sample of small size, i.e., a limited knowledge on the "true" simulator. This source of uncertainty can be taken into account by treating the GP model from a Bayesian perspective. This provides the full posterior probability distribution associated with the sensitivity measures, which can be summarized by a confidence interval to outline the regions where the GP model is "unsure." We show that this methodology is able to provide useful guidelines for the practical decision-making process and suggest further site investigations.

1. Introduction

Landslides are very complex phenomena controlled by a range of processes. Geological history, lithology and structure, slope relief and shape, weather and climate, seismicity, and human activity can be identified as the main causative factors (Crosta and Clague, 2009). The associated risk to communities can be high (e.g., Evans et al., 2002) and thus, predicting landslide behavior is a major concern.

Due to the recent advances in computer modeling (e.g., in processor performance) and in particular in the finite element method (e.g., van den Ham et al., 2009), numerical models are commonly used in practice to gain a better understanding of the landslide behavior and to predict its evolution. The main drawback of such models is the high number of input factors required for analysis. Global sensitivity analysis of complex numerical models can then be used to determine: (1) which input factors contribute the most to the output variability (within the "factors' prioritization setting" as described by Saltelli et al., 2008); (2) which input factors interact with each other; and (3) which input factors are insignificant and can be eliminated to "simplify" the model (within a "factors' fixing setting") (Saltelli et al., 2008). Such an analysis is useful in identifying which input factors require further investigations to reduce uncertainties in the computer code results, hence providing guidelines for risk management (Saltelli, 2002b).

Among the existing sensitivity methods, variance-based methods have proved to be effective (Saltelli et al., 2000). In this paper, we focus on the method of Sobol' indices (Sobol', 1993; Archer et al., 1997; Sobol' and Kucherenko, 2005). Unlike traditional linear or rank regression-based methods, these indices allow representing the sensitivity of a general model without assuming any kind of linearity or monotonicity in the model (Saltelli and Sobol', 1995). In practice, the computation of Sobol' indices uses a Monte Carlo sampling strategy. An example of application in the field of landslide modeling with applications of moderate complexity is provided by Hamm et al. (2006). Such an approach, however, appears hardly applicable for more computationally demanding models, as it requires a large number of computer code evaluations. For instance, the study of Hamm et al. (2006) required 10,000 model realizations, corresponding to about 20 h of computation time (on a 2 GHz Pentium 4 PC).

* Corresponding author. Tel.: +33 2 38 64 30 92; fax: +33 2 38 64 36 89.
E-mail addresses: j.rohmer@brgm.fr (J. Rohmer), e.foerster@brgm.fr (E. Foerster).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.02.020

The same sensitivity analysis would require 208 days using a model that takes 30 min and 2500 days using a model that takes 6 h to compute.

To overcome this difficulty, a first solution is to use a distributed parallel computing methodology, thus requiring an appropriate grid computing architecture and the optimization of computing resources (e.g., Dupros et al., 2006; Boulahya et al., 2007).

In this paper, an alternative is proposed using a limited number of computer code runs (also named "simulator") (O'Hagan, 2006), which consists in replacing (i.e., approximating) the simulator by a surrogate model with low computation time, also named a "meta-model," to compute the Sobol' indices (i.e., the sensitivity measures). Various meta-models exist (e.g., linear regression, nearest neighbor method, multivariate adaptative regression spline, neural network, and Gaussian process); see, for example, Storlie et al. (2009) for a recent review.

The meta-model uses a limited number of simulator runs, i.e., input–output pairs (corresponding to the training sample), to infer the values of the complex simulator output given a yet-unseen input configuration. Such an approximation introduces a source of uncertainty referred to as "code uncertainty" associated with the meta-model (O'Hagan, 2006), so that the sensitivity measures computed with the meta-model are "uncertain."

In the present paper, we choose to solve the described problem of approximation (and of inference) under the Bayesian formalism treating the simulator as an "unknown" function in the sense that the simulator output for any yet-unseen input configuration is unknown until the simulator is actually run for the considered configuration (Oakley and O'Hagan, 2004). We choose to use the concept of an emulator corresponding to a statistical approximation so that a prior probability distribution is assigned to the simulator outputs and updated according to the usual Bayesian paradigm given the training sample. This approach returns not only the most likely value for the output given by any input configuration but also an entire probability distribution (O'Hagan, 2006). This distribution can be used to estimate a level of confidence when the predictive quality of the meta-model is not high due to a small training data (see, for instance, Marrel et al., 2008, 2009; Storlie et al., 2009). A Gaussian process (GP) is chosen as the prior model for the simulator. It has been widely used when designing computer experiments (Sacks et al., 1989; Kennedy and O'Hagan, 2001; Santner et al., 2003).

In the first section, the Sobol' decomposition method is described in the general framework of the variance-based global sensitivity approach (Saltelli et al., 2008).

Then, the GP model used as a meta-model of the computationally intensive simulator is described in the framework of the stochastic processes for computer code experiments under the Bayesian regression formalism. The methodology for computing the Sobol' indices using the GP model is described and illustrated in two applications. The first application is a simple analytical model based on "infinite slope analysis" (Hansen, 1984). This allows us to compare the sensitivity measures computed using the "true" model with those computed using the GP model. Finally, the application of this methodology to a La Frasse (Switzerland) landslide model (Laloui et al., 2004) is presented and we show how to use the sensitivity measures to guide the decision-making process for further site investigations.

2. Global sensitivity analysis by the Sobol' decomposition method

2.1. Introduction on the variance-based sensitivity analysis

Consider the simulator g and the scalar output y determined from a vector of n input factors $\mathbf{x}=\{x_i\}_{i=1,\ldots,n}$ so that $y=g(\mathbf{x})$.

Considering the n-dimensional vector as a random vector of independent random variable X_i, then the output Y is also a random variable (as a function of a random vector). A variance-based sensitivity analysis aims at determining the part of the total unconditional variance V_Y of the output Y resulting from each input random variable X_i. The total variance V_Y can be expressed as (Saltelli et al., 2000, 2008)

$$V_Y=\sum_i V_i+\sum_{i<j} V_{ij}+\sum_{i<j<l} V_{ijl}+\cdots+V_{ij\ldots n}, \tag{1}$$

where the partial variance V_i and $V_{ij\ldots n}$ read

$$\begin{aligned} V_i &= Var[E[Y|X_i=x_i]],\\ V_{ij\ldots n} &= Var[E[Y|X_i=x_i,X_j=x_j,\ldots,X_n=x_n]]\\ &\quad -V_i-V_j-\cdots-V_n \end{aligned} \tag{2}$$

with $E[Y|X_i=x_i]$, the expectation of Y given that the ith input factor X_i has a fixed value x_i and $E[Y|X_i=x_i,X_j=x_j,\ldots,X_n=x_n]$ the conditional expectation of Y given that the ith input factor X_i has a fixed value x_i, the jth input factor X_j has a fixed value x_j, etc.

The variance of the conditional expectation V_i represents the first-order effect of the input factor X_i taken alone, whereas the higher order indices account for possible mixed influence of various input factors.

2.2. The Sobol' decomposition method

2.2.1. Presentation

To determine the partial variances of Y, Sobol' (1993) proposes the following decomposition of g into summands of increasing dimension provided that g is integrable:

$$g(\mathbf{x})=g_0+\sum_{i=1}^{n} g_i(x_i)+\sum_{i\le j}^{n} g_{ij}(x_i,x_j)+\cdots+g_{1,\ldots,n}(x_1,\ldots,x_n). \tag{3}$$

Here g_0 corresponds to the mean constant value of the function g and each term can be evaluated through multidimensional integrals as

$$g_0=E[Y]=\int_0^1\cdots\int_0^1 g(\mathbf{x})\,d\mathbf{x}=\int_0^1\cdots\int_0^1 g(x_1,\ldots,x_n)\,dx_1\ldots dx_n \tag{4}$$

$$g_i(x_i)=E[Y|X_i=x_i]-g_0=\int_0^1\cdots\int_0^1 g(\mathbf{x})\,d\mathbf{x}_{-i}-g_0 \tag{5}$$

$$\begin{aligned} g_{ij}(x_i,x_j) &= E[Y|X_i=x_i,X_j=x_j]-g_0-g_i-g_j\\ &= \int_0^1\cdots\int_0^1 g(\mathbf{x})\,d\mathbf{x}_{-(ij)}-g_0-g_i(x_i)-g_j(x_j), \end{aligned} \tag{6}$$

where $d\mathbf{x}_{-i}$ denotes the integration over all input factors except x_i and $d\mathbf{x}_{-(ij)}$, the integration over all input factors except both x_i and x_j. Similar formulae can be obtained for higher order terms.

The total variance V_Y can then be expressed as

$$V_Y=\int_0^1\cdots\int_0^1 g(x_1,\ldots,x_n)^2\,dx_1,\ldots dx_n-g_0^2 \tag{7}$$

while the partial variances read as follows:

$$V_{i_1\ldots i_s}=V[g_{i_1\ldots i_s}]=\int_0^1\cdots\int_0^1 g_{i_1\ldots i_s}(x_{i_1},\ldots,x_{i_s})^2\,dx_{i_1}\ldots dx_{i_s} \tag{8}$$

with $1\le i_1<\cdots<i_s\le n$ and $s=1,\ldots,n$.

2.2.2. Definition of the Sobol' indices

The Sobol' indices $S_{i_1\ldots i_S}$ describe which amount of the total variance is due to the uncertainties of input factors in the set $\{i_1,\ldots,i_s\}$ and is expressed as the ratio between $V_{i_1\ldots i_s}$ and V_Y, respectively, the partial and total variances.