

# Estimating the approximation error when fixing unessential factors in global sensitivity analysis

I.M. Sobol'<sup>a</sup>, S. Tarantola<sup>b</sup>, D. Gatelli<sup>b,\*</sup>, S.S. Kucherenko<sup>c</sup>, W. Mauntz<sup>d</sup>

<sup>a</sup>*Institute for Mathematical Modelling of the Russian Academy of Sciences, Moscow, Russia*

<sup>b</sup>*Joint Research Centre of the European Commission, TP361, Institute of the Protection and Security of the Citizen, Via E. Fermi 1, 21020 Ispra (VA), Italy*

<sup>c</sup>*Imperial College London, UK*

<sup>d</sup>*Department of Biochemical and Chemical Engineering, Dortmund University, Germany*

Received 20 April 2006; received in revised form 5 July 2006; accepted 6 July 2006

Available online 28 August 2006

## Abstract

One of the major settings of global sensitivity analysis is that of fixing non-influential factors, in order to reduce the dimensionality of a model. However, this is often done without knowing the magnitude of the approximation error being produced. This paper presents a new theorem for the estimation of the average approximation error generated when fixing a group of non-influential factors. A simple function where analytical solutions are available is used to illustrate the theorem. The numerical estimation of small sensitivity indices is discussed.

© 2006 Elsevier Ltd. All rights reserved.

*Keywords:* Sobol' method; Approximation error; Sensitivity analysis by groups; Sensitivity indices

## 1. Introduction

This work is related to global sensitivity analysis based on the use of ANOVA decomposition and global sensitivity indices (see [1,6–8] for theory, and [3,4] for applications). Definitions of the sensitivity indices can be found in Section 2.

Different settings for sensitivity analysis are available, depending on the modeler's needs. One of these is the factors fixing setting. It is used for identifying non-influential factors in the model (those factors that can be fixed at any value in their domains without significantly reducing the output variance). A limit with factor fixing is that of fixing unessential factors without knowing the magnitude of the approximation error that is being produced. In Section 2, we prove one new theorem which quantifies this approximation error of the model output when one factor or a group of factors is fixed. So, once we

know from total indices that a factor is unessential, we will also have an estimate of the error that is generated by fixing it.

In this paper we study a model function  $f(x_1, \dots, x_n)$ , where the factors  $x_1, \dots, x_n$  are non-random independent scaled variables:  $0 \leq x_1 \leq 1, \dots, 0 \leq x_n \leq 1$ . Thus the point  $x = (x_1, \dots, x_n)$  is defined in the  $n$ -dimensional unit hypercube with Lebesgue measure. Clearly the factors  $x_1, \dots, x_n$  can be regarded as independent random variables uniformly distributed in the unit interval [0,1]. In this case the quantities that are called variances are real variances of certain random variables.

The sensitivity analysis based on ANOVA decomposition and global sensitivity indices can be easily (mutatis mutandis) generalized to independent random factors  $x_1, \dots, x_n$  with arbitrary distribution functions  $F_1(x_1), \dots, F_n(x_n)$  (e.g., [8]). However, the requirement of independence is important.

Section 2 contains a new theorem, Section 3—an illustration of the theorem, and in Section 4 numerical estimation of small sensitivity indices is discussed.

\*Corresponding author. Tel.: +39 0332 789928; fax: +39 0332 785733.

E-mail addresses: [stefano.tarantola@jrc.it](mailto:stefano.tarantola@jrc.it) (S. Tarantola), [debora.gatelli@jrc.it](mailto:debora.gatelli@jrc.it) (D. Gatelli).

### 2. The proposed theorem

Let  $x = (x_1, \dots, x_n)$  be a point in the  $n$ -dimensional unit hypercube with Lebesgue measure. We denote by  $y$  an arbitrary subset consisting of  $s$  variables and let  $z$  be the set of  $n-s$  complementary variables,  $1 \leq s \leq n-1$ . Thus  $x \equiv (y, z)$  and  $dx = dy dz$ . All the integrals below are from 0 to 1 in each variable. The set of variables  $z$  can be regarded as non-essential if the sensitivity index  $S_z^{\text{tot}} \ll 1$ . The common practice in such a situation is to fix somehow a value  $z_0$  and to use  $f(y, z_0)$  as an approximation to  $f(x)$ . The approximation error depends on the choice of  $z_0$ :

$$\delta(z_0) = \frac{1}{D} \int [f(x) - f(y, z_0)]^2 dx, \tag{1}$$

where  $D$  is the variance of  $f(x)$ :  $D = \int f^2 dx - f_0^2$ ,  $f_0 = \int f dx$ . The model function is assumed to be square integrable.

The following theorem shows that  $\delta(z_0)$  is of the same order as  $S_z^{\text{tot}}$ .

**Theorem.** For an arbitrary  $z_0$  the error  $\delta(z_0) \geq S_z^{\text{tot}}$ . If  $z_0$  is assumed to be random and uniformly distributed, then the expected value is  $E\delta(z_0) = 2S_z^{\text{tot}}$ .

A corollary of the theorem is the following assertion from [5]: for an arbitrary  $\varepsilon > 0$  with probability exceeding  $1-\varepsilon$

$$\delta(z_0) < \left(1 + \frac{1}{\varepsilon}\right) S_z^{\text{tot}}.$$

In particular (at  $\varepsilon = 0.5$ ), the inequality  $\delta(z_0) < 3S_z^{\text{tot}}$  holds with probability exceeding 0.50.

**Proof.** The ANOVA decomposition of  $f(x)$  can be written in the form

$$f(x) = f_0 + g_1(y) + g_2(z) + g_{12}(x), \tag{2}$$

where  $g_1(y)$  is the sum of all terms that depend on  $y$  variables only and similarly  $g_2(z)$  is the sum of all terms that depend on  $z$  only;  $g_{12}$  is the remainder.

From the definition of ANOVA, one can see that  $\int g_1 dy = \int g_2 dz = \int g_{12} dy = \int g_{12} dz = 0$ .

Consider the variances  $D_y = \int g_1^2 dy$ ,  $D_z = \int g_2^2 dz$ ,  $D_{yz} = \int g_{12}^2 dx$ .

Squaring (2) and integrating over  $dx$  we obtain the relation  $D = D_y + D_z + D_{yz}$  that allows a direct definition of the sensitivity indices for the sets  $y$  and  $z$ :

$$S_z = \frac{D_z}{D}, \quad S_z^{\text{tot}} = \frac{D_z + D_{yz}}{D}, \quad S_y = \frac{D_y}{D}, \quad S_y^{\text{tot}} = \frac{D_y + D_{yz}}{D}.$$

From these definitions one can see that  $S_z^{\text{tot}} = 1 - S_y$ ,  $S_y^{\text{tot}} = 1 - S_z$ .

Now an expression for  $\delta(z_0)$  can be derived:

$$\begin{aligned} \delta(z_0) &= \frac{1}{D} \int [g_2(z) + g_{12}(x) - g_2(z_0) - g_{12}(y, z_0)]^2 dx \\ &= \frac{1}{D} \int [g_2^2(z) + g_{12}^2(x) + g_2^2(z_0) + g_{12}^2(y, z_0)] dx \\ &= \frac{1}{D} \left[ D_z + D_{yz} + g_2^2(z_0) + \int g_{12}^2(y, z_0) dy \right]. \end{aligned}$$

The final result is  $\delta(z_0) = S_z^{\text{tot}} + (1/D)[g_2^2(z_0) + \int g_{12}^2(y, z_0) dy]$ .

Both assertions of the theorem follow immediately:  $\delta(z_0) \geq S_z^{\text{tot}}$  and  $\int \delta(z_0) dz_0 = 2S_z^{\text{tot}}$ .

**Proof of the Corollary.** Consider a non-negative random variable  $\eta = \delta(z_0)/S_z^{\text{tot}} - 1$ . Clearly,  $E\eta = 1$ . A well-known Chebyshev inequality for non-negative random variables with finite expectation can be applied: for an arbitrary  $h > 0$  the probability  $P\{\eta \geq h\} \leq E\eta/h$ .

We put  $\varepsilon = 1/h$  and turn to the opposite event:  $P\{\eta < 1/\varepsilon\} > 1 - \varepsilon$ .

The last relation is equivalent to the assertion of the corollary.

### 3. Analytic example: the $g$ -function

We illustrate the theorem by using the  $g$ -function of Sobol', which is often used as a benchmark for sensitivity analysis exercises (see e.g., [2]) as the exact analytical values can be easily calculated. The function is defined as

$$f = \prod_{i=1}^n g_i(x_i), \tag{3}$$

where  $n$  is the number of independent input factors and  $g_i(x_i)$  is

$$g_i(x_i) = \frac{|4x_i - 2| + a_i}{1 + a_i}, \tag{4}$$

for  $0 \leq x_i \leq 1$  and  $a_i \geq 0$ .

The parameter  $a_i$  is set to determine the importance of the input factor  $x_i$ , given that the range of variation of  $g_i(x_i)$  depends exclusively on the value of  $a_i$ . If  $a_i = 0$ , the corresponding factor  $x_i$  is important; if  $a_i = 1$ ,  $x_i$  is relatively important, while for  $a_i = 9$  it becomes non-important and for  $a_i = 99$  non-significant.

For the function (3) the first-order partial variances are  $D_i = 1/3(1 + a_i)^2$ , the higher order partial variances are products  $D_{i_1 \dots i_s} = D_{i_1} \dots D_{i_s}$ , and the total variance  $D = \prod_{i=1}^n (D_i + 1) - 1$ .

The group variances  $D_y$ ,  $D_z$ ,  $D_y^{\text{tot}} = D_y + D_{yz}$ ,  $D_z^{\text{tot}} = D_z + D_{yz}$  are sums of partial variances. However, integral representations for these variances allow direct numerical computation of their values [6,8].

**Test 1.** We consider a model with eight input factors, where  $a_i = \{0, 1, 4.5, 9, 99, 99, 99, 99\}$ ,

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات