Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 2 (2010) 7615–7616

Procedia Social and Behavioral Sciences

Sixth International Conference on Sensitivity Analysis of Model Output

Guidelines for optimal estimation of correlation ratios based on Monte Carlo simulation of computer codes

R. Bolado[a]*, E. Plischke[b], S. Tarantola[c] and A. Badea[a]

[a]*Institute for Energy, EC, DG-JRC, Petten, The Netherlands*
[b]*Institut für Endlagerforschung, Technische Universität Clausthal, 38678 Clausthal-Zellerfeldt, Germany*
[c]*Intitute for the Protection and Security of the Citizen, EC, DG-JRC, Ispra, Italy*

Abstract

The computational cost of many computer codes is a burden that obliges users to use the same set of simulations to perform uncertainty and sensitivity analysis. Correlation ratios are a simple and straightforward tool to estimate first order sensitivity indices. Nevertheless, they demand the use of debiasing factors and of appropriate partitions of the support of each input parameter to deliver optimal estimates. The way to estimate these indices depends in fact on the actual values to be estimated. This work contains an exhaustive study developed for providing a set of guidelines about the optimal way to deliver such estimates, which involves the minimum sample sizes needed, the selection of the partition and the use of appropriate debiasing factors.

Keywords: Correlation ratios; first order sensitivity indices; unbiased estimation.

1. Main text

Computer codes used in many technical and scientific areas are usually expensive in terms of computing time. This fact obliges many computer code users to use the same set of Monte Carlo simulations to perform uncertainty and sensitivity analyses, avoiding expensive sensitivity analysis techniques that demand specific sampling techniques and large total sample sizes. Correlation ratios are a simple method to estimate first order sensitivity indices according to Sobol's High Dimensional Model Representation (HDMR). Given a model whose output variable is Y, the first order sensitivity index corresponding to a given input variable X is

$$\eta^2 = Var\left[E(Y|X)\right]/Var(Y), \tag{1}$$

The most widespread estimator for this quantity given a regression model (y) is

$$\hat{\eta}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{2}$$

* Corresponding author. Tel.: +31 224 565349; fax: +31 224 565641.
E-mail address: Ricardo.bolado-lavin@jrc.nl.

doi:10.1016/j.sbspro.2010.05.144

Creating a partition of the support of X, each interval containing the same number of sampled values (n_s), the average in each interval is used as the estimated value of Y in that interval ($\bar{y}_j$). Then, the estimator turns

$$\hat{\eta}^2 = \sum_{j=1}^{k} n_s \left(\bar{y}_j - \bar{y}\right)^2 \Big/ \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2, \tag{3}$$

where k is the number of intervals. Three main problems have to be addressed when estimating first order sensitivity indices via correlation ratios: 1) the bias of the estimator, 2) the dependence of the partition needed on the actual value of the sensitivity index and 3) the minimum sample size needed. The first problem was identified and addressed long ago by Pearson (1915) and Kelley (1935), who proposed different alternatives as debiasing factors. A simplification of both debiasing factors is the following one, which needs only simple considerations about the loss of degrees of freedom imposed in the estimation process, with k denoting the number of partitions,

$$\hat{\eta}_u^2 = \left(n\hat{\eta}^2 - k\right)\Big/\left(n - k\right) \tag{4}$$

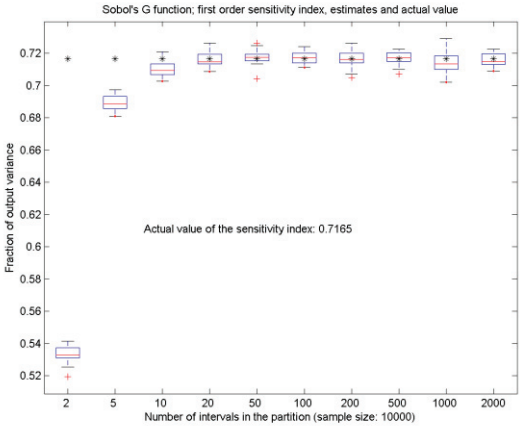

Figure 1.- First order sensitivity index estimated via formula (3); actual value: 0.7165; results obtained for 25 replicates of size 10000.

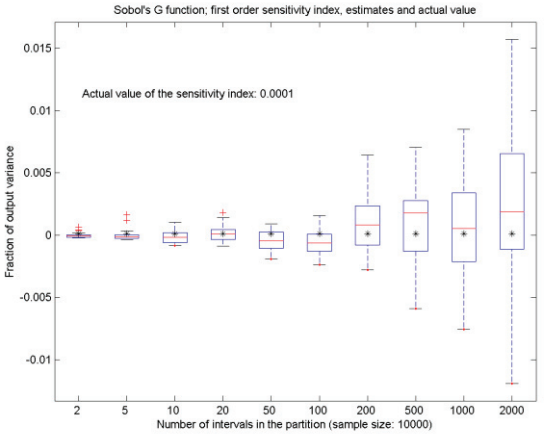

Figure 2.- First order sensitivity index estimated via formula (3); actual value: 0.0001; results obtained for 25 replicates of size 10000.

Figures 1 and 2 indicate that the debiasing factor introduced in formula (3) produces estimates with almost no bias when estimating very small sensitivity indices, whatever the number of intervals in the partition, while for large sensitivity indices this is true only if the number of intervals is above 20 (for this sample size and this model). These examples do also illustrate the problem of using the right partition in the estimation process; an intermediate number of intervals would be preferred for estimating large values while 2 intervals would deliver the most accurate estimate of very small values.

In this work we develop further the solution to the first problem and address the second problem by studying the intra-sample and inter-samples trends followed by the estimators. The mean squared error is used to discriminate between different alternatives (partitions). Results obtained so far indicate a strong dependence of estimates quality on the sample size available, a problem which is also addressed. As a result of all the work developed, clear guidelines are provided regarding the selection of debiasing factors, partitions and minimum sample sizes.

2. References

Pearson K., 1915. On the correction necessary for the correlation ratio η. *Biometrika*, 14: 492-498.
Kelley T.L., 1935. An unbiased correlation ratio measure. *Proc. Natl. Acad. Sci. USA*, 21(9):554-559.