# Lessons learnt from studies on sensitivity analysis techniques in the EU project PAMINA: A benchmark study

K.-J. Röhlig & E. Plischke
*Technische Universität Clausthal, Clausthal-Zellerfeld, Germany*
R. Bolado Lavín
*Institute for Energy, European Commission DG-JRC, Petten, The Netherlands*
D.-A. Becker
*GRS-Braunschweig, Brunswick, Germany*
P.-A. Ekström
*Facilia AB, Stockholm, Sweden*
S. Hotzel
*GRS-Köln, Cologne, Germany*

ABSTRACT: Several EU PAMINA project partners have applied different Sensitivity Analysis (SA) techniques to analytical models as well as to a simplified, though representative, PA model. The aims of the exercise were to investigate the performance of different techniques in the presence of features such as (lack of) linearity, (lack of) monotony, interactions, etc.; to check the importance of sample size; to compare different options within a given SA technique; to cross-compare the results obtained using different techniques; and to get a better understanding of the rationale behind every available technique and about their capabilities and shortcomings. This paper presents and discusses some results on the tests of global sensitivity analysis performed with variance-based sensitivity indicators.

## 1 INTRODUCTION

The EU Integrated Project PAMINA is devoting a large effort into the research of Sensitivity Analysis (SA) as a key element in the Performance Assessment (PA) of Radioactive Waste Repositories.

Among the activities was a benchmark of SA techniques designed as a two-step process. The first step is dedicated to analysing a set of mathematical functions for which most of the sensitivity measures are well known. The targets in this step are to debug SA computational tools used, to get skills in their use and to get progressively in contact with specific features of mathematical models such as (lack of) linearity, (lack of) monotony, interactions, etc., and to check the importance of sample size. The second step consists in analysing a simplified, though representative, PA model. The complex input-output relation, characterised by strong interactions among input parameters, makes it a challenging model to test SA techniques. In this case, the target is twofold: firstly to compare different options within a given SA technique (to study the added value of using more complex versions of a given technique – e.g. classical Fourier Amplitude Sensitivity Test FAST versus extended FAST, first order regressions versus higher order regressions, etc.), and secondly to cross-compare the results obtained using different techniques. The overall aim of the exercise is to get a better understanding of the rationale behind every available technique and about their capabilities and shortcomings. Recommendations concerning the calculation effort and case-specific restrictions were to be derived. This paper describes important parts of the exercise and presents selected results together with the main conclusions and recommendations. More details including a full description of the exercise and the results obtained can be found in Plischke and Röhlig (2009).

The exercise was part of a PAMINA task which also included a study investigating SA applications to realistic performance assessment models for several High-level Radioactive Waste Repositories which is described in Bolado et al. (2009).

## 2 DETAILS OF THE BENCHMARK STUDY

The plan of the benchmark study was issued in Plischke (2008), gathering the results of a meeting in which the Project partners agreed on a set of benchmark cases most of which were selected from Saltelli et al. (2000) and Saltelli et al. (2004). Since the exact results for the presented models are available in print, one can easily see if a method for the estimation of a sensitivity index works as expected.

The choice of the sensitivity analysis methods and implementations were left to the participants. Contributions were received from Andra (France),

Facilia (Sweden), JRC-Petten (The Netherlands), and TU Clausthal (Germany). A diverse range of available algorithms was in use, starting from linear regression over variance-based global sensitivity analysis to screening methods and statistical tests for performing Monte-Carlo Filtering. In order to unify the results and to draw more attention to the variance-based SA indicators a prescribed setting was specified for certain models to be analysed in a second simulation round, of which selected results are presented in this paper. For each model, 25 runs of 100, 300, 1000, 3000, and 10000 samples sizes were requested and for each run the indicators mean, variance, $R^2$, rank-based $R^{2*}$, and the variance-based sensitivity indicators (SI) first order effects and total effects (where available) were computed. The choice of the SI algorithms was left to the participants of this second round. Details of the theory behind the different sensitivity analysis algorithms are presented in Badea and Bolado (2008) and in the books Saltelli et al. (2000), Saltelli et al. (2004), Saltelli et al. (2008).

For the second round, contributions were received from Facilia (FCL, Sweden), GRS Cologne (Germany), JRC-Petten (The Netherlands), and TU Clausthal (TUC, Germany).

An example gathering some of the problems encountered in the analytical benchmarks is the PSACOIN Level E model, Nuclear Energy Agency (1989). As an optional element for the participants, a sensitivity study of the Level-E geosphere transport model was requested. Here Facilia and TUC provided results.

## 3 SA TECHNIQUES USED IN THE STUDY

As we already noted the choice of the SA methods was left to the participants. In this section we concentrate on the methods used in the second phase of the benchmark where the SA techniques were restricted to variance-based sensitivity analysis methods. The sensitivity index of first order effect is given by

$$S_i = \frac{V[E(Y|X_i)]}{V[Y]} = 1 - \frac{E[V(Y|X_i)]}{V[Y]}.$$

For total effects $S_{Ti} = 1 - S_{\sim i}$ analogous formulas apply where in $S_{\sim i}$ the condition "given $X_i$" is replaced by "given all but $X_i$". We can classify the methods for the calculation of the first order and/or total effects into four different groups.

- Correlation ratio methods. Here the conditional variance $E(Y|X_i = x)$ in the formula of the sensitivity indices is replaced by $E(Y|X_i \in I_m)$ for a suitable partition $\{I_m,\ m = 1, \dots, \ell\}$ of the range of the $i$th input parameter. To compute higher-order effects or total effects these methods suffer from the curse of dimensionality. But despite all drawbacks, these methods can work with given data; hence model evaluations available prior to the SA (e.g. from Monte Carlo simulations performed in order to carry out uncertainty analyses, i.e. to obtain statistics for the model output) can be reused. Then the choice of different partitions influences the result of such a method.
- Sobol' methods. For these methods, a special sampling scheme ensures that there are enough realisations available so that statistics for $E(Y|X_i = x)$ can be efficiently computed. A simple scheme is named after Ishigami/Homma/Saltelli (IHS), another scheme using a hyperconvergent quasi-Monte-Carlo sampling scheme is named after Sobol'. These methods can estimate first order and total effects.
- Fourier-based techniques. For these methods, the input parameter realisations have to fulfil special frequency properties. Then a frequency decomposition of the output maps different frequencies attributed to the input factors to different fractions of the variance of the output. Different frequency selection schemes have been developed, named Fourier Amplitude Sensitivity Test (FAST), Extended FAST (EFAST), and Random Balance Design (RBD). They can estimate first and/or total effects.
- Other "cheap" methods. Methods working with pre-computed model evaluations are computationally efficient. As straight-forward extensions of a linear regression method a polynomial fit of the data and a conditional linear fit have been tested. Furthermore, in the course of the benchmark an algorithm named EASI ("Effective Algorithm for variance-based Sensitivity Indices") has been developed that couples given model evaluations with Fourier-based sensitivity analysis techniques, see Plischke (2009).

## 4 MODELS CONSIDERED IN THE STUDY

The four analytical models used in the second round were chosen in such a way that a sensitivity analysis based upon linear regression analysis was bound to fail. These models are non-linear, non-monotonic, discontinuous, multi-parametric, or with input dependencies.

Moreover, a time-dependent model which has some of these properties is the PSACOIN Level-E

geosphere transport model. Some results on the SA of this complex model are also reported.

### 4.1 *The Ishigami test function*

The Ishigami test function is a three parameter model. It is in so far interesting as the second and third input factors have a Pearson Correlation Coefficient of zero. A variance-based analysis retrieves a 44% first order effect for the second input factor, but a zero effect for the third factor. Only when estimating total effects the third input factor is attributed 24% of the variance. The Ishigami function is given by

$$Y = \sin X_1 + 7 \sin^2 X_2 + 0.1 X_3^4 \sin X_1$$

where $X_i \sim U(-\pi, \pi)$ are uniformly distributed in $(-\pi, \pi)$.

The values of $R^2 \approx R^{2*} \approx 0.2$ imply that the results obtained via a standard or rank-transformed linear regression are not very powerful. Hence an analysis using other SA methods is needed.

### 4.2 *A discontinuous switch*

Discontinuities pose major numerical problems if the SA algorithm requires a smooth model.

A drastic change in the output behaviour over small variations of the input parameters is not unusual for real-world models and therefore needs further studying. Hence we analyse the following test function

$$Y = \begin{cases} -X_2, & \text{if } X_1 \le 0.5, \\ X_2, & \text{if } X_1 > 0.5, \end{cases} \qquad X_i \sim U(0,1).$$

We expect $S_1 = \frac{3}{4}$, $S_2 = 0$, $S_{T1} = 1$, and $S_{T2} = \frac{1}{4}$ as results of the sensitivity analysis.

### 4.3 *A linear model with input dependencies*

In theory, independent input parameters are required for performing variance-based SA. It is not clear what happens with the SA algorithms in the presence of dependencies between the input parameters or to what extent the results can be interpreted.

This example highlights some of the problems encountered when processing dependent data. The function under inspection is given by the linear model $Y = X_1 + X_2$ where the input parameters have a joint probability density function given by

$$p(x_1, x_2) = \begin{cases} 2, & \text{if } 0 \le x_1, x_1 \le 0.5, \\ 2, & \text{if } 0.5 \le x_1, x_1 \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

The expected values for the sensitivity indices are $S_i = \frac{13}{14}$, hence $S_{Ti} = \frac{1}{14}$, $i = 1,2$.

### 4.4 *The Sobol g test function*

Real-world models have many input parameters. Hence a test case where many input parameters are considered shows if an algorithm is robust enough to deal with such problems. A well-studied test function is the non-monotonic Sobol' g-function. Here we use its 8 parameter version which is given by

$$Y = \prod_{i=1}^{k} \frac{|4X_i - 2| + a_i}{1 + a_i}, \qquad X_i \sim U(0,1).$$

where $k = 8$ and $a = (0, 1, 4.5, 9, 99, 99, 99, 99)$. The first parameter is most influential, the influence decreases through the rest of the parameters until parameters five to eight become equally non-influential.

Due to the symmetry in the formula, we have $R^2 = R^{2*} = 0$. Hence the results based on linear regression are of no value for the sensitivity analysis.

### 4.5 *The Level-E geosphere transport model*

This computational model calculates the annual radiological dose to humans over geological time scales due to the underground migration of radionuclides from a hypothetical nuclear waste disposal site through a system of idealised natural and engineered barriers.

The Level-E code has already been used intensively for sensitivity analysis. Hence the results of a SA of this model are well-documented. It is therefore a good starting-point for an analysis of a complex model with dynamic output variables.

## 5 RESULTS

We now present and discuss some of the results of the benchmark. The benchmark exercise features all of the techniques applied to each of the examples, but in the scope of this document we only can present a representative subset. Most of the graphics are shown in form of box plots derived from the 25 available runs per sample size. The box plots show the lower quartile, the median, and the upper quartile values. The whiskers in the plots are lines showing the data range. Outliers are detected using a multiple (here: 3) of the inter-quartile range.

For the Ishigami test function, Figure 1 shows the results of eight different correlation ratio methods

for sample sizes 100, 300, 1000, 3000, and 10000 analysing the influence of the first input parameter on the output. Some of these methods study different algorithmic approaches (Variance of the Conditional Expectation VCE, Expectation of the Conditional Variance ECV), others the influence of different sampling schemes (Simple Random Sampling SRS, Latin Hypercube Sampling LHS, Latin Hypercube Sampling using conditional Median values LHS-M). A third group investigates the influence of different subsample strategies: A scheme chosen to resemble the rule-of-thumb $\ell = \lfloor \sqrt{n} \rfloor$ as close as possible from the provided data (CR) as well as schemes using a two-interval partition (CR2P) and schemes requesting a partition constructed in a way that five realisations are located in each interval $I_m$(CR5S) were studied ($\ell$: number of intervals, $n$: sample size). Using the rule-of-thumb, the partition consists of approximately $\ell$ elements with $\ell$ realisations in each element of the partition. The two methods which do not use this rule-of-thumb, CR2P and CR5S, produce non-consistent estimates.

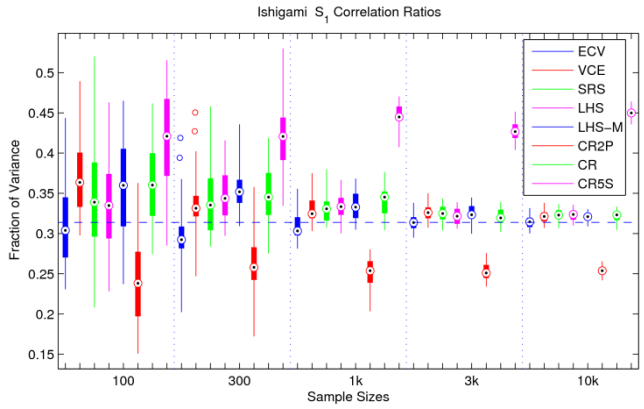


Figure 1: CR Methods for the Ishigami function, $\boldsymbol{S_1}$.

In Figures 2 and 3 this procedure is repeated for $S_2$ and for $S_3$. Most of these methods produce consistent estimates for $S_2$, with only little or no noticeable bias. The largest errors are produced by CR2P and CR5S. For example, for the estimation of $S_2$, CR2P has no advantage over a linear regression, and gives also a zero value.

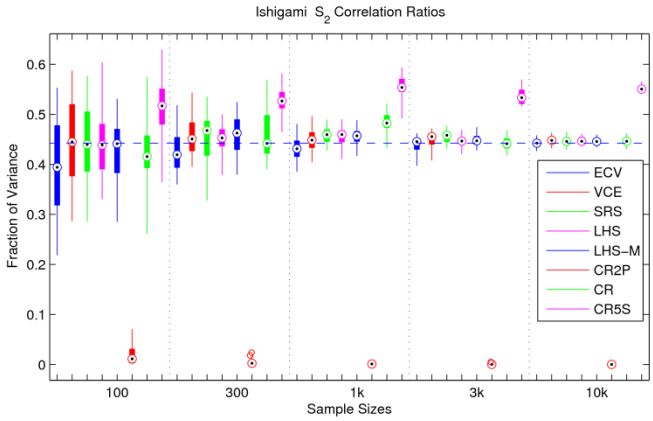


Figure 2: CR Methods for the Ishigami function, $\boldsymbol{S_2}$.

The estimation of true zero values via correlation ratio methods is difficult, only ECV and CR2P produce unbiased results for $S_3$, see Figure 3.

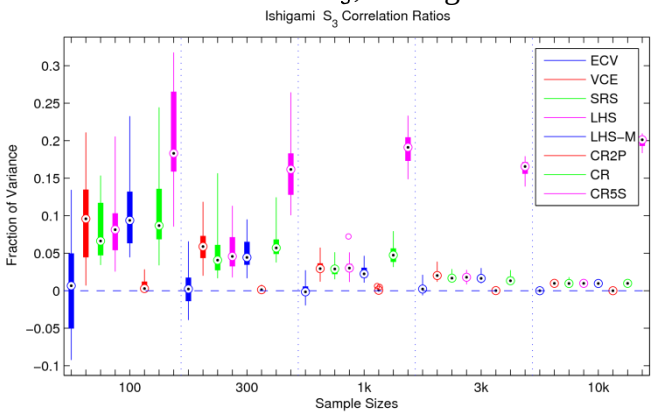


Figure 3: CR Methods for the Ishigami function, $\boldsymbol{S_3}$.

Figure 4 shows the estimates for $S_2$ using Fourier-based methods. Compared to Figure 2, the variances are in the same range but now the first four methods seem to be biased for small sample sizes, and EFAST(TUC) is not converging. This last behaviour can be explained as EFAST(TUC) is a simple implementation using no advanced frequency selection schemes. Indeed, we find the same problems in EFAST(FCL) with the small sample size of 100. Here, the TUC versions of FAST and EFAST fail as there are not enough realisations available. Note that the RBD methods offer no advantages when compared to the cheap EASI methods.

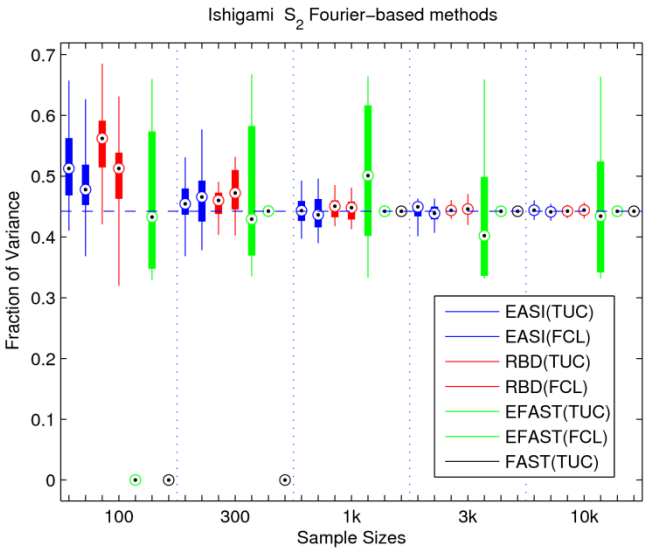


Figure 4: Fourier Methods for the Ishigami function, $\boldsymbol{S_2}$.

The results for the discontinuous switch example are illustrated in Figures 5 and 6. In Figure 5 the performance of different implementations of the Ishigami-Homma-Saltelli (IHS) method is compared. All estimators are unbiased, but they show different variations. In contrast, the variation encountered in Figure 6 for Fourier-based methods is much smaller, but the estimators are biased. Moreover, the convergence to the real value is slow for EFAST(TUC), as its maximal harmonic frequency is depending on the

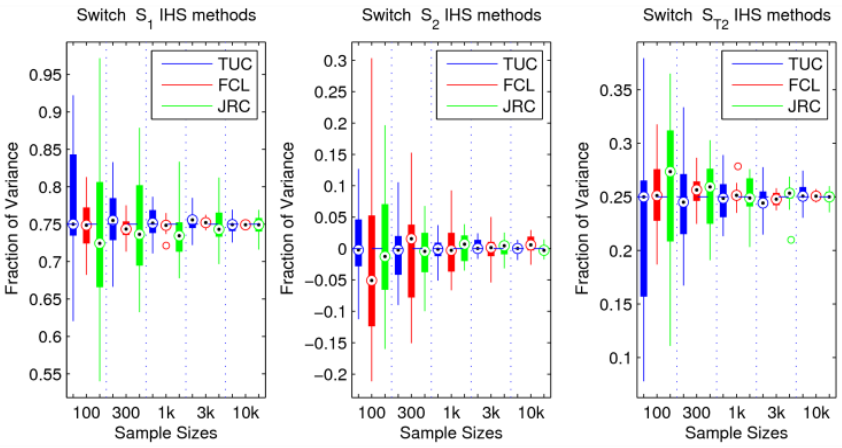


Figure 5: IHS Methods for the switch function, first order and total effects.

sample size, all other implementations have a fixed maximal harmonic frequency so that they give wrong estimates for this discontinuous example. Implementations with the same number of maximal harmonic frequency produce the same inconsistent estimate.

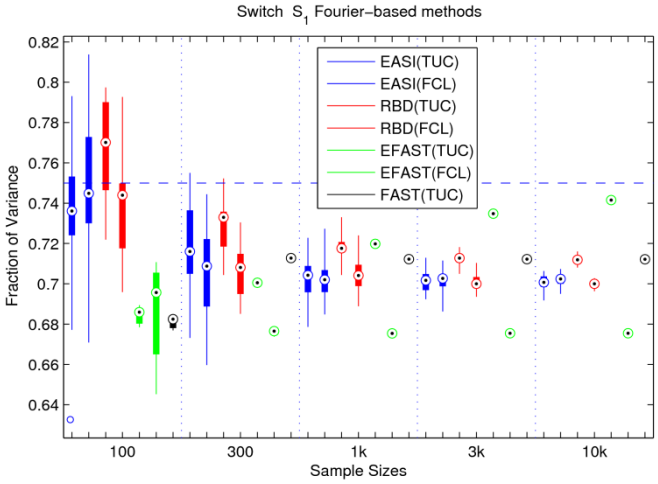


Figure 6: Fourier Methods for the switch function, $\boldsymbol{S_1}$.

Figures 7 and 8 show some of the results for the dependent input data model. Theoretically, the results for $S_1$ and $S_2$ should be the same, however only the Facilia implementations capture the right values for $S_2$, when special input sampling schemes are in use (e.g., for IHS, Sobol' and (E)FAST).

Not shown are the results from the GRS-Cologne correlation ratio implementations which also use different sampling schemes and take care of the joint input parameter distribution. These implementations also produce estimates close to the true values.

Since the model is a linear one, the estimators converge well even for small sample sizes.

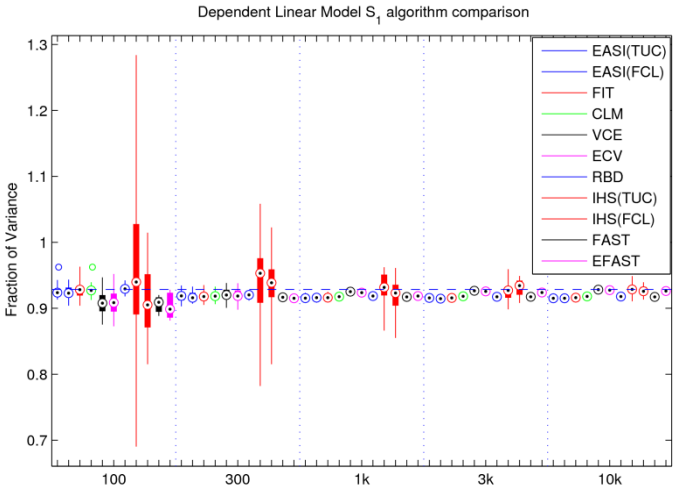


Figure 7: Overview of methods for the dependent model, $\boldsymbol{S_1}$.

The results for the Sobol' g function offer no additional information. Nearly all techniques perform well, only the EFAST(TUC) method with a simple

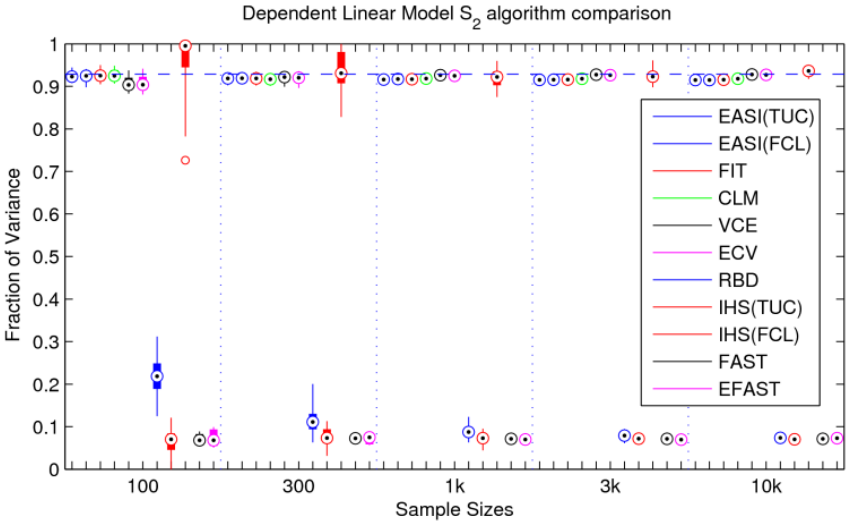


Figure 8: Overview of methods for the dependent model, $\boldsymbol{S_2}$.

frequency selection scheme does not converge.

From the experience gained during the analytical benchmark cases TUC decided to use a cheap method based on simple random sampling and the Sobol' method for the analysis of the Level-E model, Facilia performed calculations with the methods EASI, EFAST, RBD, and IHS.

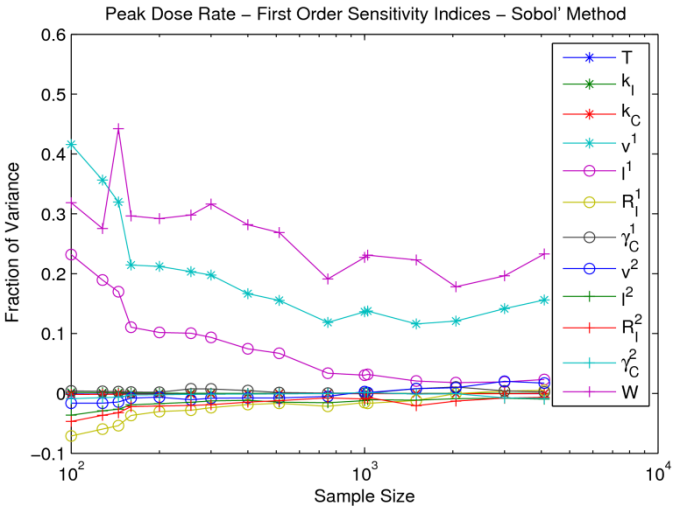


Figure 9: First order effects for the peak dose rate.

In the following, we analyse the sensitivity of the peak dose rate. Figures 9 and 10 show the results of the Sobol' calculations for first order effects and total effects which depend upon the basic sample size ranging from 100 to 4096. The most influential parameters $W$ and $v_1$ are identified even for small sample sizes. We experienced, however, considerable oscillations, which are not completely visible in the Figure due to the chosen resolution. Moreover, no convergence is apparent and $v_2$ produces large negative values for the total effects.

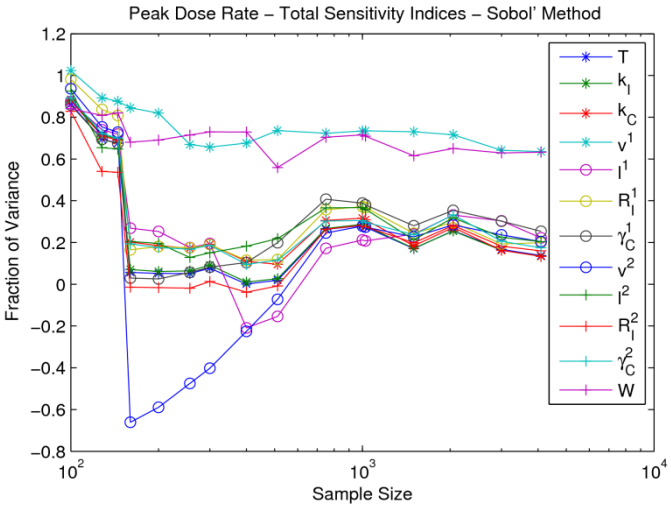


Figure 10: Total effects for the peak dose rate.

For comparison, Figure 11 shows the statistics of selected results obtained by Facilia for a sample size of up to 1000 realisations. Again, IHS shows a wide variance even for large sample sizes. It should also be noted that the IHS results differ from the outcomes of the other algorithms which are rather close to each other. Accounting for the experience that IHS results are less biased than others in a number of analytical cases one might however conclude that the IHS results are more reliable.

Figure 12 shows the Facilia results for the total effects. As for the Sobol' method, we encounter problems with the IHS method. The total effects from the parameter $v_1$ show a large negative outlier for sample size 300, and from the analysis of the parameters $v_2$ and $W$ we encounter large negative values. The results of EFAST look more promising: Their variance is small compared to the IHS methods and they seem to converge for $v_1$, $v_2$ and $W$, while the results for $R_I^1$ show sudden changes between sample sizes 300 and 1000.

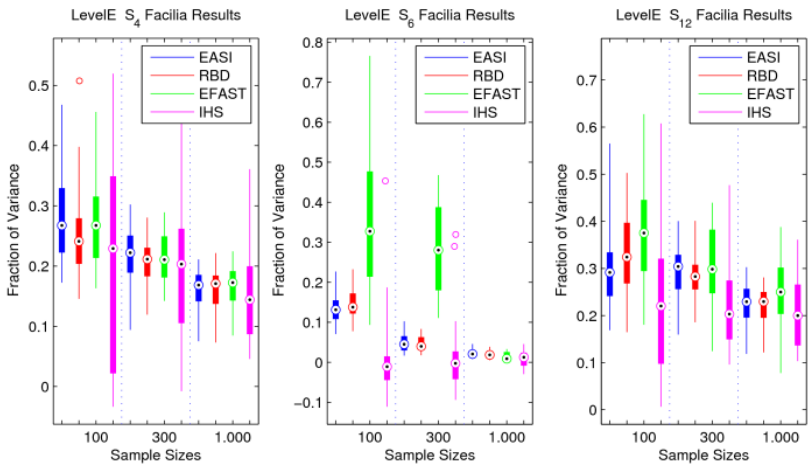


Figure 11: First order effects for $v_1$, $R_I^1$ and $W$, Facilia results.

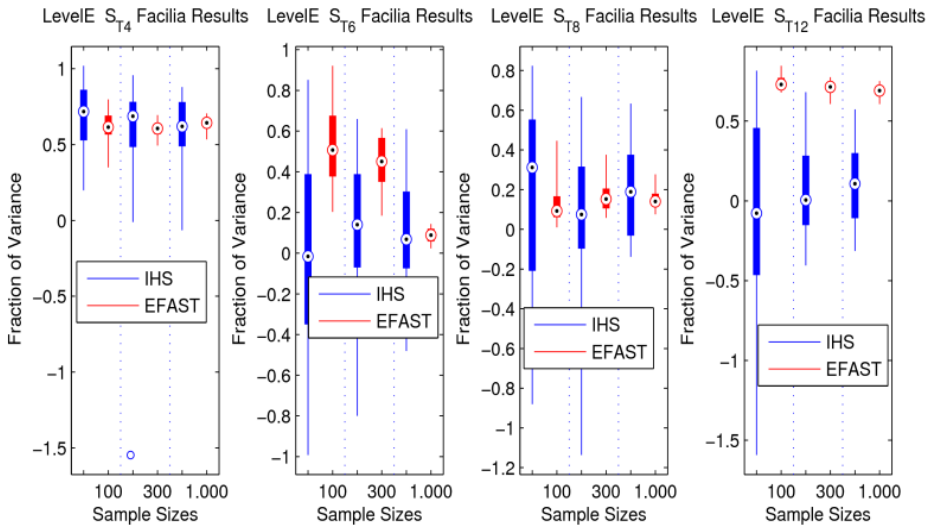


Figure 12: Total effects for $\boldsymbol{v_1}$, $\boldsymbol{R_I^1}$, $\boldsymbol{v_2}$ and $\boldsymbol{W}$, Facilia results.

## 6 LESSONS LEARNT

A lot of insight into the internals of variance-based sensitivity analysis has been gained during the course of this benchmark exercise. First of all, we noticed that for the standard algorithms the different implementations seem to be very stable and produce results with only subtle differences. In some situations, however, the results depend quite substantially on the implementation and/or the choices of the user.

The interest in cheap methods has arisen as of late in the benchmark exercise so that we were delighted that the results obtained with them are comparable to specialised methods.

However, there are some pitfalls which should be kept in mind when performing a variance-based SA.

- Sobol'/IHS without special Monte-Carlo-integration sequence performs worse than a cheap method.
- For a SA of a model with dependent inputs with methods requiring special sampling schemes care must be taken that the sampling scheme also satisfies the input distribution.
- Algorithms with fixed maximal harmonic or fixed number of intervals per partition may not capture discontinuities and may produce systematic errors by under- or over-estimating the sensitivity indices.
- Random Balance Design shows no advantages when compared with a cheap method like EASI.
- For small sensitivity indices nearly all methods show bad convergence properties. Here, IHS and Sobol' methods are positive exceptions to the rule.

## REFERENCES


Badea, A. and Bolado, R. 2008. Review of Sensitivity Analysis Methods and Experience. *Milestone 2.1.D.4, PAMINA project, Sixth Framework Programme of the EU.*

Bolado, R. et al. 2009. Lessons learnt from studies on sensitivity analysis techniques in the EU project PAMINA: Sensitivity analysis applied to different HLW PA models. *ESREL 2009.*

Nuclear Energy Agency 1989, *PSACOIN Level E Intercomparison*. Paris: OECD

Plischke, E. 2008. Plan for benchmark, including specification of synthetic PA cases. *Milestone 2.1.D.3, PAMINA project, Sixth Framework Programme of the EU.*

Plischke, E. 2009. An effective algorithm for variance-based sensitivity indices (EASI). *ESREL 2009.*

Plischke, E. and Röhlig, K.-J. 2009. Sensitivity Analysis Benchmark Based on the Use of Synthetic PA Cases (Topic Report). *Milestone 2.1.D.11, PAMINA project, Sixth Framework Programme of the EU.*

Saltelli, A. et al. (eds.) 2000. *Sensitivity Analysis*. Chichester: Wiley.

Saltelli, A. et al. 2004. *Sensitivity Analysis in Practice*. Chichester: Wiley.

Saltelli, A. et al. 2008. *Global Sensitivity Analysis*. Chichester: Wiley.