Computational Statistics & Data Analysis 15 (1993) 211–238
North-Holland

Sensitivity analysis of model output

An investigation of new techniques

A. Saltelli

Comission of the European Communities, Joint Research Centre, Italy

T.H. Andres

Atomic Energy of Canada Ltd., Pinawa Man, Canada

T. Homma

Japan Atomic Energy Research Institute of Tokai-Mura, Naka-gun Ibaraki-ken, Japan

Received March 1991
Revised July 1991

Abstract: Sensitivity Analysis (SA) of model output investigates the relationship between the predictions of a model, possibly implemented in a computer program, and its input parameters. Such an analysis is relevant for a number of practices, including quality assurance of models and codes, and the identification of crucial regions in the parameter space. This note compares established techniques with variants, such as a modified version of the Hora and Iman importance measure (SANDIA Laboratory Report SAND85-2839, 1989), or new methods, such as the iterated fractional factorial design (Andres, Hajas, Report in prep. for AECL, Pinawa, Canada, 1991). Comparison is made on the basis of method reproducibility and of method accuracy. The former is a measure of how well SA predictions are replicated when repeating the analysis on different samples taken from the same input parameters space. The latter deals with the physical correctness of the SA results. The present article is a sequel to an earlier study in this journal (Saltelli, Homma, *Comp. Stat. and Data Anal.* **13** (1) 1992, 73–94 of limitations in existing SA techniques, where the inadequacy of existing schemes to deal with non-monotonic relationships within the model was pointed out.

International benchmark test models were taken from the field of performance analysis of nuclear waste disposal in geological formations. The results based on these models show that the modified version of the Hora and Iman method proposed in this paper is extremely robust, when compared with the other existing statistics, even in the presence of model non-monotonicities. This importance measure is accurate, although its application is demanding – in terms of computer time – for system with very large numbers of input parameters. Its predictions are also among the most reproducible. The newly proposed iterated fractional factorial design appears to score the best in reproducibility. The accuracy of this latter method demands further investigation.

Keywords: Sensitivity analysis; Uncertainty analysis; Nonparametric statistics; Model nonmonotonicity.

Correspondence to: A. Saltelli, Environment Institute, CEC-JRC ISPRA Establishment, I-21020 Ispra (Varese) Italy.

1. Introduction

The present study investigates the performances of statistical techniques which are currently used in Sensitivity Analysis (SA) for computer codes.

More specifically Probabilistic System Assessment (PSA) computer models are considered, which are used to estimate the distribution function of model output functions by repeatedly sampling input parameter values from assigned distributions [26]. For these models an analysis of the output/input relationship requires more than a point derivative of the output with respect to the model input parameters, given the multivariate nature of the input and the fact that the range of uncertainty which affects the input parameters can be large.

When dealing with this type of model it is customary to use SA estimators which yield a global measure of sensitivity, where the influence of a given parameter on the output is averaged both on the distribution of the parameter itself and on the distributions of all the remaining parameters (global SA techniques). On the other hand methods such as differential analysis or the Adjoint Method give sensitivity measures relative to single points in the multidimensional space of the sample [15,14].

It can be observed that the use of global SA techniques is not limited to PSA models and codes. In fact every model is likely to have uncertain input parameters; similarly the range of uncertainty is likely to change from parameter to parameter, and a thorough analysis of model response over the entire space of input parameter values may well be considered as an indispensable part of model/code quality assurance process. For a discussion of Monte Carlo based SA see also [12,15] and Appendix in [7].

Several SA techniques are described in the literature, and a few intercomparison studies are also available. A study of the performance of sensitivity analysis techniques on different test models was performed by Iman and Helton [14,15]. These studies pointed out the effectiveness of the regression based non-parametric techniques such as the Standardised Rank Regression Coefficient (SRRC) and Partial Rank Correlation Coefficient (PRCC). Differential analysis and response surface analysis based on fractional factorial design were also investigated by these authors.

The present work partially reproduces another intercomparison exercise based on the Level 0 test case [30], where a variance analysis was done on the predictions of several SA techniques. The main finding of this previous work was a confirmation of the higher reproducibility of the SRRC, PRCC estimators both with respect to their parametric equivalent and to the other non-parametric tests being investigated. The reason for repeating the analysis on a different sample is mainly to compare the performances of these tests with the new ones introduced in Section 3 on a larger set of test cases.

The search for better estimators arises from another study where three different test cases were used to discuss the inadequacy of available nonparametric statistics in presence of model non-monotonicity [29]. The same paper also mentioned in its conclusions that the Hora and Iman importance (see

Section 3) measure might overcome these difficulties, since this method – at least in principle – is not affected by the degree of monotonicity of the model. The present work is intended to test such an hypothesis.

It should be pointed out that SA method performance is model dependent. SA is easier for linear models than for non-linear ones, and for monotonic rather than for non-monotonic ones. SA is also dependent on how the output variables are considered. Rank or logarithmic transformations of the output are often employed, which have a great influence on the relative importance of the input parameters. Also, different SA estimators point to different types of variables. For instance a linear correlation measure will give higher weight to a parameter influencing the output in a linear fashion, whereas a rank based version of the same measure will highlight parameters whose rank is linearly correlated to the ranked output (this is the case with the Pearson and Spearman coefficients, see Section 3). When the output distribution ranges over more than one order of magnitude this can lead to diverging results, as the linear SA estimators will be strongly sensitive to the values in the upper tail of the output distribution, while rank-based estimators will give equal weight to all the values in the distribution [29,30]. This can lead to a considerable degree of the subjectivity in the interpretation of the results as well as in the choice of the proper SA estimator.

Comparing the performances of SA techniques requires a definition of the properties on which the comparison must be based. This has been attempted in the present paper by considering an SA estimator as a measuring instrument. In experimental physics instruments are normally characterised with respect to three main properties: precision, reproducibility and accuracy. Precision usually indicates how many digits can be effectively read out of the instrument display (or which is the smallest value of the quantity under examination which the instrument can detect or differentiate). Reproducibility deals with the instrument capacity of yielding the same result when measuring repeatedly the same sample. Accuracy is "the extent to which the results of a calculation... approach the true values of the calculated... quantities, and are free of error" [18].

When using global SA estimators as in the present study the precision deals with how many variables can be considered as successfully 'ranked', [1] or identified as influential; this can be done using hypothesis testing [5]: for any SA estimator a test statistic is generally available to indicate, at a given significance level, which variables have significant values and which do not. The threshold estimator value depends upon the significance level as well as upon the size of the sample available for the analysis. In the context of hypothesis testing the relative efficiency (precision) of different tests can be compared for given values

[1] There is a possible source of confusion in the fact that the term "ranking" is used here both for the process of replacing data with ranks and for the ordering of the influential variables (eg. the most influential variable has rank one). The context should make clear which is which.

of significance level and of power [5]. This aspect of the problem is not addressed in the present paper.

The reproducibility of an estimator is closely related to its precision, and, in the present context, it can be defined as the extent to which the test produces the same variable ranking when repeating the analysis on a different sample of the same size.

The approach taken in the present study focuses on technique reproducibility and accuracy. The reproducibility of the different sensitivity analysis estimators is measured through an empirical variance analysis on selected models, in which the SA estimator predictions are compared over different samples of the same size. The analysis is then repeated at a different sample sizes (see Section 4).

The analysis of a technique's accuracy is a more delicate point. Because different techniques look at different aspects of the input–output relationship it is possible that different predictions (rankings) are in fact both correct, providing complementary information on the system. As mentioned above a given input parameter can influence the output more than the rank of the output, or the output mean more than the output variance and so on. At the same time a given technique can fail while another gives the correct answer (see Section 8). There is no automated procedure to study the technique accuracy (in fact there are no standards to compare with). In the present work deductions largely rely on the knowledge of the mathematical structure of the model being investigated.

Two test models are considered in this study. Both of them have been used in previous works [29,30] and pertain to the field of nuclear waste disposal safety analysis. The first one can be considered as a worst case model for SA, being non-linear, non-monotonic and censored (many output values are zero). This test case is mainly used to investigate technique reproducibility. The second one is less pathological, though it displays interesting non-monotonic features which are suitable for discussion of technique accuracy. These models are used to discuss the effectiveness of the modified Hora and Iman method proposed in this work as compared with those of the existing SA estimators with respect to both accuracy and reproducibility. Results are available for the iterated fractional factorial method for only the first of these models.

2. Test cases

The test cases employed in the present work have been the object of international intercomparison exercises. They pertain to the field of the analysis of the environmental impact of radioactive waste disposal in deep geological formations using Monte Carlo codes. The exercises, named Level 0 and Level E, are the first of a series of benchmarks being conducted within the PSAC (Probabilistic System Assessment Code) User Group. This body, active since 1985, is coordinated by the Nuclear Energy Agency (NEA) of the OECD [4,25,28].

Because of their nature as benchmarks, both models are well documented. Furthermore the software employed in the present study has been tested in the

Table 1

Description of parameters to be treated as random variables in the Level 0 exercise.

Notation	Definition	Distribution	Attributes [a]	Units
RLEACH	leach rate	loguniform	/0.00269, 12.9/	kg/m^2/a
XBFILL	buffer thickness	uniform	/0.5, 5/	m
XPATH	geosphere path length	uniform	/1000, 10000/	m
V	ground water velocity	loguniform	/0.001, 0.1/	m/a
DIFFG	geosph. diff. coeff.	normal	mean = 0.04, std = 0.001	m^2/a
ADISPG	dispersivity in the geosph.	loguniform	/2, 200/	m
ABSR	water extraction rate	uniform	/5 × 10^5, 5 × 10^6/	m^3/a
RMW	water ingestion rate	uniform	/0.7, 0.9/	m^3/a
BD(Cs)	sorpt. const. in the buffer	lognormal	mean = 0.46, std = 0.86	m^3/kg
BD(I)	sorpt. const. in the buffer	lognormal	mean = −5.07, std = 1.34	m^3/kg
BD(Pd)	sorpt. const. in the buffer	lognormal	mean = −1.91, std = 0.669	m^3/kg
BD(Se)	sorpt. const. in the buffer	lognormal	mean = −2.38, std = 0.143	m^3/kg
BD(Sm)	sorpt. const. in the buffer	lognormal	mean = −2.13, std = 0.605	m^3/kg
BD(Sn)	sorpt. const. in the buffer	lognormal	mean = −1.77, std = 0.729	m^3/kg
BD(Zr)	sorpt. const. in the buffer	lognormal	mean = −0.71, std = 0.5	m^3/kg
KD(Cs)	sorpt. const. in the geosph.	lognormal	mean = −1.46, std = 1.6	m^3/kg
KD(I)	sorpt. const. in the geosph.	lognormal	mean = −6.07, std = 2.6	m^3/kg
KD(Pd)	sorpt. const. in the geosph.	lognormal	mean = −2.91, std = 1.4	m^3/kg
KD(Se)	sorpt. const. in the geosph.	lognormal	mean = −3.38, std = 0.3	m^3/kg
KD(Sm)	sorpt. const. in the geosph.	lognormal	mean = −3.13, std = 1.2	m^3/kg
KD(Sn)	sorpt. const. in the geosph.	lognormal	mean = −2.77, std = 1.4	m^3/kg
KD(Zr)	sorpt. const. in the geosph.	lognormal	mean = −1.71, std = 1.0	m^3/kg

[a] For uniform and loguniform distributions the Attributes are the interval endpoints. For the normal distributions these are the mean and the standard deviation. For the lognormal distributions mean and standard deviation refer to the logarithm (base 10) of the variable.

intercomparison process. Another advantage of these models is that they appear quite complex as far as the input–output relationship is concerned, whilst being computationally not too expensive to run.

The test models involve the computation of the dose to man resulting from migration of selected radionuclides through a multi-barrier system (waste form, near field, far field, biosphere). Model input parameters can either be given as constants or as distributed parameters. The output being compared is the frequency distribution of the output dose. Comprehensive reports describe both Level 0 [19] and Level *E* [20]. A more succinct description can be found in [29]. In this note only the essential elements are given.

The Level 0 model contains very simple barrier sub-models, in the form of easy to compute analytical formulae. Seven uncorrelated isotopes are considered and a total of 22 input parameters are taken as uncertain. The distribution characteristics are given in Table 1. It can be seen that large ranges of variation (orders of magnitude) are involved. The mean annual dose as a function of time is shown in Figure 1 for the sample employed in the present study (2500 runs). The same figure also gives the confidence bounds on the mean [24] and the output annual dose from the 5 simulations yielding the highest peaks. For each

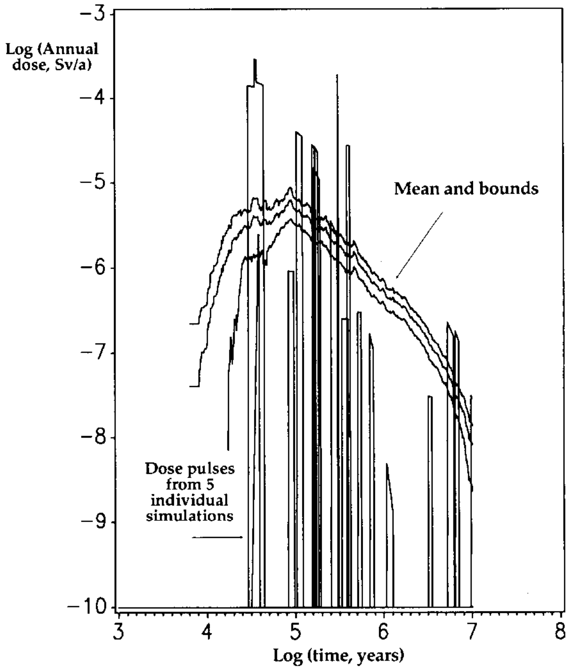

Fig. 1.

simulation the output vector contains mostly zeros (no dose), the dose pulses reaching the biosphere as sharp peaks or square waves. Consequently for each time point there is a large percentage of zero output, which results in ties when the rank of doses are taken. Model coefficients of determination (Section 3) are also very low for this model (Figure 2).

For these reasons the variable 'maximum dose between 0 and the considered time point' has been used in the present study. The percentage of non-zero outputs for this variable is given in Figure 3. It can be seen that in spite of the variable transformation the percentage of non-zero outputs is still low for the first time points. The same figure also plots the model coefficients of determina-

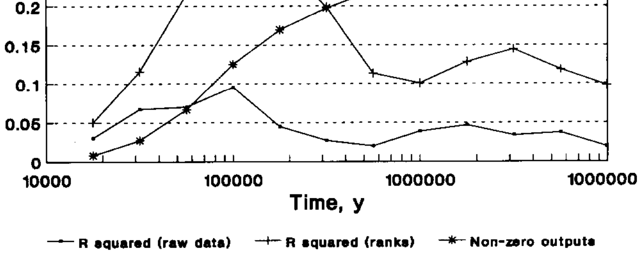

Fig. 2. R squared coeff. for Level 0 – dose.

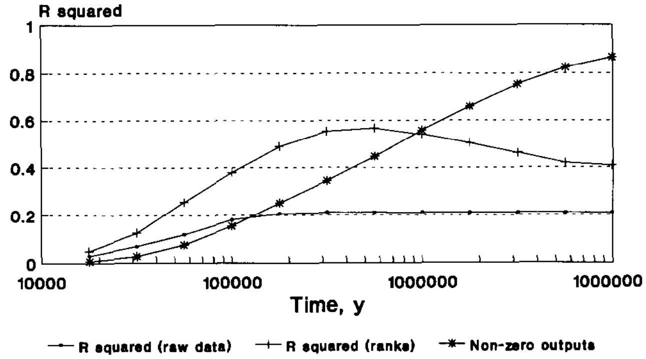

Fig. 3. R squared coeff. for Level 0 – max. dose.

tions for the regression models based on the raw values and on the ranks. The large difference between the values of these two coefficients demonstrates the non-linearity of the model.

The Level *E* test case is computationally more demanding. The geosphere model includes a two layers pathlength where dispersion, advection, decay and chemical retention have to be modeled for the I129 isotope and for the Np237–U233–Th229 radionuclide decay chain. The characteristics of the twelve distributed parameters are given in Table 2. It can be seen that the parameter

Table 2

Description of parameters to be treated as random variables in the Level *E* exercise.

Notation	Definition	Distribution	Attributes (endpoints)	Units
CONTIM	containment time	uniform	/100, 1000/	a
RELRI	leach rate for Iodine	log-uniform	$/10^{-3}, 10^{-2}/$	a^{-1}
RELRC	leach rate for Np chain nuclides	log-uniform	$/10^{-6}, 10^{-5}/$	a^{-1}
FLOWV1	water velocity in geosphere's first layer	log-uniform	$/10^{-3}, 10^{-1}/$	m/a
PATHL1	length of geosphere's first layer	uniform	/100, 500/	m
RETF1I	geosphere retardation coeff. for Iodine (first layer)	uniform	/1, 5/	–
RETF1C	factor to compute geosphere retardation coeff. for Np chain nuclides (first layer)	uniform	/3, 30/	–
FLOWV2	water velocity in geosphere's second layer	log-uniform	$/10^{-2}, 10^{-1}/$	m/a
PATHL2	length of geosphere's second layer	uniform	/50, 200/	m
RETF2I	geosphere retardation coeff. for Iodine (second layer)	uniform	/1, 5/	–
RETF2C	factor to compute geosphere retardation coeff. for Np chain nuclides (second layer)	uniform	/3, 30/	–
STFLOW	stream flow rate	log-uniform	$/10^{5}, 10^{7}/$	m^{3}/a

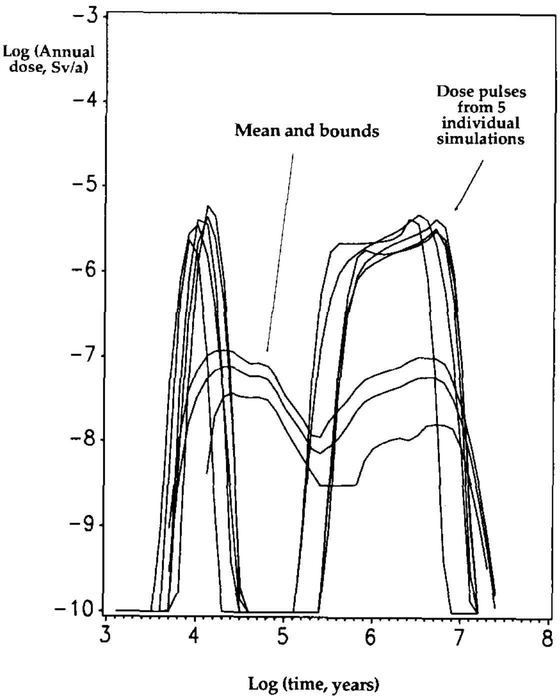

Fig. 4.

variation is less pronounced than for Level 0, and the spread in results among the various runs is also less severe.

The mean total dose for a sample of size 2500 is shown in Figure 4, where the first pulse is due to the I129 contribution and the second one to the Np237 chain. In this figure the annual dose from the 5 'highest pulse' simulations is also given.

The model coefficients of determination and the percentage of non-zero output are given in Figure 5. Because the percentage of non-zero runs is much higher than for the Level 0 case, both the total dose and the maximum total dose (Figure 6) have been taken for the variance analysis. Interesting in this test

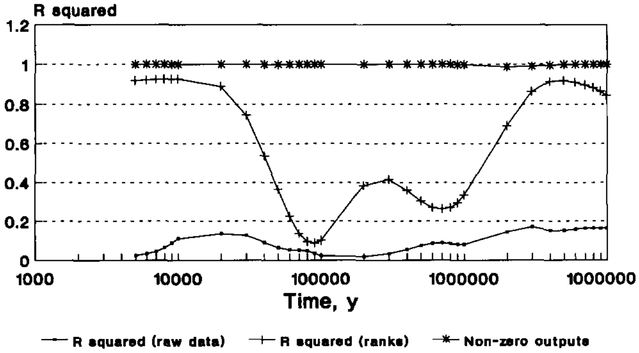

Fig. 5. R squared for Level E – dose.

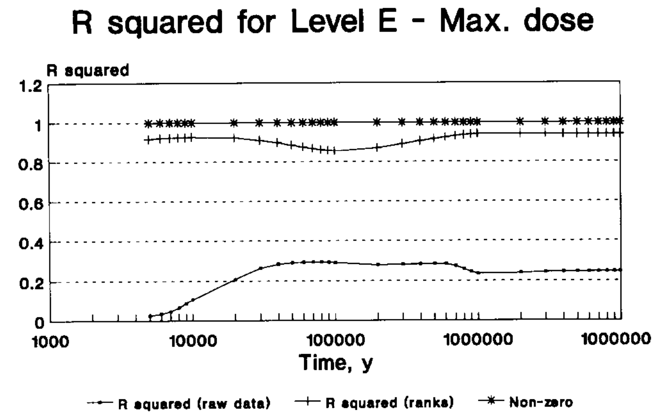

Fig. 6. R squared for Level E – max. dose.

case is the multi-modal shape of the model coefficient of determination on ranks, which oscillates between values higher than 0.9 (very efficient regression model) and values below 0.1 (useless regression model) for the output variable 'total dose'.

Model computation times ranged from a less than a hundredth of second per run for the Level 0 to an average of 0.6 seconds per run for Level E using the JAERI Fujitsu FACOM-M780 computer (about 33 mips).

3. Sensitivity analysis methods

The Sensitivity Analysis estimators investigated in the present work include several parametric and non-parametric techniques, based on regression–correlation measures, as well as some two-sample tests, some variance reduction methods, the 'Importance Measure' test and an iterated fractional factorial design method. For the sake of conciseness the tests are referred to in the text by the abbreviated name used in the computer program. A list of the tests and abbreviations is given below.

Pearson Correlation Coefficient	PEAR
Spearman Rank Correlation Coefficient	SPEA
Partial Correlation Coefficient	PCC
Partial Rank Correlation Coefficient	PRCC
Standardised Regression Coefficient	SRC
Standardised Rank Regression Coefficient	SRRC
Smirnov Test Statistic	SMIR
Cramer–Von Mises Test Statistic	CRAM
Mann–Whitney Test Statistic	TMWT
Two-sample t Test Statistic	TTST
Input distribution shrinking approach	SHRI
Input distribution shifting approach	SHFT

Hora and Iman Importance Measure	HIM
Modified Hora and Iman Importance Measure	HIM*
Iterated Fractional Factorial Design	IFFD

The reproducibility of the first 10 tests (PEAR though TTST) has already been investigated in the variance analysis study mentioned in the Introduction [30], using a different sample of the same test model. The tests are all extensively described in the literature and only their essential features are recalled here.

The last four methods, SHRI, SHFT, HIM and IFFD have been recently proposed and – to the authors' knowledge – their reproducibility has not yet been investigated.

PEARS and SPEA are the usual correlation coefficients based on the raw values and on the ranks, respectively [5].

PCC and its rank equivalent PRCC are also correlation measures which look at the degree of correlation between output and any input variable by removing the effect due to existing correlations between this variable and any other input variable [10,16]. It should be mentioned that in the present study no correlation is imposed on the input variables, apart from that arising stochastically from the sampling.

SRC and SRRC are regression based measures, i.e., they are the coefficient for the regression model of the system. The coefficients refer to the standardised input and output variables obtained by subtracting the sample mean from the original values and dividing by the sample standard deviation [10,16]. When using these coefficients it is useful to consider the model coefficient of determination R_y^2 (on raw values or ranks) which gives the percentage of the variance of the input data reproduced by the regression model. R_y^2 values close to one indicate an effective regression model, and hence the possibility of ranking the model parameters based on regression model coefficients. Low R_y^2 values indicate a poor regression model; a low percentage of the data variance is accounted for, so that the ranking of the parameters based on their contribution to this fraction loses significance.

SRC and PCC always produce the same ranking, unless significant correlations are imposed on the input variables [16], which is not the case of the present study. The same applies to SRRC and PRCC, so that the use of both methods is redundant for this particular application. Previous studies [29,30] have pointed out the high correlation existing between the predictions of SRRC (and PRCC) with the other non-parametric tests such as SPEA and, to a lesser extent, SMIR, CRAM, TMWT. This implies that when the R_y^2 (on ranks) coefficient flags an inadequate regression model, also the predictions from these other non-parametric tests are impaired.

SMIR and CRAM tests are both 'two-sample' tests designed to check the hypothesis that two samples belong to the same population [5]. They are used in SA by partitioning the sample of the input variables according to the quantiles of the output variable distribution. In the present application one subsample collects all the input values for the selected input variable which correspond to the 10% highest output values. The second sub-sample collects the remaining values. The tests are based on differences in the cumulative distributions of the two sub-samples [5].

TTST and TMWT are also two-sample tests, used in this application with a 10–90% splitting of the input sample as described above. These tests check the means of the two sub-samples. Their interest mainly lies in the fact that TTST is the exact parametric equivalent of TMWT, i.e., TMWT is the same as TTST if the sample values are replaced by their ranks [5].

The SHRI and SHFT estimators originate from another benchmark of the PSAC group, the Level S [22]. This exercise addresses sensitivity analysis on the Level *E* test model. The participants in Level *S* are asked to rank the parameters of the Level *E* exercise based on the answer to two distinct questions:

(1) What is the reduction in output variance associated with a 10% reduction in the range of the input parameter distribution (5% on each side)?

(2) What is the shift in output mean associated with an upward 5% shift in the mean of the input parameter distribution?

The exercise is facilitated by the fact that the input distributions for the Level *E* model are either uniform or loguniform, so that an analytical, hence exact, solution is available for both questions (1) and (2) [22]. Although the results from Level *S* are not discussed in the present work it was felt that an analysis of the performances of SHFT and SHRI might be of interest. In the present study the numerical value of SHRI is computed as follows: For each parameter the output sample is censored by excluding all the values which correspond to values of the input variable under consideration below the 5th percentile and above the 95th percentile of its distribution. The variance of the censored sample is computed and SHRI is evaluated as the ratio of the censored versus the uncensored sample variance. SHFT is similarly computed as follows: for each input variable the inputs are sorted in ascending order and values in the lower tail are dropped till a 5% increase in the parameter mean is achieved. The output is censored by excluding the values corresponding to the dropped inputs. The mean of the censored sample in computed. Its ratio to the mean of the uncensored sample constitutes the numerical value of SHFT.

The importance measure proposed by Hora and Iman [8] is used in this work following a computational scheme suggested by Ishigami and Homma [17]. Its derivation is repeated here as it will be needed when discussing the modified version of the test.

Let the output variable *Y* be a function of *K* variables

$$Y = h(X_1, X_2, \dots X_K). \tag{3.1}$$

Assuming that the input is composed of independent random variables the joint Probability Density Function (PDF) of the input is

$$f(X_1, X_2, \dots X_K) = \prod_1^K f_i(X_i). \tag{3.2}$$

Mean and variance of Y can then be expressed as

$$\langle Y\rangle = \int\int \ldots \int h(X_1, X_2, \ldots X_K) \prod_1^K f_i(X_i)\, \mathrm{d}X_i, \tag{3.3}$$

$$\begin{aligned} V_Y &= \int\int \ldots \int \{h(X_1, X_2, \ldots X_K) - \langle Y\rangle\}^2 \prod_1^K f_i(X_i)\, \mathrm{d}X_i \\ &= \int\int \ldots \int \{h(X_1, X_2, \ldots X_K)\}^2 \prod_1^K f_i(X_i)\, \mathrm{d}X_i - \langle Y\rangle^2. \end{aligned} \tag{3.4}$$

Let now $V_Y(\tilde{x}_j)$ represent the output variance when the input variable X_j is fixed to its value $\tilde{x}_j$

$$\begin{aligned} V_Y(\tilde{x}_j) &= \int\int \ldots \int \left\{h(X_1, X_2, \ldots \tilde{x}_j, \ldots X_K) - \langle h(\tilde{x}_j)\rangle\right\}^2 \prod_{\substack{i=1\\ i\neq j}}^K f_i(X_i)\, \mathrm{d}X_i \\ &= \int\int \ldots \int \left\{h(X_1, X_2, \tilde{x}_j, \ldots X_K)\right\}^2 \prod_{\substack{i=1\\ i\neq j}}^K f_i(X_i)\, \mathrm{d}X_i - \langle h(\tilde{x}_j)\rangle^2, \end{aligned} \tag{3.5}$$

where $\langle h(\tilde{x}_j)\rangle$ is the mean of the output Y when the variable X_j is fixed to the value $\tilde{x}_j$, i.e.,

$$\langle h(\tilde{x}_j)\rangle = \int\int \ldots \int h(X_1, X_2, \ldots \tilde{x}_j, \ldots X_K) \prod_{\substack{i=1\\ i\neq j}}^K f_i(X_i)\, \mathrm{d}X_i. \tag{3.6}$$

The dependence of the variance $V_Y(\tilde{x}_j)$ upon the specific value $\tilde{x}_j$ can be eliminated by averaging $V_Y(\tilde{x}_j)$ according to the PDF of X_j to yield

$$V_Y^j = \int V_Y(\tilde{x}_j) f_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j. \tag{3.7}$$

Substituting Equation 3.5 in Eq. 3.7 gives

$$V_Y^j = \int\int \ldots \int \{h(X_1, X_2, \ldots X_K)\}^2 \prod_{i=1}^K f_i(X_i)\, \mathrm{d}X_i - \int \langle h(\tilde{x}_j)\rangle^2 f_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j. \tag{3.8}$$

Comparing Equations 3.4 and 3.8 leads to the relation

$$V_Y - V_Y^j = U_j - \langle Y\rangle^2, \tag{3.9}$$

where

$$U_j = \int \langle h(\tilde{x}_j)\rangle^2 f_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j. \tag{3.10}$$

According to Hora and Iman the measure of importance is defined as the square root of the difference in Equation 3.9, i.e.,

$$I_j = \sqrt{V_Y - V_Y^j} = \sqrt{U_j - \langle Y\rangle^2}. \tag{3.11}$$

Since the quantity $\langle Y \rangle^2$ is a constant, the variable ranking is in fact based on the values of U_j, i.e., variable k is more important than variable j if $U_k > U_j$.

The importance measure is hence related to the computation of the integral in Equation 3.10. Assuming that the model does not allow an analytical determination of this integral a Monte Carlo method can be applied. The computation requires the average $h(\tilde{x}_j)$ to be computed, and the integral (3.10) to be evaluated by integrating on all the possible values of $\tilde{x}_j$. The number of simulations required for this computation is hence M^*N, where M is the mesh size used to compute the integral (3.10) and N is the sample size used in evaluating $h(\tilde{x}_j)$ for the specified $\tilde{x}_j$ value. This calculation method is too expensive and impractical.

In order to find a more effective computation scheme U_j can be rewritten as [17]

$$\begin{aligned} U_j &= \int \left\{ \int\int \cdots \int h(X_1, X_2, \ldots \tilde{x}_j, \ldots X_K) \prod_{\substack{i=1 \\ i \neq j}}^{K} f_i(X_i)\, \mathrm{d}X_i \right\}^2 f_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j \\ &= \int\int \cdots \int h(X_1, X_2, \ldots \tilde{x}_j, \ldots X_K) \times h(X_1', X_2', \ldots \tilde{x}_j, \ldots X_K') \\ &\qquad \times \left(\prod_{\substack{i=1 \\ i \neq j}}^{K} f_i(X_i)\, \mathrm{d}X_i \right) \left(\prod_{\substack{i=1 \\ i \neq j}}^{K} f_i(X_i')\, \mathrm{d}X_i' \right) f_j(\tilde{x}_j)\, \mathrm{d}\tilde{x}_j \\ &= \int\int \cdots \int h(X_1, X_2, \ldots X_K) \times h(X_1', X_2', \ldots X_j, \ldots X_K') \\ &\qquad \times \left(\prod_{i=1}^{K} f_i(X_i)\, \mathrm{d}X_i \right) \left(\prod_{\substack{i=1 \\ i \neq j}}^{K} f_i(X_i')\, \mathrm{d}X_i' \right). \end{aligned} \tag{3.12}$$

The above equations shows that U_j is nothing more than the expectation value of the function

$$\begin{aligned} H(X_1, X_2 \ldots X_K, X_1', X_2', \ldots X_{j-1}', X_{j+1}', \ldots X_K') &= h \times h^* \\ = h(X_1, X_2, \ldots X_K) \times h(X_1', X_2', \ldots X_{j-1}', X_j, X_{j+1}', \ldots X_K'), \end{aligned} \tag{3.13}$$

of a set of $(2K-1)$ independent variables. Assuming again that N is the sample size usually employed to estimate the expectation value of the output, the number of model executions needed to compute the expectation value of this function is simply $2N$, because the function H is expressed as a twofold product of the original model function h. For any given run of the PSA computation H can be evaluated by multiplying h, computed from a sampled vector of the K input parameters, with h^*, computed by resampling all the parameters but the jth one (see next section). The total number of model executions needed to rank all the variables is N times $(1+K)$, where K is the number of variables.

$$\begin{array}{l}\textit{Base sample}\\ \textit{matrix}\end{array} \underline{\underline{x_B}} = \left\{\begin{array}{ll} x_{1,1}, x_{1,2}, \ldots & x_{1,K} \\ x_{2,1}, x_{2,2}, \ldots & x_{2,K} \\ . & \\ . & \\ . & \\ x_{N,1}, x_{N,2}, \ldots & x_{N,K} \end{array}\right\} \rightarrow \begin{array}{l}\textit{output}\\ \textit{vector}\end{array} \overline{Y_B} : \left\{\begin{array}{l} y_1 \\ y_2 \\ . \\ . \\ . \\ . \\ y_N \end{array}\right\}$$

$$\begin{array}{l}\textit{sample matrix}\\ \text{for } \textit{variable } X_j\end{array} \underline{\underline{x_j}} = \left\{\begin{array}{l} \xi_{1,1}, \xi_{1,2}, \ldots\ldots\ldots\ x_{1,j}\ \ldots\ \xi_{1,K} \\ \xi_{2,1}, \xi_{2,2}, \ldots\ldots\ldots\ x_{2,j}\ \ldots\ \xi_{2,K} \\ . \\ . \\ . \\ \xi_{N,1}, \xi_{N,2}, \ldots\ldots\ldots\ x_{N,j}\ \ldots\ \xi_{N,K} \end{array}\right\} \rightarrow \begin{array}{l}\textit{output}\\ \textit{vector}\end{array} \overline{Y_j} : \left\{\begin{array}{l} y^j_1 \\ y^j_2 \\ . \\ . \\ . \\ . \\ y^j_N \end{array}\right\}$$

$$\begin{array}{l}\textit{sample matrix}\\ \text{for } \textit{variable } X_l\end{array} \underline{\underline{x_l}} = \left\{\begin{array}{l} \xi_{1,1}, \xi_{1,2}, \ldots\ x_{1,l}\ \ldots\ldots\ldots\ \xi_{1,K} \\ \xi_{2,1}, \xi_{2,2}, \ldots\ x_{2,l}\ \ldots\ldots\ldots\ \xi_{2,K} \\ . \\ . \\ . \\ \xi_{N,1}, \xi_{N,2}, \ldots\ x_{N,l}\ \ldots\ldots\ldots\ \xi_{N,K} \end{array}\right\} \rightarrow \begin{array}{l}\textit{output}\\ \textit{vector}\end{array} \overline{Y_l} : \left\{\begin{array}{l} y^l_1 \\ y^l_2 \\ . \\ . \\ . \\ . \\ y^l_N \end{array}\right\}$$

Fig. 7. Base sample matrix and HIM sample for two generic variables X_j and X_l. (Note that the input matrix for the two generic variables only differ for columns "j" and "l". N and K indicate the number of runs and the number of variables respectively. The HIM for variable X_j is simply: $\text{HIM}(j) = \Sigma_{i=1}^{N} y_i y_i^j / N$.)

In the following the convention of Figure 7 will be taken, using $\overline{Y_B}$ to indicate the base vector of the values of the function h above, and $\overline{Y_j}$ for the h^* function vector. The elements of $\overline{Y_B}$ and $\overline{Y_j}$ will be indicated with y_i and y_i^j respectively, with $i = 1, 2, \ldots N$. The summation for HIM in Figure 7 will be always greater for correlated $\overline{Y_B}$ and $\overline{Y_j}$ vectors than for uncorrelated ones.

The IFFD estimator was developed by Andres [3] for sensitivity analysis of the Canadian SYVAC3-CC3 model of a nuclear fuel waste disposal site [6]. Critical issues for that application were:

(a) SA of a model having thousands of parameters;

(b) SA with a minimal number of simulations because of the computer time involved in running the model;

(c) Identification of parameters having a nonmonotonic effect on an output variable.

IFFD differs from the other methods tested in that parameter values are chosen according to a statistical design, rather than being randomly sampled. Three levels of each parameter are selected: a low value, a middle value, and a high value. To standardize the treatment of each parameter, these levels are chosen to be specific quantiles of the probability distribution used in the Monte Carlo approach for sampling the parameter. The results quoted in sections 5 and 6 were based on the 0.05, 0.50 (median), and 0.95 quantiles. The benefit of using discrete parameter values is to reduce the variance of any SA estimator. The disadvantage is that it is possible to fail to identify a parameter that has an effect on the output variable over a small part of its domain. The 3-level design chosen can detect linear and quadratic effects.

The statistical design consists of multiple iterations of a simple 2-level fractional factorial design. In each iteration, parameters are randomly grouped together (aliased). The design is further randomized in each iteration by randomly assigning groups to variables of the design, and by randomly orienting each parameter with others in its group (i.e., some parameters in a group take high values when other parameters take low values).

The iterated design is converted from a 2-level design to a 3-level design by setting a parameter to its middle value in all simulations of a few randomly selected iterations.

A typical iterated design uses 256 simulations to identify up to 8 important parameters out of as many as 4000 parameters. The design consists of 16 iterations of a 16-simulation 2-level fractional factorial design. Each parameter takes a middle value in 4 randomly selected iterations out of the 16. In each iteration, parameters are assigned to 8 groups. The fractional factorial design is of Resolution IV, so that the linear effects of each group can be determined without being confounded with two-way interactions among the groups. If the number of important parameters is very small, it is possible by analyzing just one iteration (i.e., 16 simulations) to identify the groups containing those parameters. After several iterations, the important parameters can be identified as the parameters belonging to a succession of important groups.

This screening process is best accomplished using stepwise regression. Each important parameter will give rise to copycats; these are unimportant parameters that by chance shared a group with the important parameter several times. Their linear effects will be correlated with that of the important parameter. By using stepwise regression, we may remove the contribution of each important parameter as it is identified, thereby eliminating the copycats. It is possible to apply the screening process in an efficient way based on the structure of the design, without actually applying stepwise regression to a matrix with 256 rows and 4000 columns. The details are described in [3]. IFFD works well in applications where there are a few important parameters, hidden among a large set of unimportant parameters.

In situations where there are many equally important parameters, IFFD will tend to give spurious results. One effective way to determine if IFFD is giving meaningful results in a particular application is to use dummy parameters. For example, the 256-simulation method described above could be applied to a data set containing 2000 true parameters used by a model, and 2000 dummy parameters that are not used. As long as the stepwise regression procedure generates parameters from the true parameter set, one can be confident that real effects are being found. When the procedure identifies a dummy parameter as important, all subsequent parameters generated by stepwise regression are suspect.

This procedure was used in the comparison described below to rank a subset of the parameters being studied.

4. Computational scheme

The reproducibility of the techniques has been investigated by an empirical variance analysis conducted on the two selected test cases.[2] The variance analysis approach is conceptually very simple, though its application in conjunction with the HIM method is somewhat laborious. Letting aside for the moment the complications introduced by HIM, the procedure can be described as follows (see Figure 8).

First a large sample of size N is generated which contains, for each of the N runs, the output time series and the sample input parameter values. In the Level 0 case, for example, the input data set contains 2500 runs. For each run the values assigned to the 22 sampled variable values and total dose values for 12 selected time points are given. The large sample is then partitioned, i.e., subdivided into smaller samples of identical size N_i, such that $\Sigma N_i \leq N$. In the Level 0 example, 6 such partitions are generated:

1st partition : 5 sub-samples of size 500;
2nd partition : 6 sub-samples of size 416;
3rd partition : 7 sub-samples of size 357;
4th partition : 10 sub-samples of size 250;
5th partition : 25 sub-samples of size 100;
6th partition : 50 sub-samples of size 50.

All the partitions are made of the same 2500 runs of the starting large sample. A separate variance analysis is performed for each partition. Take the first partition and the statistical technique SRRC as an example; the SRRC values for each variable at each time point are computed for each of the 5 sub-samples in the first partition. Based on the SRRC estimator's value at each time point and in each partition the input variables are ranked (rank 1 for the most important, rank 22 for the least important). These ranks are then converted into Savage scores [15] and the variables' score for the five partitions at each time point are stored.

For each variable and each time point the mean and variance of the scores over the 5 subsamples are computed. Call this variance var(SRRC, *ntime, nvar*, 500). The mean value over all the variables and time points of this quantity is the desired variance of SRRC at the sample size 500. Call it var(SRRC, 500). The same is repeated for all the techniques (PEAR through IFFD) and for all the partitions (sample sizes 50 through 500).

The discussion of Sections 5–7 is based on the values of var(technique, sample size) so obtained. The reason for the use of the Savage scores (Figure 8)

[2] The expression "variance analysis" does not refer in this context to ANOVA-like tests, but to a mere measurement of variance over different samples.

Step 1) Compute sample of large size N. For instance, in Level 0, N=2500.

Step 2) Break it down into subsamples of smaller size N_i, such that

$$\sum_i N_i = N$$

(for instance N_i = 500; 5 sub-samples).

Step 3) For each of these subsamples S_i perform sensitivity analysis on the selected output variables (here dose at time point) using the SA techniques under analysis. Eg for SRRC compute

SRRC(*nvar*,*ntime*,S_i)

(value of SRRC computed for subsample S_i, variable *nvar* and dose at time point *ntime*).

Step 4) Convert SRRC(*nvar*,*ntime*,S_i) to rank, ie to the order of importance given by SRRC to variable *nvar* at time point *ntime* obtaining

Rank(SRRC,*nvar*,*ntime*,S_i).

In Level 0 the most important variable is assigned rank 1 and the least important rank 22.

Step 5) Convert rank to Savage score as

$$\text{Score}(nvar,ntime,S_i) = \sum_{m=R}^{K} \frac{1}{m}$$

where R = Rank(SRRC,*nvar*,*ntime*,S_i). Given K=22, for R=1 score=3.691; for R=22 score=0.0455.

Step 6) Determine the mean and the variance of Score(*nvar*,*ntime*,S_i) over the 5 subsamples S_i; call this latter

var(SRRC,*nvar*,*ntime*,N_i)

where N_i indicates that the value is relative to the sample size (eg 500) under consideration.

Step 7) Average var(SRRC,*nvar*,*ntime*,N_i) over the K variables and 12 selected time points t o obtain

var(SRRC,N_i)

which is the statistic presented in Figures 10,14,15.

Step 8) Repeat steps 3) to 7) above for a different subdivision of the base sample, eg 10 subsamples of size 250.

Figure 8 - Variance Analysis Scheme
Procedure adopted for the Variance Analysis as exemplified by the application to the Level 0 test case.

The output from SPOP is (generally) time dependent. For each time point

```
/TIME =    9.0E+04 Y/
```

SPOP outputs the mean and standard deviation of the output under consideration

```
EXPECTED VALUE OF YVAR =  0.271E-07 SV/Y +/- 0.988E-08 (TCHEBYCHEFF BOUNDS). ST. DEV. IS :  0.110E-06
```

Then for the given sample size and significance level

```
SAMPLE SIZE: 2500.  SIGNIFICANCE LEVEL ALPHA = 0.050
```

the quantiles of the test distributions, eg

```
QUANTILE(ALPHA/2) FOR THE SPEARMAN TEST DISTR. = -0.392E-01
```

the model coefficients of determination based on SRC, SRRC

```
MODEL COEFFICIENT OF DETERMINATION = 0.036 (ON ROW VALUES) and 0.088 (ON RANKS)
```

the value of the SA estimators

```
         PEAR  SPEA  PCC   PRCC  SRC   SRRC  SMIR  CRAM  TTST  TMWT  SHRI  SHFT   HIM       HIM*
FLOWV1   0.03 -0.05 -0.03 -0.06 -0.03 -0.06  0.24  5.00  5.10 -0.66  0.05  0.05   1.28E-15  1.13
PATHL1   0.02 -0.21 -0.02 -0.22 -0.02 -0.21  0.06  0.23  1.36  1.33  0.24  0.04   7.10E-16  1.02
...
```

and the corresponding rank table.

```
           PEAR  SPEA  PCC   PRCC  SRC   SRRC  SMIR  CRAM  TTST  TMWT  SHRI  SHFT  HIM   HIM*
   FLOWV1  4.00  6.00  4.00  6.00  4.00  6.00  2.00  2.00  2.00  9.00  5.00  4.00  2.00  1.00
   PATHL1  9.00  1.00  9.00  1.00  9.00  1.00  8.00  8.00  8.00  7.00  4.00  6.00  6.00  2.00
...
```

Based on the columns of the previous table the statistic/statistic score correlation coefficients are computed:

```
         PEAR  SPEA  PCC   PRCC  SRC   SRRC  SMIR  CRAM  TTST  TMWT  SHRI SHFT  HIM   HIM*
   PEAR  1.00  0.44  1.00  0.40  1.00  0.40  0.86  0.86  0.87  0.86  0.20  0.82  0.74  0.36
   SPEA  ----  1.00  0.44  0.99  0.44  0.99  0.39  0.39  0.38  0.48  0.05  0.38  0.45  0.51
...
```

Fig. 9. Selected output from the SPOP code, Level E test case. The HIM statistic values are generated as described in Figure 7. The HIM* values have been normalised by a factor equal to $N((N+1)/2)^2$. The statistic/statistic score correlation coefficients are computed by replacing the variable ranks in the upper table by their Savage scores, as described in Figure 8, then taking the linear (Pearson) correlation coefficient between columns.

in place of the ranks is that using the scores gives greater weight to the most important variables (low rank) in computing the estimators' variance.

A computational complication is introduced by the need to compute the HIM estimator. As described in Section 3 this estimator needs a sample size of size $N \times (1+K)$, where N is the number of runs and K is the number of variables.

All computations are performed using the LISA package. The input matrices (X_B and X_j, $j = 1, 2, \dots K$ in Figure 7) are generated by the PREP utility [9]. The output vectors ($\bar{Y}_B$ and $\bar{Y}_j$, $j = 1, 2, \dots K$ in Figure 7) are obtained using different versions of the LISA code [21]. SA estimators and variance analysis are computed using the SPOP statistical post processor [27]. A representative output from SPOP is presented in Figure 9 for the Level E analysis and the $t = 9 \times 10^4$ time point.

IFFD was computed in a different manner from the other SA estimators. Sample sets were generated by the program SAMPLE [2] based on methods described in [1]. Because of the structured nature of the samples, it was not possible to use the partitioning approach. Instead, five new samples were generated for each sample size. The sample sizes used were 512 and 256.

The Level 0 results were computed using a version of the SYVAC3-LZ code [6], rather than LISA. The Level E part of the experiment was not carried out.

Since IFFD ranked only the most important parameters, it was necessary to fabricate rankings for the rest to complete the comparison. These rankings were randomly generated from the unassigned ranks. That is, if IFFD ranked seven

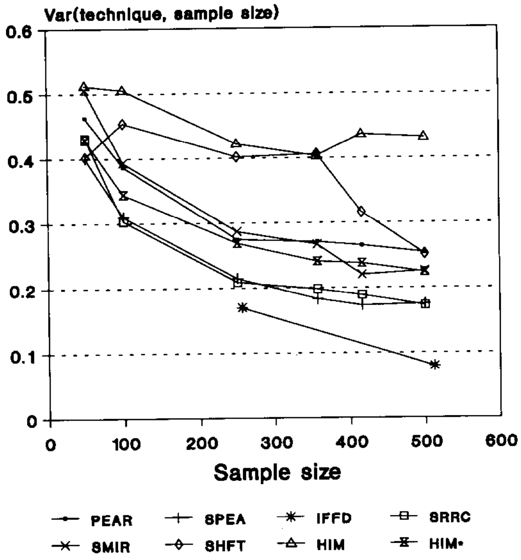

Fig. 10.

out of 22 parameters, the remaining 15 parameters would be assigned a random permutation of the numbers from 8 to 22.

Because of the procedural differences from the other estimators, and the incomplete results, the IFFD results should be considered preliminary. Nevertheless they were included because this estimator shows promise of good performance in comparison with the others.

5. Variance analysis for the Level 0 test case

Values of var(technique, sample size) obtained as described in Section 4 are given in Figure 10. The following observations can be made:

(1) As a general trend variances decrease when increasing sample size, i.e. technique reproducibility increases with sample size as expected.

(2) As described in [30] all the parametric tests yield higher variances than their non-parametric counterpart. This is true of PEAR with respect to SPEA, SRC with SRRC, PRC with PRRC and, to a reduced extent, of TTST with TMWT. The non-parametric tests SPEA, SRRC and PRCC are consistently more reproducible than their parametric equivalent. The SRRC and PRCC methods perform identically [16,29,30].

(3) The other two-sample non-parametric tests SMIR and CRAM perform poorly, sometimes below the parametric PCC, SRC and PEAR.

(4) The SHRI method appears by far the worst as far as reproducibility is concerned. Values of var(SHRI, sample size) range between 0.64 and 0.76 and have not been plotted in the figure. This confirms previous work done on estimators based on sample variance, such as the Klotz test and the Sum of the

Squared Ranks test (see [5] for a description of these methods and [23] for a variance analysis of these estimators).

(5) The HIM method appears to be the second worst.

(6) The SHFT method is also poor as far as reproducibility is concerned. Its performance closes on that of PEAR at the highest sample sizes, but is much worse (higher variances) for the other sample sizes.

The IFFD showed better performance than the other estimators, although it was applied at only two samples sizes.

Figure 10 shows that the estimators SRRC and PRCC are the most reproducible over the entire sample size region explored. One might wonder why, then, a better method is needed.

In fact, as shown in [29], these methods do not provide a base for ranking the input variables in the presence of model non-monotonicity, regardless of the sample size, so that the search for a better estimator is motivated by the need for a better accuracy rather than improved reproducibility.

The remark that non-parametric tests based on rank are systematically better than their parametric equivalent based on the raw values suggests the development of a rank-based version of HIM. This is easily achieved if both the $\bar{Y}_B$ and Y_j vectors are replaced by their ranks before the integration for the expectation value (Equation 3.12 or summation in Figure 7).

The estimator so obtained has been indicated as HIM* in Figure 10. It can be seen that although its performances do not challenge those of SRRC, PRCC and SPEA, they are close to those of the parametric tests SRC, PRC, PEAR.

6. The Level *E* test case

As mentioned in Section 2 the Level E test case displays interesting non-monotonic features which are evidenced by the multi-modal shape of the R_y^2 versus time curve shown in Figure 5.

As discussed in [29] this is due to the fact that the variables which govern the transit time in the geosphere (and hence the dose) have a positive correlation with the output at early time, which becomes negative at later time. Taking water velocity (FLOWV1) as an example, at early times high doses are obtained for high water velocities (positive correlation) whereas at later times the output is depleted unless low values of FLOWV1 are involved (negative correlation). At intermediate times, such as the $t = 90\,000$ y time point shown in Figure 11, a non-monotonic relation is evident between the rank of dose and the rank of FLOWV1, which results in a very low value of SRRC for FLOWV1 at this time point. The thin horizontal line in Figure 11 represent the rank regression between the two variables. For this time point R_y^2 is as low as 0.09.

Comparing Figure 4 (Level E mean dose) with Figure 5 (R_y^2) it can be inferred that the first local minimum of the R_y^2 curve ($t = 90\,000$ y) corresponds exactly to the point in which the correlation between Iodine dose and FLOWV1 passes from positive to negative (SRRC = 0.). This is confirmed by Figure 12,

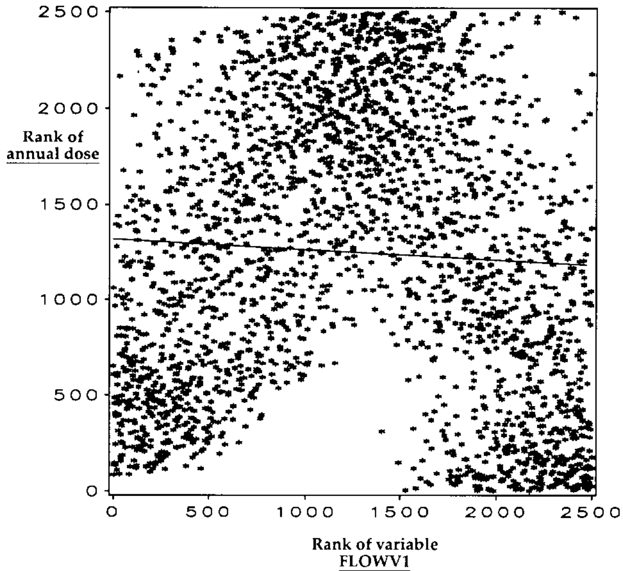

Fig. 11. Scatterplot of rank of dose vs rank of variable for FLOWV1 at time = 9×10^4.

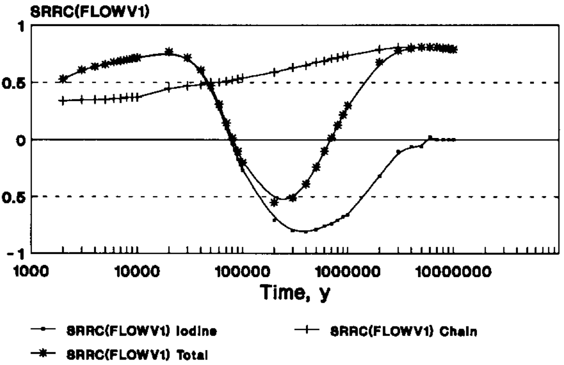

Fig. 12.

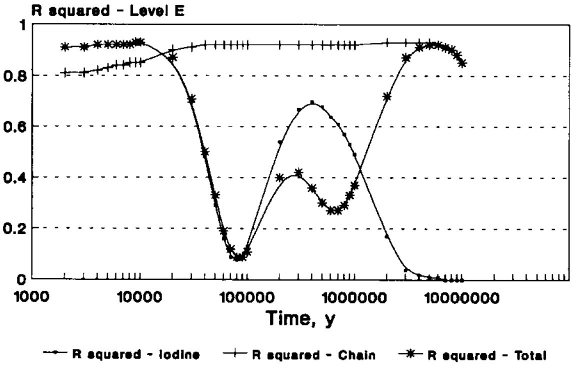

Fig. 13.

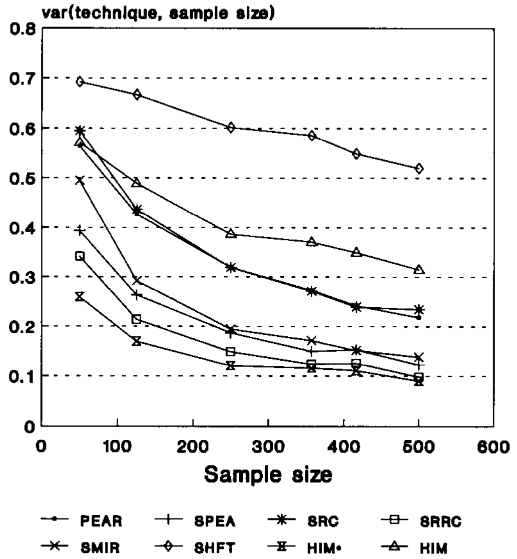

Fig. 14.

where the standard rank correlation coefficient for the variable FLOWV1 is plotted as function of time, and by Figure 13, where the R_y^2 coefficient is plotted for the Iodine and Np chain dose separately. In correspondence of the second local minimum of R_y^2 ($t = 700000$ y, Figure 5) the SRRC for Iodine dose is negative, while that for Np chain dose is positive (Figure 12). This also results in a non-monotonic pattern between 'rank of total dose' and 'rank of FLOWV1' which is evidenced by the rank scatterplot for this time point. In other words there is a correspondence between the minima of the R_y^2 curve and these non-monotonicities which escape the detection of all the rank based correlation/regression estimators (SPEA, SRRC, PRCC...).

This phenomenon is much easier to observe in Level E, where Iodine and Np chain peaks are well resolved, than in Level 0, where many superimposed nuclides contribute to the average dose and the effect is blurred.

The inadequacy of the rank based estimators does not depend upon the sample size, i.e. R_y^2 does not increase when increasing the number of runs.

The results of the variance analysis for Level E are shown in Figure 14. The partitions employed in this case are the same as for the Level 0 analysis.

The variances for all the estimators appear to depend upon the model (compare Figure 10 with Figure 14). Passing from Level 0 to Level E a few estimators worsen in performance, i.e.: variance for PEAR, PCC, and SRC increase moderately (on average); the variance of SHFT increases considerably; for most of the tests better performances are obtained, and in particular: the variance of SHRI decreases slightly: values of var(SHRI, sample size) range between 0.60 and 0.73 (not shown in Figure 14). The variance of SPEA decreases appreciably; SRRC, PRCC, SMIR and CRAM show moderate decreases; HIM*'s variance decreases considerably.

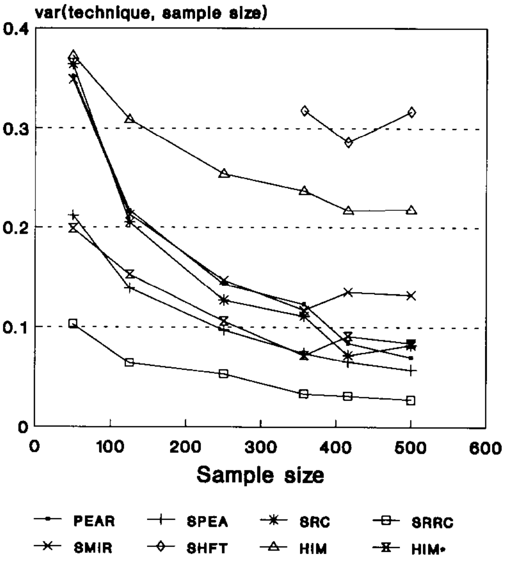

Fig. 15.

As a result of the HIM* reduced variance this estimator is, for Level *E*, the most reproducible at all the sample sizes, followed by SRRC, PRCC and, at some higher variance values, by SPEA. It can be speculated that the improved performance of HIM* as compared with SRRC is due to the fact that HIM* is not affected by model non-monotonicity as much as SRRC. In order to see if this is the case the Level *E* variance analysis has been repeated on the maximum doses (as for Level 0). The effect of taking the maximum of dose between $t = 0$ and the time point, rather than the dose at the time point, is of removing the model non-monotonicity (compare Figure 5 with Figure 6). The results are plotted in Figure 15. As a general trend all the estimators improve their performance, and the spread in results is lower. HIM* is almost unaffected whereas SRRC, PRRC and SPEA become considerably more reproducible. Evidence of the fact that the increased reproducibility of Figure 15 with respect to Figure 14 is due to the removal of the non-monotonicity is the fact the most unaffected technique is HIM*, whose predictions do not rely on sample monotonicity, followed by the two sample tests where also non-monotonicity plays a lesser role.

7. Technique accuracy

The discussion of the technique accuracy is based on the Level *E* test case, taking ‘dose at the time point’ as output variable. The analysis of the technique accuracy mainly relies on the knowledge of the model structure. In any specific application, further measures can be used to test the appropriateness of a technique, including:

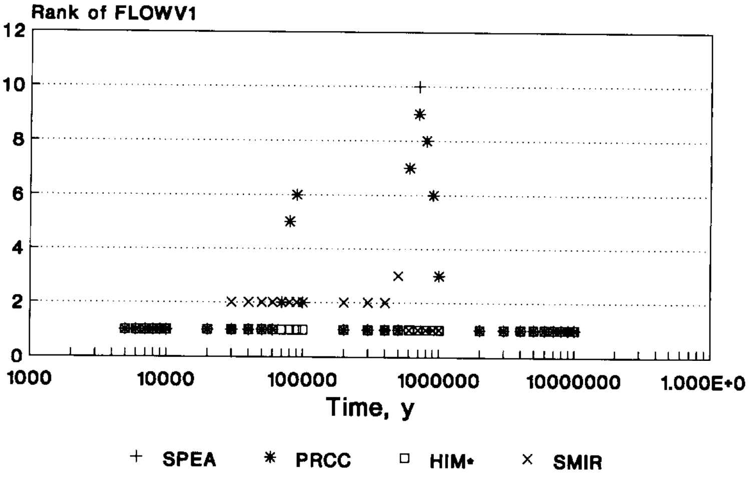

Fig. 16.

- Scatterplots (as that shown in Figure 11), which highlight possible non monotonic trends and the degree of dependence among variables [13].
- Analysis of the range of applicability of tests, i.e., are the data compatible with the test being used? The analysis of the model coefficient of determination for instance indicates whether the regression model for the model output can be used as a basis for ranking the parameters. Especially when using stepwise regression [13] the PRESS (Predicted Sum of Squares) test statistics can be used to select among competing regression models.
- Score correlation tables. These tables provide a measure of the agreement between the rankings produced by different estimators, and are computed by replacing the ranks (such as those plotted in Figure 11) with their Savage scores [15]. These latter allow the correlation between two different tests (e.g. SMIR and SPEA) to be computed by giving the highest weight to agreement (or disagreement) on the most influential variables. One such table is shown in Figure 9, which gives selected SA statistics for the $t = 9 \times 10^4$ time point.

In Figure 16 the technique ranking as a function of time is plotted for selected estimators for the Level E test case. It can be seen that for the $t = 9 \times 10^4$ time point SPEA, PRCC and SMIR all fail to identify FLOWV1 as the most influential variable, while it is still the most influential variable for HIM* (Figure 16). The same happens for the second local minimum of the R_y^2 curve.

The question which arises now is which estimator should be trusted more. It is evident that, even without using the hypothesis testing, PRCC and SRRC predictions should be disregarded for the low R_y^2 points, where the regression model does not have predictive capability. The same can be said of SPEA, given that the score correlation coefficient between SPEA and SRRC is almost one. In fact these estimators implie a linear relationship between ranks which is not satisfied here. The score correlation coefficient between SMIR and PEAR for

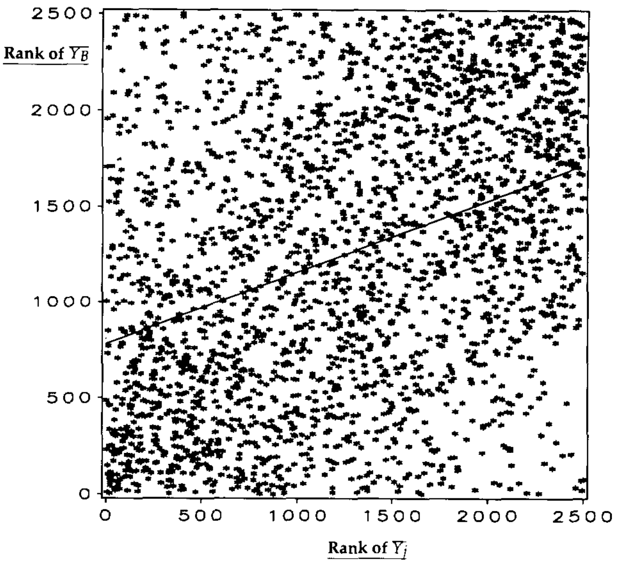

Fig. 17. Scatterplot of the ranked vectors $\overline{Y}_B$ and $\overline{Y}_j$ at time = 9×10^4.

the $t = 9 \times 10^4$ time points is 0.86, indicating that PEAR and SMIR are similarly sensitive to the upper tail of the output distribution. In fact the value of the R_y^2 coefficient based on the raw values for this time point is below 0.1 (Figure 9), indicating a poor performance of PEAR.

The above discussion points out that the only estimator whose use is legitimate for the Level E output 'dose at the time point' over all the time range is the HIM, either in its raw or rank version. The observation that HIM* gives rank = 1 to the variable FLOWV1 over the entire time span (Figure 16) is in agreement with our understanding of the model behaviour; FLOWV1 controls the transit time in the first segment of the geosphere and is in fact the most influential parameter.

The comparative effectiveness of HIM* versus SRRC/PRCC/SPEA can be visualised by comparing Figure 11 (rank of dose against rank of FLOWV1 scatterplot) with the new Figure 17. What is shown here is the scatterplot of the two ranked vectors $\overline{Y}_B$, $\overline{Y}_j$ used in the computation of the HIM* statistics for the same variable and time point as Figure 11. It is easy to see that the non-monotonic trend of Figure 11 becomes an evident monotonic trend in Figure 17. The thin line gives the linear regression on the points, which shows the positive correlation existing between the two vectors. This line should not be confused with the HIM* statistics; HIM* is not the correlation between $\overline{Y}_B$, $\overline{Y}_j$, but the sum of the pairwise products Rank(y_i) × Rank($y_{i,j}$), $i = 1, 2, \ldots NRUNS$, as described in Figure 7. It can be mentioned that in the process of the optimization of the HIM* estimator the $\overline{Y}_B$, $\overline{Y}_j$ rank correlation has also been attempted as a possible alternative to HIM*. This attempt resulted in accuracy values intermediate between HIM and HIM*, so that the pure rank based HIM* was finally retained.

8. Conclusions

An analysis has been made of the performances of selected SA estimators with respect to two characteristics: estimator reproducibility and estimator accuracy. It has been shown that although existing nonparametric techniques such as the SRRC, PRCC and SPEA are fairly reproducible and accurate when the model output varies linearly or at least monotonically with each independent variable their accuracy becomes dubious in the presence of model non-monotonicity. This poses a severe limitation to their application as the existence of non-monotonic relationships within the model cannot in general be determined a priori; a combined study of data scatterplots and of statistics such as the model coefficient of determination R_y^2 is needed to identify the problem. The modified Hora and Iman importance measure HIM*, on the contrary, appears capable of overcoming the difficulties posed by model non-monotonicity.

While HIM* demonstrates the existence of such methods, it is relatively expensive to apply. To carry out a HIM* analysis, even using the new technique described in Section 3, requires $N \times (1+K)$ simulations, where N is the sample size and K is the number of variables being analyzed. For models with large numbers of input parameters, where SA is most needed, the cost of performing the simulations can be prohibitive. This was not the case with the models employed in the present study.

Some preliminary results with IFFD were described in this paper because IFFD has been developed to analyze efficiently non-monotonic models with large numbers of variables. The results of the Level 0 analysis showed that IFFD had good reproducibility in that test case. It has not yet been applied to the Level E model to determine if it can deal effectively with the non-monotonic behaviour occurring there. The accuracy of IFFD shall be the object of further investigation.

References

[1] T.H. Andres. Statistical sampling strategies. In *Proceedings of Uncertainty Analysis for Performance Assessments of Radioactive Waste Disposal Systems*, an NEA workshop held in Seattle, February 24–26, 1987.

[2] T.H. Andres. User manual for SAMPLE. In preparation as a report for AECL, Pinawa, Canada R0E 1L0.

[3] T.H. Andres and W.C. Hajas. Sensitivity analysis of the SYVAC3-CC3 model of a nuclear fuel waste disposal system using iterated fractional factorial design. In preparation as a report for AECL, Pinawa, Canada R0E 1L0.

[4] S.G. Carlyle. The activities, objectives and recent achievements of the NEA Probabilistic System Assessment Codes User Group. In *Proceedings of Waste Management '87*, Ed. R.G. Post and M.E. Wacks, Tucson 1987.

[5] W.J. Conover, *Practical Non-parametric Statistics*. 2nd Edition (Wiley, New York, 1980).

[6] B.W. Goodwin, T.H. Andres, P.A. Davis, D.M. LeNeveu, T.W. Melnyk, G.R. Sherman and D.M. Wuschke. Post-closure environmental assessment for the Canadian nuclear fuel waste management program. *Radioactive Waste Management and the Nuclear Fuel Cycle*, **8** (2–3), 241–272 (1987).

[7] J.C. Helton, R.L. Iman, J.D. Johnson, and C.D. Leigh. Uncertainty and sensitivity analysis of a dry containment test problem for the MAEROS aerosol model. *Nucl. Sci. Eng.*, **102**, 22–42 (1989).

[8] S.C. Hora and R.L. Iman. A comparison of Maximum/Bounding and Bayesian/Monte Carlo for fault tree uncertainty analysis. *SANDIA Laboratory report SAND*85–2839, (1989).

[9] T. Homma and A. Saltelli. LISA package user guide. Part I. PREP (Statistical Pre-Processor) Preparation of input sample for Monte Carlo Simulation; Program description and user guide. CEC/JRC Nuclear Science and Technology Report EUR 13922/EN, Luxembourg 1991.

[10] R.L. Iman and W.J. Conover. The use of rank transform in regression, *Technometrics*, **21** (4) 499–509, (1979).

[11] R.L. Iman, J.M. Davenport, E.L. Frost, and M.J. Shortnecarier. Stepwise regression with PRESS and rank regression. Program User's guide. SANDIA National Laboratory report SAND 79–1472 (1980).

[12] R.L. Iman, J.C. Helton and J.E. Campbell. An approach to sensitivity analysis of computer models, Parts I and II. *Journal of Quality Technology*, **13** (3,4), 174–183 and 232–240, (1981).

[13] R.L. Iman and J.M. Davenport. Rank correlation plots for use with correlated input variables. *Comm. Statist. Simulation Comput.* **11**(3), 335–360 (1982)

[14] R.L. Iman and J.C. Helton. An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis*, **8**, 1, 71–90 (1988).

[15] R.L. Iman and J.C. Helton. A comparison of uncertainty and sensitivity analysis techniques for computer models. *Sandia Natl. Laboratories report NUREG/CR-3904, SAND* 84–1461, (1985).

[16] R.L. Iman, M.J. Shortencarier and J.D. Johnson. A FORTRAN 77 program and user's guide for the calculation of partial correlation and standardized regression coefficients. Sandia Natl. Laboratories report NUREG/CR 4122, SAND 85–0044 (1985).

[17] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis. *Japan Atomic Energy Research Institute report JAERI-M* 89–111, (1989).

[18] McGrow–Hill *Dictionary of Scientific and Technical Terms*, (New York) 1978.

[19] OECD – NEA, PSACOIN Level 0 Intercomparison. An international Code Intercomparison Exercise on a Hypothetical Safety Assessment Case Study for Radioactive Waste Disposal Systems. Prepared by A. Saltelli, E. Sartori, B.W. Goodwin and S.G. Carlyle. OECD –NEA publication, Paris (1987).

[20] OECD – NEA, PSACOIN Level *E* Intercomparison. An international Code Intercomparison Exercise on a Hypothetical Safety Assessment Case Study for Radioactive Waste Disposal Systems. Prepared by B.W. Goodwin, J.M. Laurens, J.E. Sinclair, D.A. Galson and E. Sartori. OECD – NEA publication, Paris (1989).

[21] P. Prado, A. Saltelli and T. Homma. LISA package user guide. Part II. LISA; Program description and user guide. CEC/JRC Nuclear Science and Technology Report EUR 13923/EN, Luxembourg (1991).

[22] P. Robinson et al.. Probabilistic System Assessment Codes Group; Level S; Specification for a sensitivity analysis exercise based on the level E test case. PSAC draft report, OECD – NEA, Paris, (February 1989).

[23] A. Saltelli, J. Marivoet. Performances of nonparametric statistics in sensitivity analysis and parameter ranking. CEC/JRC Nuclear Science and Technology Report EUR 10851 EN, Luxembourg (1986).

[24] A. Saltelli and J. Marivoet. Safety assessment for nuclear waste disposal. Some observations about actual risk calculations, *Radioactive Waste Management and the Nuclear Fuel Cycle*, **9** (4), (1988).

[25] A. Saltelli. The role of the code intercomparison exercise: Activities of the Probabilistic System Assessment Codes Group, in *Proceeding of the Ispra Course on Risk Analysis in Nuclear Waste Management*, May 30th–June 3rd, Kluwer Academic Publisher, Dordrecht, EUR 11969 EN, p. 69–95 (1989).

[26] A. Saltelli. Techniques for uncertainty and sensitivity analyses, in *Proceeding of the Ispra Course on Risk Analysis in Nuclear Waste Management*, May 30th–June 3rd, Kluwer Academic Publisher, Dordrecht, EUR 11969 EN, p. 129–160, (1989).

[27] A. Saltelli and T. Homma: LISA package user guide. Part III. SPOP; Uncertainty and sensitivity analysis for model output. Program description and user guide. CEC/JRC Nuclear Science and Technology Report EUR 13924/EN, Luxembourg (1991).

[28] A. Saltelli, T.H. Andres, B.W. Goodwin, E. Sartori, S.G. Carlyle and B. Cronhjort. PSACOIN Level 0 intercomparison; an international verification exercise on a hypothetical safety assessment case study, in *Proceedings of the Twenty-Second annual Hawaii conference on System Sciences*, Hawaii, January 3–6 (1989).

[29] A. Saltelli and T. Homma. Sensitivity analysis for model output. Performance of black box techniques on three international benchmark exercises. *Computational Statistics and Data Analysis* **13** (1) (1992) 73–94.

[30] A. Saltelli, J. Marivoet. Nonparametric statistics in sensitivity analysis for model output; a comparison of selected techniques, in *Reliability Engineering and System Safety*, **28**, 229–253 (1990).