Computational Statistics & Data Analysis 13 (1992) 73–94
North-Holland

Sensitivity analysis for model output

Performance of black box techniques on three international benchmark exercises

A. Saltelli

Commission of the European Communities, Joint Research Centre, Ispra, Italy

T. Homma

Japan Atomic Energy Research Institute, Tokai Research Establishment, Department of Environmental Safety Research, Tokai-Mura, Ibaraki, Japan

Received February 1990
Revised October 1990

Abstract: The paper analyses the difficulties of performing sensitivity analysis on the output of complex models. To this purpose a number of selected non-parametric statistics techniques are applied to model outputs without assuming knowledge of the model structure, ie as to a black box. The techniques employed are mainly concerned with the analysis of the rank transformation of both input and output variables (eg standardised rank regression coefficients, model coefficient of determination on ranks...). The test models taken into consideration are three benchmarks of the Probabilistic System Assessment Code (PSAC) User Group, an international working party coordinated by the OECD/NEA. They describe nuclide chain transport through a multi-barrier system (near field, geosphere, biosphere) and are employed in the analysis of the safety of a nuclear waste disposal in a geological formation. Due to the large uncertainties affecting the system these models are normally run within a Monte Carlo driver in order to characterise the distribution of the model output. A crucial step in the analysis of the system is the study of the sensitivity of the model output to the value of its input parameters. This study may be complicated by factors such as the complexity of the model, its non-linearity and non-monotonicity and others. The problem is discussed with reference to the three test cases and model non-monotonicity is shown to be particularly difficult to handle with the employed techniques. Alternative approaches to sensitivity analysis are also touched upon.

1. Introduction

The analysis of the sensitivity of the model response to the value of its input parameters is an essential element in the study of model performance. Such an analysis is part of model verification; for example, it helps to ensure that the coded version of the model works according to its specifications and that the

model responses to the variation in the input parameters is physically reasonable. Especially when the model has to be used for prediction under uncertainty, as, for instance, in the analysis of the environmental impact of pollutants, sensitivity analysis can rank the importance of the various uncertain parameters thus suggesting research priorities.

Another area of recent interest is the analysis of aggregation problems, where the degree of resolution of a computer model is calibrated against the desired level of resolution in the prediction. As an example, SA techniques can be used to optimize model gridding [1,2].

Sensitivity Analysis (SA) plays an important role in the stochastic computer codes used in Probabilistic System Assessment (PSA). These codes are usually run in a Monte Carlo fashion; the input sample consists of different sets (vectors) of input parameters. For each set the model is executed once, to produce a distribution of values for the output variable under consideration. SA attempts to determine which input variables are most important in causing the observed variation in the dependent variables.

Different SA techniques are described in the literature [3–5] and their relative performances have already constituted the object of intercomparison studies [6–8]. Non-parametric statistics based on ranks, such as the standardised rank regression coefficients and the partial rank regression coefficients appear to be among the most robust and reliable. Other non-parametric techniques such as the Smirnov test are also commonly used [9]. One of the main advantages of these techniques is that they do not require knowledge of the system structure, ie they can be applied to the model under consideration as to a black box, simply comparing model output with input. It is the purpose of this note to highlight the difficulties of the black box approach, with reference to some worked examples. The test models taken into consideration are three benchmarks of the Probabilistic System Assessment Code (PSAC) User Group [10–14], an international working party coordinated by the Nuclear Energy Agency of the Organisation for the Economic Cooperation and Development. These models analyse the performance of nuclear waste disposal in a geological formation. They are non linear, involve many uncertain parameters and have time dependent outputs whose variation covers orders of magnitude.

In section 2 a description of these models is given, together with the definition of the methods employed. Previous work on the use of non-parametric statistics in sensitivity analysis is also reviewed. The main results of the analysis are given in section 3. Some alternative approaches to SA are also discussed.

2. Models and methods

2.1 Test models

The PSAC group was established in early 1985 to coordinate the activities of teams involved in performance assessment using Monte Carlo codes. Up to now

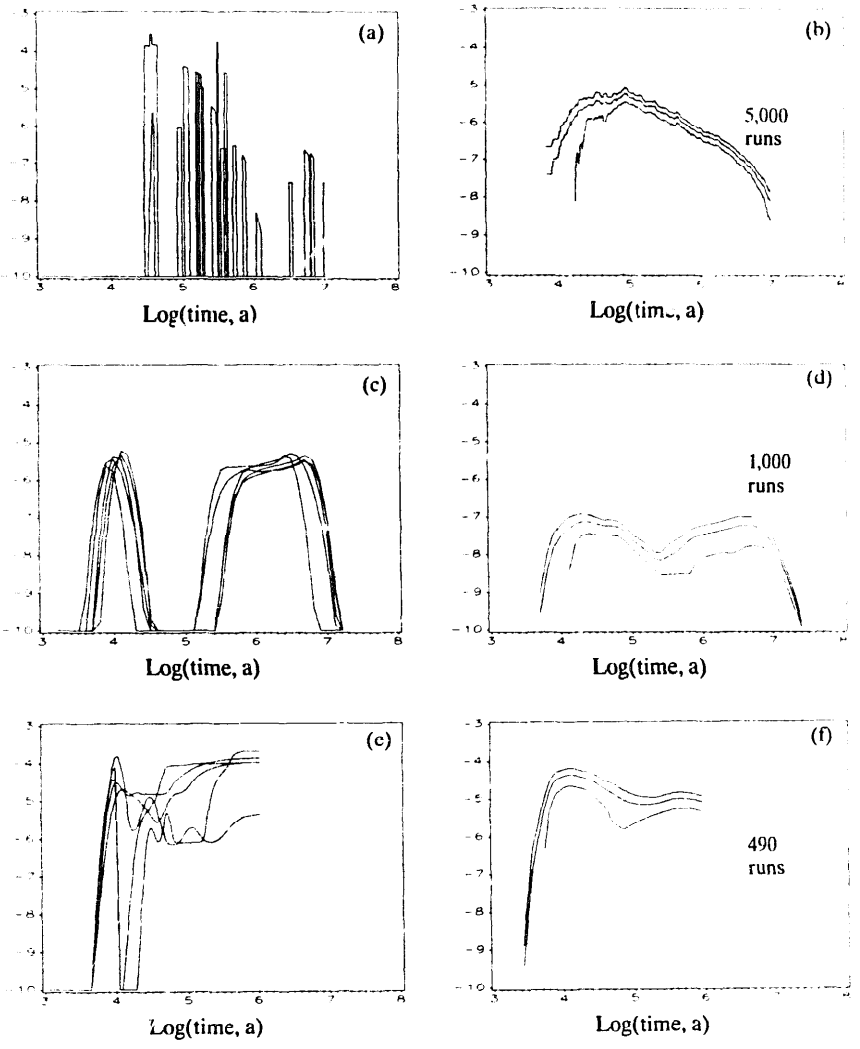

Fig. 1. Simulations highest outputs (left) and mean dose with Tchebycheff confidence bounds (right) for Levels 0 (a,b), E (c,d) and 1A (e,f). In all plots the ordinate axis is the logarithm of dose rate in Sv/a.

three full scale intercomparison exercises have been run by the group, on an increasing scale of complexity. The test models all involved the computation of the dose to man resulting from the migration of radionuclide chains or isotopes through a multi-barrier system including a nuclear waste repository (near field), a geosphere (far field) and a simplified biosphere.

In the Monte Carlo scheme the computation was repeated many times as to yield a distribution of the output under consideration (in this case dose at a time point). The objective of the analysis was to quantify the output distribution: output mean, confidence bounds and percentiles were sought. The sensitivity of the output to the model input parameters was also investigated. Figure 1 (a to f) offers a synoptic view of relevant results from the three benchmarks.

Table 1
Description of parameters to be treated as random variables in the Level 0 exercise

Notation	Definition	Distribution	Value	Units
RLEACH	leach rate	log-uniform	/0.00269, 12.9/	kg/m^2/a
XBFILL	buffer thickness	uniform	/0.5, 5/	m
XPATH	geosphere path length	uniform	/1000, 10,000/	m
V	ground water velocity	log-uniform	/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.	log-uniform	/2, 200/	m
ABSR	water extraction rate	uniform	/5.105, 5.106/	m^2/a
RMW	water ingestion rate	uniform	/0.7, 0.9/	m^3/a
BD(Cs)	sorpt. const. in the buffer	log-normal	mean = −0.46, std = 0.26	m^3/kg
BD(I)	sorpt. const. in the buffer	log-normal	mean = −5.07, std = 1.34	m^3/kg
BD(Pd)	sorpt. const. in the buffer	log-normal	mean = −1.91, std = 0.669	m^3/kg
BD(Se)	sorpt. const. in the buffer	log-normal	mean = −2.38, std = 0.143	m^3/kg
BD(Sm)	sorpt. const. in the buffer	log-normal	mean = −2.13, std = 0.605	m^3/kg
BD(Sn)	sorpt. const. in the buffer	log-normal	mean = −1.77, std = 0.729	m^3/kg
BD(Zr)	sorpt. const. in the buffer	log-normal	mean = −0.71, std = 0.5	m^3/kg
KD(Cs)	sorpt. const. in the geosph.	log-normal	mean = −1.46, std = 1.6	m^3/kg
KD(I)	sorpt. const. In the geosph.	log-normal	mean = −6.07, std = 2.6	m^3/kg
KD(Pd)	sorpt. const. in the geosph.	log-normal	mean = −2.91, std = 1.4	m^3/kg
KD(Se)	sorpt. const. in the geosph.	log-normal	mean = −3.38, std = 0.3	m^3/kg
KD(Sm)	sorpt. const. in the geosph.	log-normal	mean = −3.13, std = 1.2	m^3/kg
KD(Sn)	sorpt. const. in the geosph.	log-normal	mean = −2.77, std = 1.4	m^3/kg
KD(Zr)	sorpt. const. in the geosph.	log-normal	mean = −1.71, std = 1.0	m^3/kg

For the first exercise, named Level 0, the barrier submodels were extremely simple, the exercise being mainly meant to test the sampling subroutines, the executive (or driver) of the code and the code statistical post-processor [12]. Seven radionuclides, ^{135}Cs, ^{129}I, ^{79}Se, ^{151}Sm, ^{126}Sn and ^{93}Zr were considered in the exercise. All the barriers were described with very simple analytical equations; for instance, in the geosphere sub-model, the Gaussian transfer function corresponding to transport by advection and dispersion was simplified to a rectangular transfer function, the width of which simulates the effect of dispersion.

The Level 0 model considered 22 distributed parameters (ie parameters whose value is sampled for each run). The parameter characteristics are given in Table 1. The large range of variability can be noticed; uniform and normal distributions, on both linear and logarithmic scale are considered. Figure 1a shows the total dose (summed over all the nuclides) for the five runs yielding the highest output in a simulation composed of 5,000 runs. It can be seen that for many time points there is no output at all for any of the runs considered. This results in ties when the ranks of the output are computed, and complicates the sensitivity analysis (see next section). The mean dose originating from the same simulation is shown in Figure 1b, together with the 95th percent Tchebycheff's confidence bounds [15]. In spite of the large number of runs the output does not show a smooth profile.

Table 2
Description of parameters to be treated as random variables in the Level E exercise

Notation	Definition	Distribution	Value	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

The second PSAC exercise was characterised by a more complex geosphere sub-model: two layer, mono-dimensional, including dispersion, advection, decay and chemical retention for nuclide chain. The geosphere sub-model was dealt with numerically by most of the participating PSA codes. For this exercise a quasi-analytical solution is available for the stochastic output, based on an accurate numerical inversion of the exact solution in the Laplace space [16]. For this reason the exercise was named Level E (after Exact). Again ^{129}I is considered, together with the ^{237}Np-^{233}U-^{229}Th chain [13]. Twelve distributed parameters were considered (Table 2). It can be seen that the range of parameter variation is somewhat less severe than for Level 0. Only uniform type distributions (linear or logarithmic) are considered. Typical model output are shown in Figure 1c. The spread in the results among the various runs is less pronounced than for the previous exercise; the two separate peaks identify the Iodine and the Np chain doses. The dose mean, relative to a simulation of size 1000, has a more regular shape (Figure 1d).

In the third exercise, Level 1A/14/, the complexity of the near field sub-model was increased, including a solubility limited release from the repository vault, while the far field remains substantially unchanged. Furthermore, for this exercise, the model specification did not include mathematical expressions, letting the user decide on how to interpret the description of the problem. Dose computations were sought for the nuclides ^{14}C, ^{59}Ni, ^{79}Se, ^{99}Tc, ^{129}I and for the chains ^{237}Np-^{233}U-^{229}Th and ^{238}U-^{234}U-^{230}Th-^{226}Ra-^{210}Pb. The presence of

several isotopes of the same element made the computation of the solubility limited transport in the vault more difficult to handle.

Input data for the exercise are given in Table 3(a and b). Fifteen nuclide independent parameters plus 40 element-specific parameters make a total of 55 random variables for the exercise. Uniform, log-uniform, normal and log-normal distributions are considered. The spread in the input data is comparable to that of the Level 0 exercise.

Due to the higher number of isotopes and to the increased complexity of the vault submodel, Level 1A was more computer time consuming than the other two exercises. Some high dose outputs for a simulation of 490 runs are given in Figure 1e, and the output mean in Figure 1f.

Neglecting the effect of solubility limits, considered only in the last exercise, the model output for the three test models is roughly given by

$$Dose(t) = (\Sigma_i Release_i(t)) \times Dilution,$$

where the summation is extended to all the nuclides, *Dose* is the total dose rate, *Dilution* is a dilution factor related to the biosphere compartment and *Release* is generally a pulse function whose position on the time axis is determined by the nuclides transit time in the barrier system. *Dilution* is roughly equal to 1/ABSR

Table 3a
Description of non-element specific parameters to be treated as random variables in the Level 1A exercise. The format of data input is different from that adopted for Levels 0 and E. Values A and B below refer to the extremes of the distributions for uniform and log-uniform data, and to the 0.001–0.999 quantiles for the normal and log-normal distributions. Mean μ and standard deviation a can be obtained as $\mu = (A+B)/2$, $\sigma = (B-A)/6.18$

Notation	Definition	Distribution	Value A	Value B	units
TDRUM	maximum container lifetime	normal	100	500	a
TMATR	matrix degradation time	uniform	200	400	a
PORVLT	effective porosity in vault	uniform	0.10	0.20	–
PORL1	effective porosity in first layer	uniform	0.05	0.10	
PORL2	effective porosity in second layer	uniform	0.20	0.25	–
VDCYL1	Darcy velocity in first layer	uniform	0.001	0.01	m/a
VDCYL2	Darcy velocity in second layer	uniform	0.1	1	m/a
DCYVLT	ratio of Darcy velocity in vault to that of first layer	log-uniform	0.01	2	–
LPATH1	pathlength of first layer	normal	90	110	m
LPATH2	pathlength of second layer	normal	17,000	23,000	m
CDIF1	effective diffusion coefficient for first layer	normal	0.002	0.005	m^2/a
CDIF2	effective diffusion coefficient for second layer	normal	0.005	0.02	m^2/a
LDISP1	longitudinal dispersion length in first layer	uniform	1	10	m
LDISP2	longitudinal dispersion length in second layer	uniform	50	500	m
FDILUT	dilution factor	log-uniform	10^{-5}	10^{-3}	–

Table 3b
Description of element specific parameters to be treated as random variables in the Level 1A exercise. The range limits A and B (see previous table) for the solubility limits are obtained by multiplying the values below with 1/100 and 100 respectively. Similarly the range limits A and B for the exchange constants K_D's are computed by multiplying the values below with 1/10 and 10

Element	K_D values (litre/kg)			Solubility limits (mol/litre)
	Vault	First layer	Second layer	
C	0.5	1	0.5	$1.0\,10^{-5}$
Ni	5	1	5	$1.0\,10^{-5}$
Se	1	1	1	$1.0\,10^{-2}$
Tc	0.1	0.5	0.5	$1.0\,10^{-5}$
I	0	0	0	1.0
Np	50	10	10	$1.0\,10^{-7}$
U +	10	5	1	$1.0\,10^{-5}$
Th	1,000	500	1,000	$1.0\,10^{-8}$
Ra	1	1	5	$1.0\,10^{-4}$
Pb	5	10	100	$1.0\,10^{-5}$

in Level 0, 1/STFLOW in Level E and *FDILUT* in Level 1A. Almost all the other uncertain parameters considered in the three exercises enter into the equations which govern the transit time (path lengths, flow velocities, retention coefficients,...); the the dependence of *Dose* upon these parameters is strongly nonlinear. A purely linear relationship exists between *Dose* and *FDILUT* in Level 1A; a linear correlation exists between *Dose* and both *ABSR* and *STFLOW* (Levels 0 and E).

2.2. Methods

Sample generation and computation. Random sampling has been systematically used for all the test cases discussed here, although Latin Hypercube Sampling was also used for comparison. The computations have been performed using different versions of the LISA code [17–18] and of its statistical post-processor SPOP [19].

Sensitivity analysis techniques. In a previous paper ([8], see section 2.3) some parametric and non-parametric sensitivity analysis techniques were intercompared using selected outputs from the Level 0 benchmark as a test case. The non-parametric techniques which proved there to be the most robust have been applied here. These are the standardized regression coefficient (on ranks) and the partial rank correlation coefficient [20]. The Spearman coefficient and the Smirnov test are also considered, together with a few parametric tests. A short description of these tests is given here for convenience.

The Pearson product moment correlation coefficient (*PEAR*, in the following) is the usual linear correlation coefficient computed on the x_{ij}, y_i's ($i = 1,2 \ldots N$), where x_{ij} is the sampled value for the variable X_j in the run i and y_i is the corresponding value of the output variable Y. For non-linear models the Spear-

man coefficient (*SPEA*) is preferred as a measure of correlation, which is essentially the same as *PEAR*, but using the ranks of both Y and X_j values instead of the raw values [9]:

$$SPEA(Y, X_j) = PEAR(R(Y), R(X_j)).$$

The basic assumptions underlying the Spearman test are:

(a) Both the x_{ij} and the y_i are random samples from their respective populations.
(b) The measurement scale of both variables is at least ordinal.

SPEA can be used for hypothesis testing [9].

Partial Correlation Coefficients (*PCC*, in the following) and Standardized Regression Coefficients (*SRC*) are correlation estimators which can also be used on the ranks of the (Y, X_j) values (Partial Rank Correlation Coefficients *PRCC* and Standardized Rank Regression Coefficient *SRRC*) [20]. The $SRC(Y, X_j)$ are the coefficients of the regression model for Y; they provide an approximation to Y in the form:

$$Y^* = \sum_{j=1}^{K} SRC(Y, X_j) X_j^*,$$

where X_j^* are the normalised variables

$$X_j^* = \frac{(X_j - \bar{X}_j)}{S(X_j)},$$

and $\bar{X}_j$ and $S(X_j)$ are respectively the sample mean and standard deviation.

When using the *SRC*'s it is also important to consider the model coefficient of determination R_y^2. R_y^2 provides a measure of how well the linear regression model based on SRC's can reproduce the actual output vector Y. In particular:

$$R_y^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the output values y_i and the $\hat{y}_i$ are the model prediction based on the *SRC*'s, so that R_y^2 represents the fraction of the variance of the output vector explained by the regression. The closer R_y^2 is to unity, the better is the model performance.

The coefficients $SRC(Y, X_j)$ can themselves provide a very effective measure of the relative importance of the input variables. Of course the validity of the *SRC*'s as a measure of sensitivity is conditional to the degree to which the regression models fits the data, i.e. to R_y^2.

When the value of the R_y^2 coefficient computed on the raw values is low, it is usually worth trying the rank equivalent of *SRC*, i.e., the *SRRC*. These are simply obtained by replacing both the output variable values y_i and the input

vectors X_i's by the respective ranks. If the new R^2 coefficient (on rank) is higher, then the *SRRC* can be used for *SA*. It should be clear that a successful regression analysis based on the *SRC*'s must be preferred to one based on the *SRRC*'s: in the example of the present study a regression model for *Dose* is more informative than a regression model for *Rank of Dose*. Yet when the regression on the raw data is poor, a rank transformation might be appropriate. All that is needed for the purpose of *SA* is that the regression model which is eventually used to rank the data is an effective one. This can be verified by looking at the R^2_y value; other regression statistics such as the Predicted Error Sum of Squares (PRESS, [21]) might be used to ensure that the regression model is not overfitting the data. PRESS is especially useful when performing stepwise regression to discriminate between competing models. For a discussion of raw and rank regression see also Appendix in [1]. A comparison between rank regression and other techniques is given in [22].

The *PCC* can be considered as an extension of the usual correlation coefficients and represents that part of the interdependence between two variables which is not due to correlation between these two variables and the remaining ones. When *PCC*'s are used they can provide a ranking of the various variables by indicating the strength of the linear relationship between Y and X_j. When *PRCC*'s are used, the linear relationship between the ranks of Y and X_j is measured. This gives an effective estimation of sensitivity.

The Smirnov test [9] (*SMIR* in the following) is a "two-sample" test originally designed to check the hypothesis that two different samples belong to the same population. The application of a "two-sample" test to sensitivity analysis comes from the idea of partitioning the sample of the parameter X_j under consideration into two sub-samples according to the quantiles of the output (Y) distribution. If the distributions of X_j in the two sub-samples can be proved to be different then the parameter under consideration is recognized as influential. For instance, the values x_{ij}'s corresponding to output y_i's above the 90th quantile of the $F(Y)$ distribution may constitute one sub-sample, and all the remaining x_{ij}'s the other sub-sample.

For *SMIR* to be applicable, the following assumptions must be satisfied by the two sub-samples under consideration, viz:

(a) the two sub-samples are random samples;
(b) the two sub-samples are mutually independent;
(c) the measurement scale is at least ordinal;
(d) the random variables must be continuous.

When using *SMIR* the empirical cumulative distributions $F(X_j)$ are computed on the two samples and the two distributions compared with each other. If the two distributions are different, it can be said that the parameter influences the output, and that high outputs are preferentially associated with high, or low, parameter values. More quantitatively the Smirnov statistic is defined as the maximum vertical distance between the empirical cumulative distribution functions of the two samples. *SMIR* can be used for hypothesis testing [9].

An important observation has to made here. The use of *PRCC* and *SRRC* is somewhat redundant for the test cases under consideration. In fact these two techniques usually produce identical variable ranking. To take an example, referring to the Level 0 test case (Table 1), if variable V is ranked one (most important) and variable $KD(Sm)$ is ranked 22 (least important) by the *SRRC*'s, an identical ranking is provided by the *PRCC*'s (see reference [20]). *PRCC*'s and *SRRC*'s rankings only differ when significant correlations are involved among the input variables, which is not the case for any of the three benchmarks. Furthermore, *SRRC*'s predictions are also strongly correlated to those of *SPEA* and *SMIR* ("score" correlation coefficients for the couples *SMIR/SRRC* and *SPEA/SRRC* are in general close to 0.8 or higher [8]). For this reason the R_y^2 coefficient plays a crucial role, as it indicates the degree of reliability of the *SRRC*'s as well as of the other techniques. In other words, when R_y^2 flags an inadequate regression model, and hence a failure of the *SRRC*'s, also the ranking provided by *SPEA* and *SMIR* should be regarded with suspicion.

2.3. Previous work: Intercomparing sensitivity analysis techniques

Iman and Helton [6,7] have analysed the performances of a number of sensitivity analysis techniques for model output, including

- Response Surface Replacement, used in conjunction with Fractional Factorial Design,
- Differential Analysis,
- Partial Rank Correlation Coefficient in conjunction with Latin Hypercube Sampling (LHS).

Their conclusions were that the non-parametric techniques used in conjunction with LHS are the more robust, being able to cope with model non-linearity (better than in the case of fractional design) and to scan all the space of the input variables (differential analysis only provides information around a point in the space of the variables). For a broader discussion of Monte Carlo based SA see also final section of [1].

An intercomparison of non-parametric statistics can be found in [8], where the Level 0 test model was also used. A test of the various SA techniques was made by analysing the fluctuations in the prediction of the various tests when repeating the sensitivity analysis on different samples. The main finding of this study was that the *SRRC*'s and the *PRCC*'s were the most stable estimators, followed by *SPEA* and *SMIR*. The predictions from the linear estimators (e.g., *PEAR*) were considerably more erratic.

3. Results of SA for the test models

3.1. Analysis of the model coefficient of determination

As mentioned in the previous section, the value of the model coefficient of determination is crucial to the interpretation of the results, as it provides a

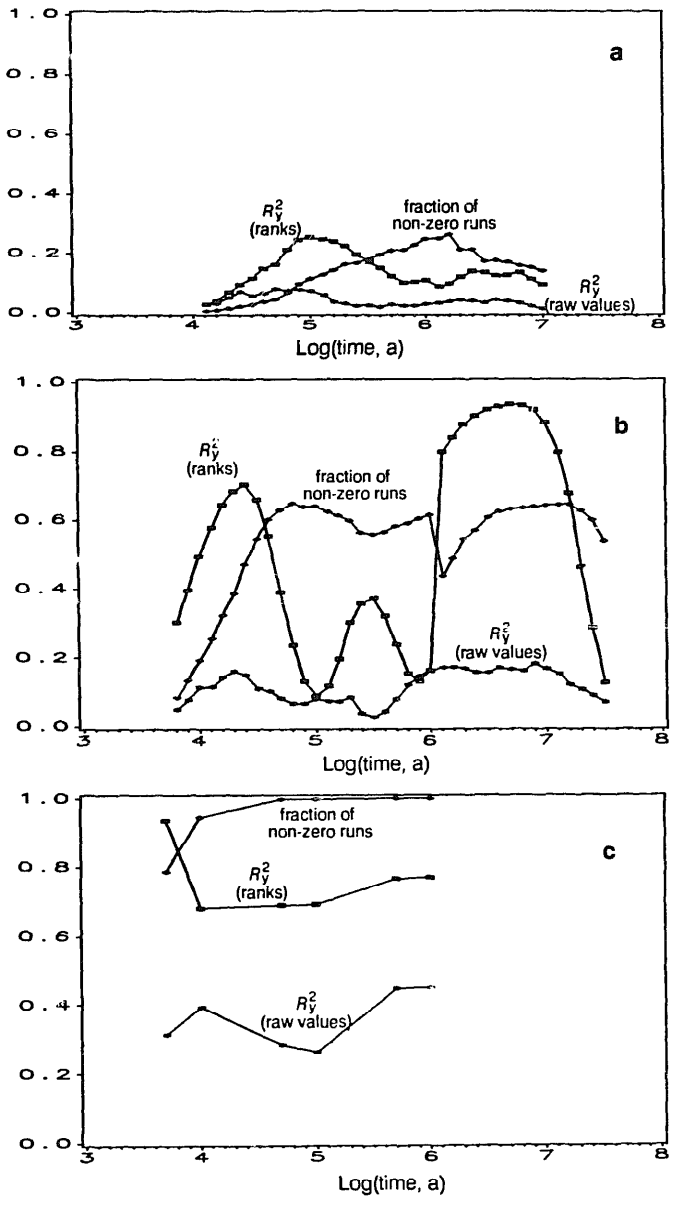

Fig. 2. Model coefficient of determination (on raw value and ranks) and percentage of non-zero runs for Level 0 (a), Level E (b) and Level 1A (c).

measure of the adequacy of the regression model based on the ranks and thus an indication of the performances of the non-parametric tests. The R_y^2 coefficient can also be computed on the raw values, ie on the *SRC*'s instead of on the *SRRC*'s, thus providing an indication on the performances of the linear correlation/regression techniques (*PCC, SRC, PEAR* ...).

The model coefficients of determination for the output variable "dose at the time point" has been plotted in Figure 2(a to c) as function of time. Values of R_y^2

close to one indicate a good performance of the regression model. Three quantities have been plotted:

- the R_y^2 values based upon the dose rate *raw values* (R_y^2 based upon the *SRC*'s)
- the R_y^2 values based upon the *ranks of the input value* (R_y^2 based upon the *SRRC*'s);
- the percentage of non-zero output for each time point.

All the results have been obtained using the same simulations of Figure 1 (5,000 runs for Level 0, 1,000 for Level E and 490 for Level 1A).

For the Level 0 case the percentage of non zero runs never exceeds 27% and is as low as a few percent for the lowest time point. The model coefficients of determination are also very low, never exceeding 0.26 for the regression based on the ranks. R_y^2 values for the regression based upon the raw values are even lower, indicating that a sensitivity analysis based upon a linear regression technique is not really worth being pursued. It is quite difficult in this context to establish the relative importance of the input parameters; if all the parameters taken together account for only 26% of the data variance, it may not be worthwhile to determine how much variance each of them can account for individually. Thus, Figure 2a suggests that the *SRRC*'s should not be used to rank the Level 0 input parameters for the output variable under consideration. Because of the correlation among non-parametric techniques mentioned above also *SMIR* and *SPEA* prediction should be considered as nonreliable. Also interesting in Figure 2a is the non-monotonic trend of the R_y^2 estimator, which passes through a minimum ($R_y^2 = 0.08$) in correspondence to the $t = 10^6$ time point.

Figure 2b displays the results for the Level E exercise (same statistics and scales as Figure 2a). The percentage of non-zero runs is much higher in this case (0.5–0.6 for most of the time span). The R_y^2 on ranks exhibits a multi-modal pattern which appears completely un-correlated with the percentage of non-zero runs, and passes through local minima where R_y^2 is close to 0.1. The R_y^2 on the raw values is, even in this case, very low.

The results for the Level 1A (Figure 2c, same scales) are qualitatively different from those of Levels 0 and E. First, the percentage of non-zero runs is always close to one. This is due to the solubility limited release from the vault, which causes the output fluxes from geosphere to be broad smooth pulses rather than sharp peaks. The R_y^2 on ranks is quite high on all the time range explored (larger than 0.6). Also the R_y^2 on raw values are higher than from the two other test cases.

3.2. *Variable ranking as function of time*

For each time point considered in the analysis the *SA* statistics (*PEAR* through *SMIR*) have been applied, producing a variable ranking. Normally different statistics produce different rankings. Linear techniques tend to give lower rank (most important variables) to parameters having a linear influence on the output, such as the biosphere dilution factors (*ABSR* in Level 0, *STFLOW*

in Level E and FDILUT in Level 1A). These variables roughly represent a flow by which the output concentration in the geosphere must be divided before it is converted to dose. They can change linearly the output by one to two orders of magnitude (see section 2). The nonparametric tests instead give the lowest ranks to the variables capable of producing a dose at that time point, such as the variables which govern the nuclide transit time (water velocity, path length, retention). These parameters can change the output by many orders of magnitude, especially for Levels 0 and E, but are not always detected as influential by the linear statistics (*PEAR*) as these latter concentrate on the upper tail of the output distribution.

For each variable the ranking produced by the different estimators can be plotted as a function of time. In Figure 3 selected results from the Level 0 exercise have been given. The ordinate axis represents the rank given by the technique (eg PRCC) to the variable (rank 1 = most important variable, rank 9 = least important variable). The ranking produced by *SMIR, SPEA, SRRC* and *PRCC* for the three most important variables shows that the disagreement among techniques is higher for the lowest R_y^2 point ($t = 10^6$). It is puzzling at first that for this time point the variable "dispersivity in the geosphere" (*ADISPG*) becomes more important than the "water velocity in the geosphere" (*V*) as indicated by all the techniques (see next section). It should be kept in mind that the rankings of Figure 3 are of little use, at least as far as *SPEA, PRCC* and *SRRC* are concerned, because of the scarce predictive ability of the regression model (Figure 2a).

A similar trend appears for the Level E results (Figure 4). Here the two spikes at $t = 10^5$ and $t = 10^6$ for the variable *FLOWV1* (water velocity in the first geosphere layer) closely correspond to the two local minima of R_y^2 (Figure 2b). These results are also somewhat suspect, as they show that a variable which influences the output linearly (*STFLOW*) becomes more important than the variables linked to the geosphere transit time.

For the Level 1A exercise the transit time is less important than for the two others; nuclides reach the biosphere as broad pulses rather than as sharp peaks (compare Figure 1e with 1a). This explains why the dilution factor at the geo/biosphere interface remains the dominant parameter for all the time span explored (*FDILUT* variable, Figure 5), followed by the Darcy velocities in the three sections which constitute the pathway (vault plus two geosphere layers; variables *DCYVLT, VDCYL1* and *VDCYL2*). Nothing pathological shows up in this figure, in agreement with the regular trend of R_y^2 (Figure 2c).

3.3. Analysis of the input / output scatterplot

When performing sensitivity analysis, the variable/variable scatterplots provide a valuable information. In Figure 6 the rank of the output dose has been plotted against the rank of the variable *V* (Level 0 exercise) for three different time points. It can be recalled that this variable is generally the most influential, except for strange discontinuity at $t = 10^6$ a (Figure 3). This corresponded to a

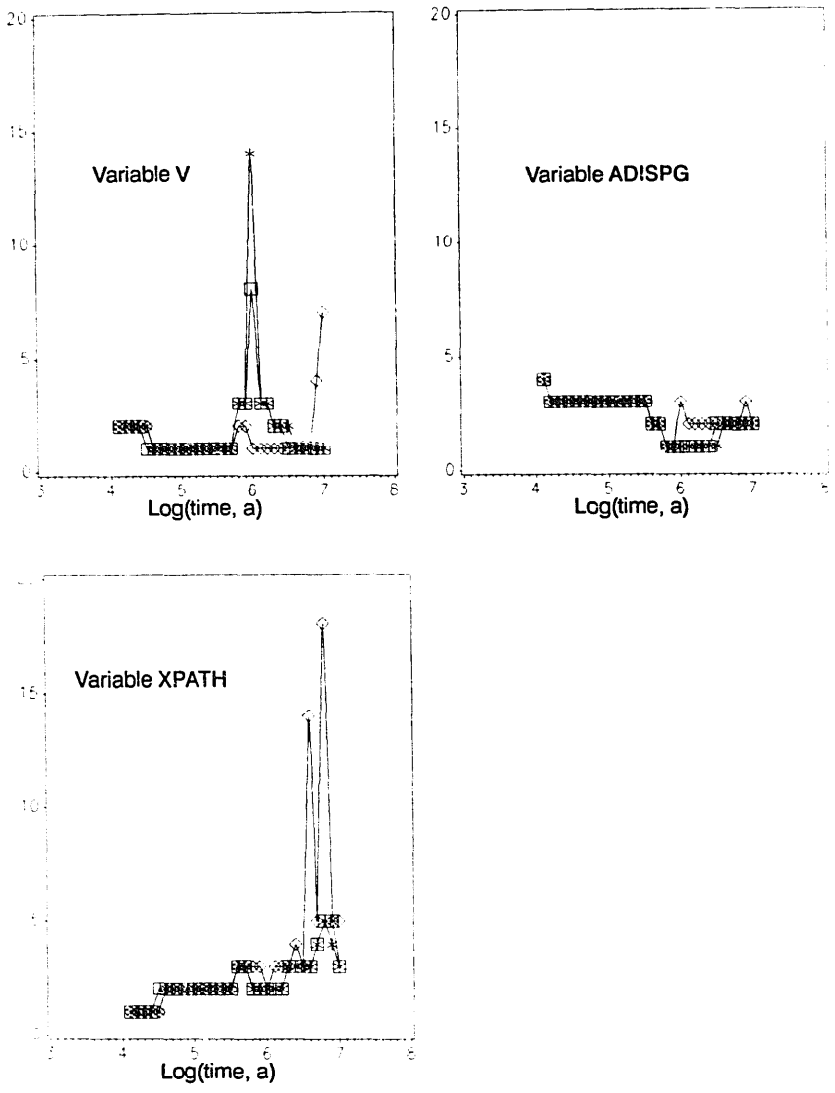

Fig. 3. Level 0 exercise. Variable ranking for three influential variables. Diamond = Smirnov test; Asterisk = Spearman coefficient; Square = *PRCC* and *SRRC* (always superimposed). The ordinate axis is the variable rank (= 1 for the most important variable; = 22 for the least important one).

minimum of the R_y^2 (Figure 2a). Both these facts can be easily explained by looking at the three plots in Figure 6.

For early time ($t = 10^5$) only high V values can generate a non-zero output, so that a linear positive correlation exists between rank of dose and rank of V (upper-left plot in Figure 6). For late times ($t = 10^7$ a) high V values result in zero dose, as the nuclide peak has arrived the biosphere earlier, i.e., only low V values are associated to non-zero runs; hence the negative slope of the rank regression line in the bottom plot of Figure 6. For the intermediate time point $t = 10^6$ the situation is in between (rightmost plot in Figure 6) and the input/output relationship is non-monotonic. This would be even more evident if one could

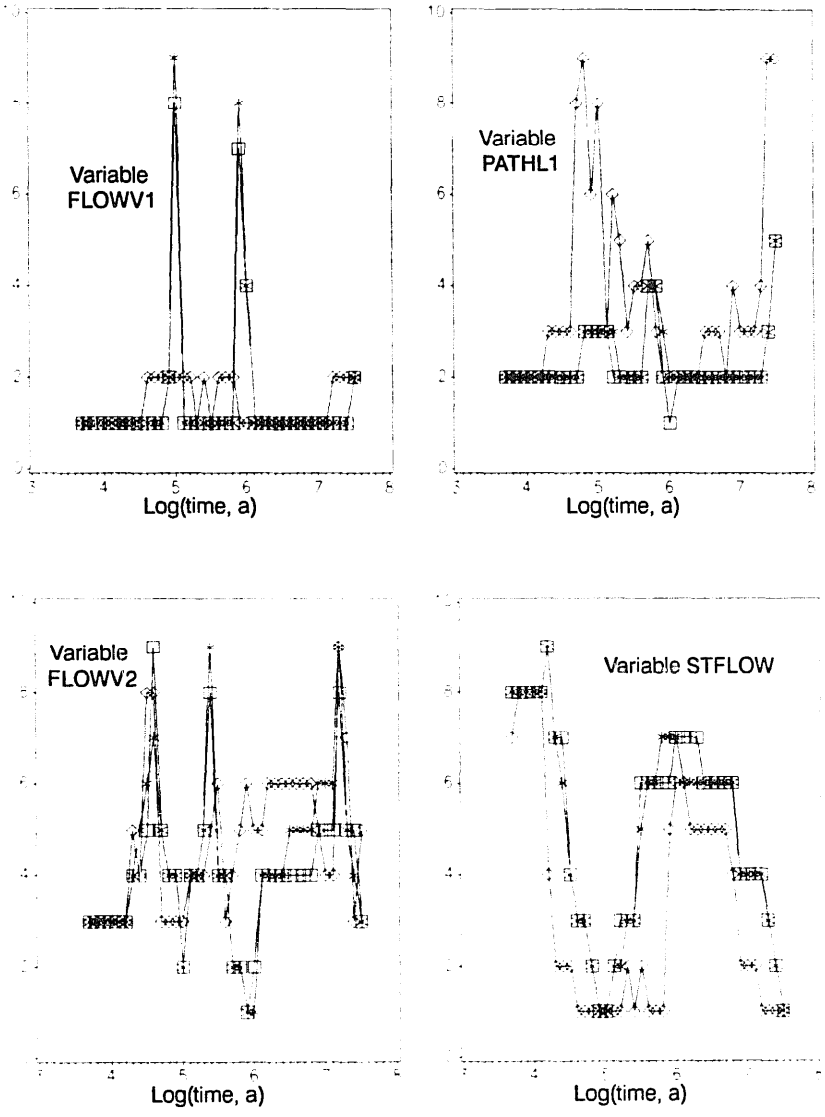

Fig. 4. Level E exercise. Variable ranking for four influential variables Diamond = Smirnov test; Asterisk = Spearman coefficient; Square = *PRCC* and *SRRC* (always superimposed). The ordinate axis is the variable rank.

count the ties in the figure: there are many more zero outputs at the extreme of this plot than in its middle.

It is well known that rank-based non-parametric techniques are able to handle model non-linearity, by linearising the input/output relationship [5]. This requires the relationship to be monotonic. When this is not the case, as shown in Figure 6, the rank techniques are doomed to fail. In other words both the scatterplots in Figure 6 and the low values of the R_i^2 coefficient in Figure 2a reveal that this model is more complicated than the linear rank regression model that is being used to approximate it; no automated "black box" SA of the Level 0 model can be achieved using these techniques.

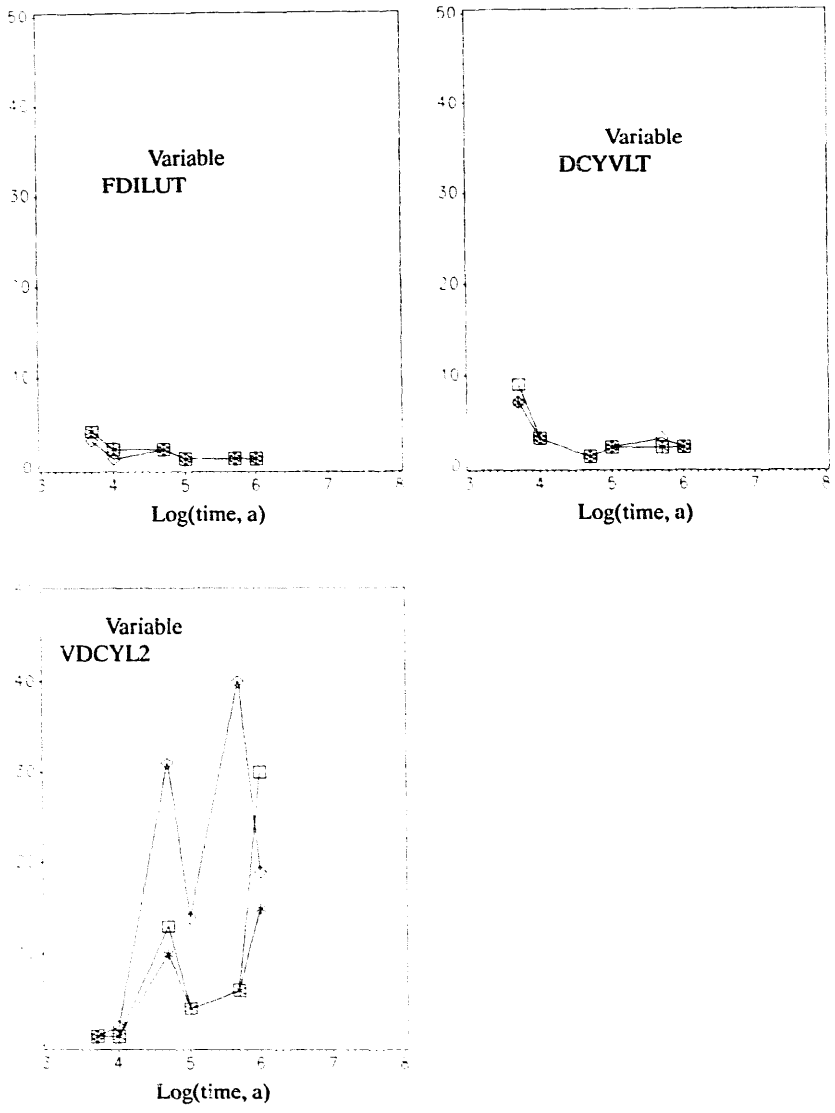

Fig. 5. Level 1A exercise. Variable ranking for three influential variables. Diamond = Smirnov test; Asterisk = Spearman coefficient; Square = *PRCC* and *SRRC* (always superimposed). The ordinate axis is the variable rank.

This pattern is even more evident for the Level E results. The plots of Figure 7 are rank scatterplots (as in Figure 6) for the most influential variable (*FLOWV1*) at three time points. The three plots span the area corresponding to the first local minimum of the R_y^2 curve in Figure 2b, and show clearly the non-monotonicity of the input/output relationship at the $t = 10^5$ time points. Also the second local minimum of R_y^2 at $t = 10^6$ in Figure 2b can be shown to depend upon model non-monotonicity.

It can be mentioned that in the Level E example the zero outputs (the ties in Figure 7) are generated by a dose cut off control in LISA which sets dose rates below 10^{-15} Sv/a to zero. Even when this censoring effect is removed (bypassing

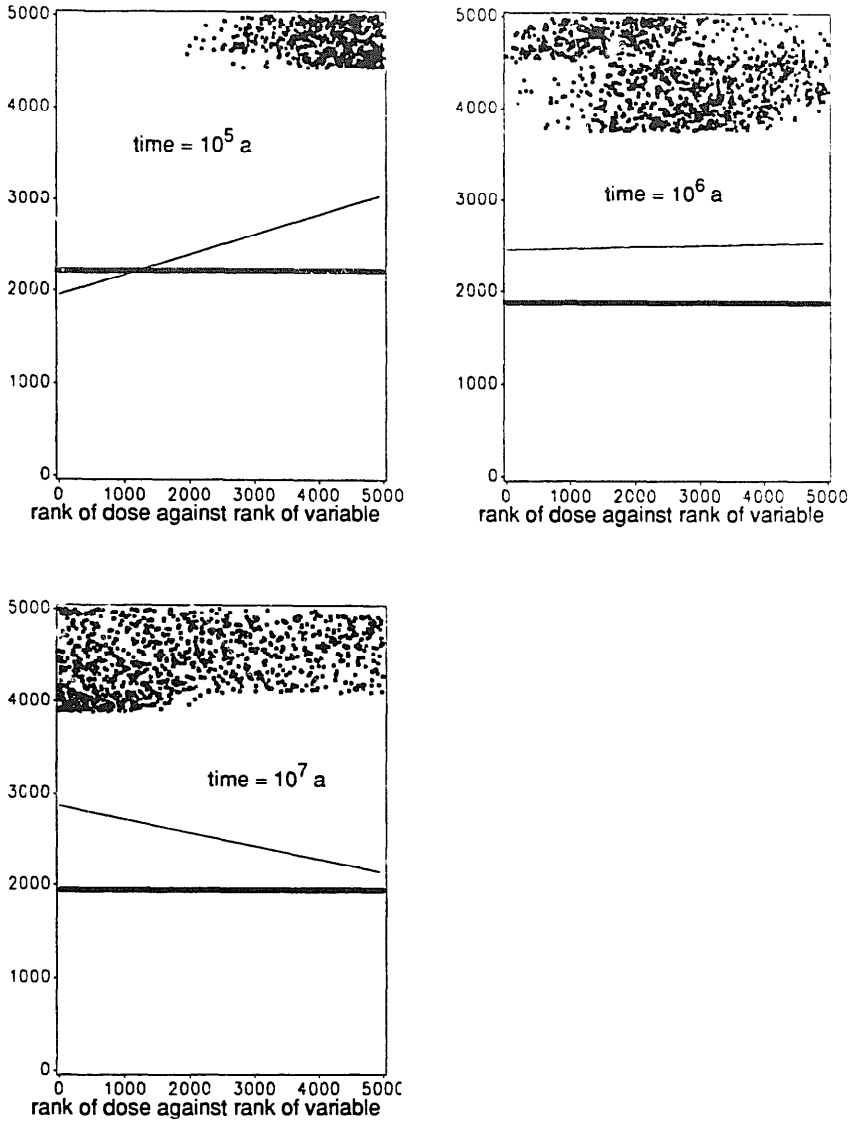

Fig. 6. Level 0 exercise. Input output scatterplots for the variable V; The rank of the output dose at three different time points is plotted against the rank of the variable V. The thick horizontal line represents the contribution of the observations that had no discharge and hence no dose; these result in ties when the ranks are taken; for instance, if out of 5,000 runs 4,400 yield zero dose, all these are given the average (tie) rank 2,200 (upper-left figure). The thin line represents the linear regression over all the ranks (non-zero plus ties).

the cut-off control and hence eliminating the ties in Figure 7) still the scatterplot at $t = 10^5$ a exhibits a bell shape and the regression line is horizontal.

When the variable *FLOWV1* fails to be identified as influential due to the model non-monotonicity, then the *STFLOW* variable, which affects the output linearly, may show up as influential as can be seen from the last plot of Figure 4. This should be considered as an artifact of the data, as the variables which govern the transit time in the geosphere are surely the most influential in the Level E test model, as shown by the scatterplots.

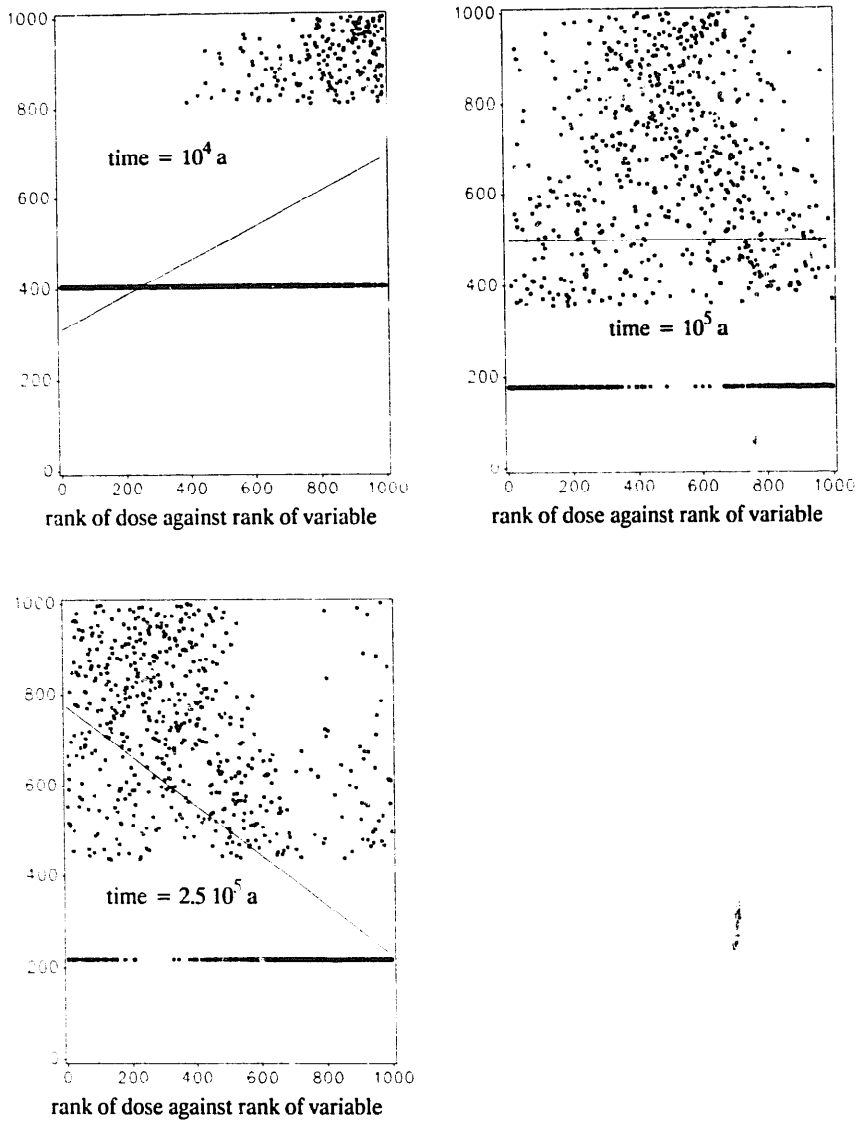

Fig. 7. Level E exercise. Scatterplots of the rank of the output dose against the rank of the variable *FLOWV1* at three different time points. As in Figure 6 the thick horizontal line represents the contribution of the ties and the thin line represents the linear rank regression.

The situation is quite different for the Level 1A exercise. Here the *FDILUT* variable (dilution factor at the geo/biosphere interface) is really the most influential one for the entire time span considered (Figure 5). The most influential geosphere parameter (the Darcy velocity in the second layer *VDCYL2*) is only influential at short times (see scatterplots in Figure 8), when high doses are only associated to high values of the Darcy velocity. For longer times (second and third plot in Figure 8) doses are insensitive to the *VDCYL2* values due to the extremely slow release from the vault, so that the only parameters which matter are *FDILUT* and *DCYVLT* which also control the release rate from the vault.

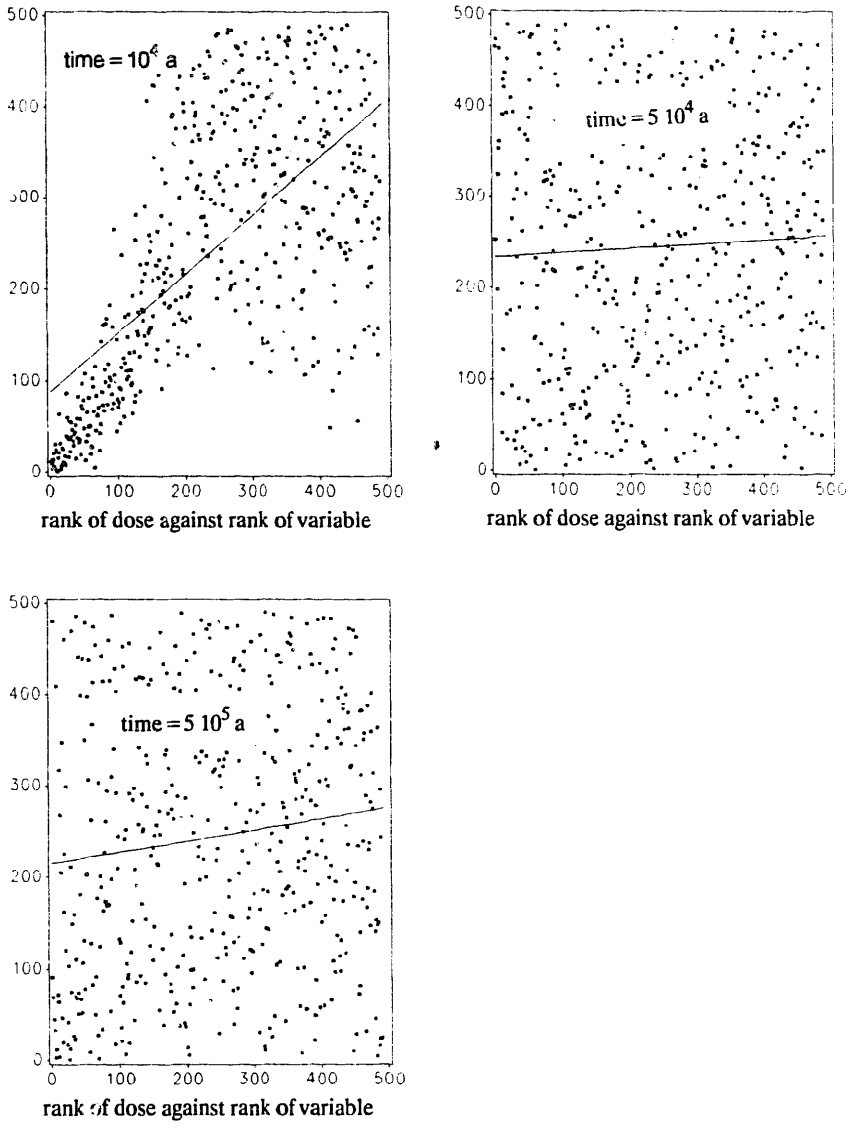

Fig. 8. Level 1A exercise. Scatterplots of the rank of the output dose against the rank of the variable *VDCYL2* at $t = 10^4$, $5\ 10^4$ and $5\ 10^5$ a. The thin line represents the linear rank regression. There appear to be no zero discharges.

3.4. An alternative approach

The "Importance Measure" proposed by Hora and Iman [23] is based on the concept of uncertainty reduction. It was applied by Ishigami and Homma to the Level 0 exercise [24]. The idea behind this method is that the variance V_Y of the output variable Y can be reduced by fixing the value of any generic input variable X_j to a constant value x_j. Since this conditional variance $V_Y(x_j)$ depends on the selected x_j value, $V_Y(x_j)$ should be averaged according to the distribution function of X_j, to obtain

$$V_Y^j = \int V_Y(x_j) f_j(x_j) \mathrm{d}x_j,$$

Table 4
Ranking of Level 0 three most important parameters as function of time based on the variance reduction technique

Time Rank	$10^{4.5}$	10^{5}	$10^{5.5}$	10^{6}	$10^{6.5}$	10^{7}
1	*XPATH*	*V*	*V*	*V*	*ADISPG*	*ADISPG*
2	*V*	*ABSR*	*ABSR*	*ABSR*	*V*	*V*
3	*ABSR*	*DIFFG*	*XPATH*	*RLEACH*	*RLEACH*	*DIFFG*

where f_j is the density function for X_j. Then the importance measure is defined as [23].

$$I_j = V_Y - V_Y^j.$$

With some manipulation I_j can be described as

$$I_j = U_j - \langle Y \rangle^2,$$

where $U_j = \int \langle h(x_j) \rangle^2 f_j(x_j) \mathrm{d}x_j$ and $\langle h(x_j) \rangle$ denotes the mean of Y obtained by fixing X_j to the specific value x_j.

In [24] a computationally efficient method is presented to compute the above statistic. The three most important variables from the Level 0 exercise have been determined on the basis of this importance measure (Table 4). This table shows that the variable ranking produced by the uncertainty reduction method exhibits the same trend of the estimators shown in Figure 3. This ranking is physically reasonable because the variable "geosphere path length (*XPATH*)" is more important at early times and the "water velocity in the geosphere (*V*)" is constantly important at all the time points. As expected, the variable "dispersivity in the geosphere (*ADISPG*)" becomes important at late time points. This importance measure is general and widely applicable, as no assumption has been made in its derivation but that of independence of input variables. It should not be affected by model non-monotonicity.

4. Conclusions and future work

The analysis of the three worked examples has demonstrated that SA techniques, used in a black box fashion, can lead to misinterpretation of the results. This is particularly true when the output under consideration is a time dependent function of the input parameters.

The main obstacle to a fully automated SA procedure appears to be the existence of model non-monotonicities. These can be revealed by a scanning of the time axis for the dependent variable. The following estimators are particularly useful:

(1) Model coefficient of determination on ranks. Low values of this coefficient flag a possible inadequacy of a regression model based on ranks and of all the associated non-parametric estimators.

(2) Variable ranking plots as function of time. They give the general time evolution of the model governing parameters.
(3) Input–output rank scatterplot. They detect non-monotonicity regions.

It has been observed that in the regions of non-monotonicity variables with a linear relationship with the output can be given higher importance from the sensitivity estimators. Strictly speaking this is not a mistake of the estimators, since for the particular time point at hand a dilution factor might well happen to be the most influential parameter. This is the case in Level 1A (Figure 5). On the other hand Figure 7 suggests that the low value of the regression coefficient for *FLOWV1* in Level E might indeed become a high one if a non-linear model were to be regressed on the ranks.

It has also been observed that the linear sensitivity estimators yield responses qualitatively different from those of the non-parametric tests, emphasizing the role of the variables linearly correlated with the output.

The importance measure proposed by Hora and Iman appears promising, especially in view of the efficient computation scheme suggested by Ishigami and Homma. This technique does not appear to be affected by model non-monotonicity. A proper investigation of the performances of the new technique should include an analysis of the variance of the technique predictions over different simulations. An example of such an analysis is given in [8] for the classical estimators *PEAR* through *SMIR*.

References

[1] 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).
[2] J.C. Helton and J.D. Johnson. An uncertainty/sensitivity study for the station blackout sequence at a Mark I boiling water reactor. *Rel. Eng. Syst. Safety*, **26**: 293–328 (1989).
[3] R.L. Iman, J.C. Helton and J.E. Campbell, Risk methodology for geological disposal of radioactive waste: sensitivity analysis techniques. Sandia Natl. Laboratories report, SAND 78-0912 (1978).
[4] R.L. Iman and W.J. Conover, Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Commun. Statist. Theor. Meth.*, **A9(17)**, 1749–1842 (1980).
[5] 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).
[6] 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).
[7] 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).
[8] 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).
[9] W.J. Conover, Practical non-parametric statistics. 2nd Edition John Wiley & Sons Ed., New York (1980).

[10] 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.

[11] 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. 129–160, 1989.

[12] 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).

[13] 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).

[14] OECD-NEA, PSACOIN Level 1A Intercomparison. An international Code Intercomparison Exercise on a Hypothetical Safety Assessment Case Study for Radioactive Waste Disposal Systems. Prepared by A. Nies, D.A. Galson, J.M. Laurens and S. Webster. OECD-NEA publication, Paris (1990).

[15] 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.

[16] P.C. Robinson and D.P. Hodgkinson: exact solutions for radionuclide transport in the presence of parameter uncertainty. *Radioactive Waste Management and the Nuclear Fuel Cycle*, **8**, 1987.

[17] 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 EUR13922EN, Luxembourg 1991.

[18] 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 EUR13923EN, Luxembourg 1991.

[19] 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 EUR13924EN, Luxembourg 1991.

[20] 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 coefficient. Sandia Natl. Laboratories report NUREG/CR 4122, SAND85-0044 (1985).

[21] R.L. Iman, J.M. Davenport, E.L. Frost and M.J. Shortencarier. Stepwise regression with PRESS and rank regression. Program User's guide. SANDIA National Laboratory report SAND79-1472 (1980).

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

[23] 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).

[24] T. Ishigami and T. Homma: an importance quantification technique in uncertainty analysis. Japan Atomic Energy Research Institute report JAERI-M 89-111 (1989).