Reliability Engineering and System Safety **28** (1990) 229–253

Non-parametric Statistics in Sensitivity Analysis for Model Output: A Comparison of Selected Techniques

A. Saltelli

Commission of the European Communities, Joint Research Centre—Ispra Establishment, 21020 Ispra (Varese), Italy

&

J. Marivoet

Belgian Nuclear Research Establishment SCK/CEN, Boeretang 200, B-2400, Belgium

(Received 26 May 1989; accepted 3 August 1989)

ABSTRACT

The use of Statistics in risk assessment studies is an expanding field where the selection of the proper technique is often difficult to make. This is the case with the sensitivity analysis methods used in conjunction with Monte Carlo computer codes. The Monte Carlo approach is commonly used in risk assessment, where it can be used to estimate the uncertainty in the model's output due to the uncertainty in the model's input parameters. This treatment is referred to as Uncertainty Analysis, and is generally complemented with a Sensitivity Analysis, which is aimed at the identification of the most influential system parameters. Often different sensitivity analysis techniques are used in similar contexts, and it would be useful to identify (a) whether certain technique(s) perform better than others and (b) when two or more techniques can provide complementary information.

In this article a number of sensitivity analysis techniques are compared in the case of non-linear model responses. The test models originate from the context of the risk analysis for the disposal of radioactive waste, where sensitivity analysis plays a crucial role. The statistics taken into consideration include:

Pearson Correlation Coefficient
Partial Correlation Coefficient

Standardized Regression Coefficient
Smirnov Test Statistic
Mann–Whitney Test Statistic
Spearman Rank Correlation Coefficient
Partial Rank Correlation Coefficient
Standardized Rank Regression Coefficient
Cramer–Von Mises Test Statistic
Two sample t Test Statistic

All the techniques are applied to output from the same Monte Carlo simulations, where random sampling is used for the sample selection. Hypothesis testing is systematically applied to quantify the degree of confidence in the results given by the various sensitivity estimators. Although the problem of relative efficiency is not touched upon explicitly, the estimators are ranked according to their robustness and stability for the test case under consideration, and qualitative differences in the prediction of the various tests are pointed out.

1 INTRODUCTION

1.1 The problem

The analysis of the sensitivity of model response to the value of input parameters is a crucial step in the analysis of model performance, especially when the model itself is complex and involves many variable parameters. The difficulty of Sensitivity Analysis (SA) increases when:

(a) the model is non-linear;
(b) the model is non-monotonic;
(c) the model's output is a time dependent function of the input variables;
(d) the distribution functions of the input parameters range over many orders of magnitude;
(e) there are many 'ties' in the output vector (a possibility is that many output data are zero);
(f) the computer code where the model has been implemented is time consuming and expensive to run.

It is well known that non-parametric statistics based on the ranks of both input and output vectors are an appropriate tool for tackling sensitivity analysis problems.[1] A great number of different techniques of this type are described in the literature.[1–6]

Since different research groups are currently using one or another of these techniques in similar contexts the problem arises of investigating the relative performance of the various statistics.

In this work a number of statistics have been considered, which includes the Spearman 'rho' coefficient, the partial rank correlation coefficient, the standardized rank regression coefficient and some two-sample statistics such as the Mann–Whitney test. A few parametric statistics have also been included for the sake of comparison.

1.2 Previous work

The problem of intercomparing uncertainty and sensitivity analysis techniques has already been addressed by Iman and Helton,[4] who used three different complex test models to intercompare the performances of (a) response surface replacement for the computer model, (b) modified Monte Carlo as exemplified by Latin Hypercube sampling with and without regression analysis and (c) differential analysis.

Some of the conclusions of that article are summarised here:

(a) *Response surface replacement for the computer model.* Fractional factorial design has been used to generate the response surface. Such a technique is an optimal choice for the input selection if the output behaves in a linear fashion. Because this is not generally the case with complex models, the response surface might not be adequate in approximating these models.

(b) *Latin hypercube sampling.* It is the easiest to implement, especially when the number of variables is large. This type of design can easily handle complex multivariate input structures, when the input variables are not independent from each other (see also Ref. 7). The input space is well represented with this technique. As far as sensitivity analysis is concerned several methods for ranking the input variables can be adopted, including the powerful partial rank correlation coefficient. This is not always the case when using the above design (a).

(c) *Differential analysis.* This technique is intended to provide information with respect to a small perturbation about a point. 'Problems arise, however, in an uncertainty analysis or in sensitivity analysis when large uncertainties are present and attempts are made to extend the results from the small perturbation in the input variables, for which the differential analysis is intended, to a broader global interpretation'.[4] In other words, for complex models with large uncertainties the results might be too sensitive to the choice of the base-case point. The implementation of this technique can be difficult, depending on the nature of the model. The partial rank correlation coefficient cannot be used for sensitivity analysis.

The same authors have subsequently condensed their study in a different

article,[8] confirming their conclusions and recommending the use of LHS in conjunction with non-parametric rank correlation/regression techniques for sensitivity analysis.

1.3 Approach

Because design (b) above appeared as the most promising as far as sensitivity analysis is concerned it has been used in the present article, by expanding on the use of different SA estimators.

A test case has been selected which displays most of the difficulties mentioned in section 1.1. It pertains to the radioactive waste management field, and has been used already in a recent code intercomparison exercise promoted by the Nuclear Energy Agency of the Organization for Economic Cooperation and Development (OECD/NEA)[9] in order to check the statistical sampling techniques of different computer codes used for probabilistic safety assessments for nuclear waste repositories.

The input data sample has been generated using purely random sampling, as the Latin Hypercube was not suited to some of the SA techniques being compared (Smirnov test, for example; see Appendix). As usual, when running a model in a Monte Carlo fashion, the input sample consists of different sets (vectors) of input parameters. For each set the model is executed once. The ensemble of these executions (runs) is called a simulation (or case), and yields a distribution of values for the output variable(s) under consideration. Uncertainty and sensitivity analyses aim at characterising this distribution(s) with respect to the distributions of the input parameters.

The approach taken in the present study has been to realize different 'simulations', changing each time the seed used for the random number generation, and to compare the variances of the SA estimator prediction over the various simulations. Estimator–estimator score correlation coefficients have also been computed. A short description of the formulae used in the sensitivity analysis study is given in the Appendix.

2 METHODS

2.1 Sampling

Random sampling has been systematically used for all the simulations described. Because of the random pairing of the input values undesired correlations among the input variables may be introduced. Such correlations are particularly undesirable when the sample has also to be used for purposes of sensitivity analysis. Although a technique to eliminate

spurious correlations[7] is available in our sampling subroutine it was not used for the test model under consideration.

2.2 Sensitivity analysis

Various statistical techniques have been applied to the model output data in order to rank the input parameters as a function of their influence on the output distribution. It is understood that the ranking will not be unique. Different estimators (techniques) can focus on different kinds of input–output correlation, thus assigning different ranks to the parameters.

For the sake of conciseness the tests applied in this exercise are referred to by the abbreviated name used in the computer program. A list of these abbreviations is given below.

Pearson Correlation Coefficient	PEAR
Spearman Rank Correlation Coefficient	SPEA
Partial Correlation Coefficient	PCC
Partial Rank Correlation Coefficient	PRCC
Standardized Regression Coefficient	SRC
Standardized Rank Regression Coefficient	SRRC
Smirnov Test Statistic	SMIR
Cramer–Von Mises Test Statistics	CRAM
Mann–Whitney Test Statistics	TMWT
Two Sample t Test Statistic	TTST

A description of the above techniques is given in the Appendix, while their main characteristics are summarized in Table 1.

TABLE 1

Classification of the Statistics Used. P = parametric; NP = non-parametric; X_j is the Generic Input Variable ($j = 1, 2, \ldots, K$), which takes the Value x_{ij} in the ith Runs ($i = 1,2, \ldots, N$). Y_i is the Corresponding Value of the Output Variable

Objective	*Statistic type*		*Data disposition*
Determination of correlation within the sample	PEAR	P	1 bivariate sample (y_i, x_{ij})
	SPEAR	NP	$i = 1, 2, \ldots, N$ for each variable X_j
	PCC	P	1 $(K+1)$ variate sample
	PRCC	NP	$(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$
	SRC	P	$i = 1, 2, \ldots, N$
	SRRC	NP	for each variable X_j
Test for the location of the two samples	TTST	P	2 monovariate samples
	SMIR	NP	(x_{ij}) $i = 1, 2, \ldots, N_1$
	CRAM	NP	(x_{nj}) $n = 1, 2, \ldots, N_2$
	TMWT	NP	for each variable X_j

2.3 Computations

The results from the test model have been obtained with the LISA.SCK code.[10] This code has been developed at the Centre d'Etudes Nucleaires (SCK/CEN) of Mol (B) by adapting the LISA code, from the Joint Research Centre (JRC) of the CEC in ISpra (I).[11] The sensitivity analysis computations have been made with the SPOP code developed at the JRC.[12]

3 THE TEST MODEL

The 'Level 0' model was set up by the Probabilistic System Assessment Code (PSAC) group of the OECD/NEA in order to run one of the group's

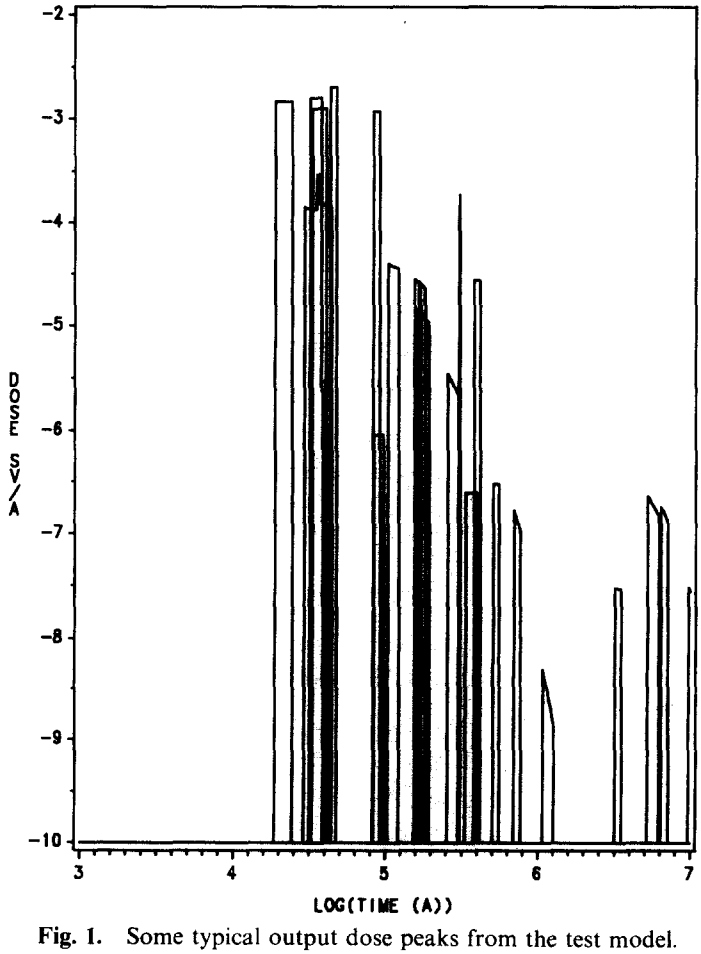

Fig. 1. Some typical output dose peaks from the test model.

benchmark exercises.[9] Twelve different organizations from the United Kingdom, Belgium, West Germany, Sweden, Finland, Canada, USA and Japan have participated in the exercise, which was mainly focused on uncertainty analysis, but also on sensitivity analysis.

The test model describes a hypothetical disposal system making use of simplified mathematical equations. The following compartments can be identified: the waste form, a buffer material, the geosphere and the biosphere. Seven radionuclides, ^{135}Cs, ^{129}I, ^{107}Pd, ^{79}Se, ^{151}Sm, ^{126}Sn and ^{93}Zr are considered in the exercise.

The rate of radionuclides' release from the waste form is assumed to be constant until depletion of the source. The transport through the buffer is described as a pure delay function, i.e. the output from the buffer sub-model

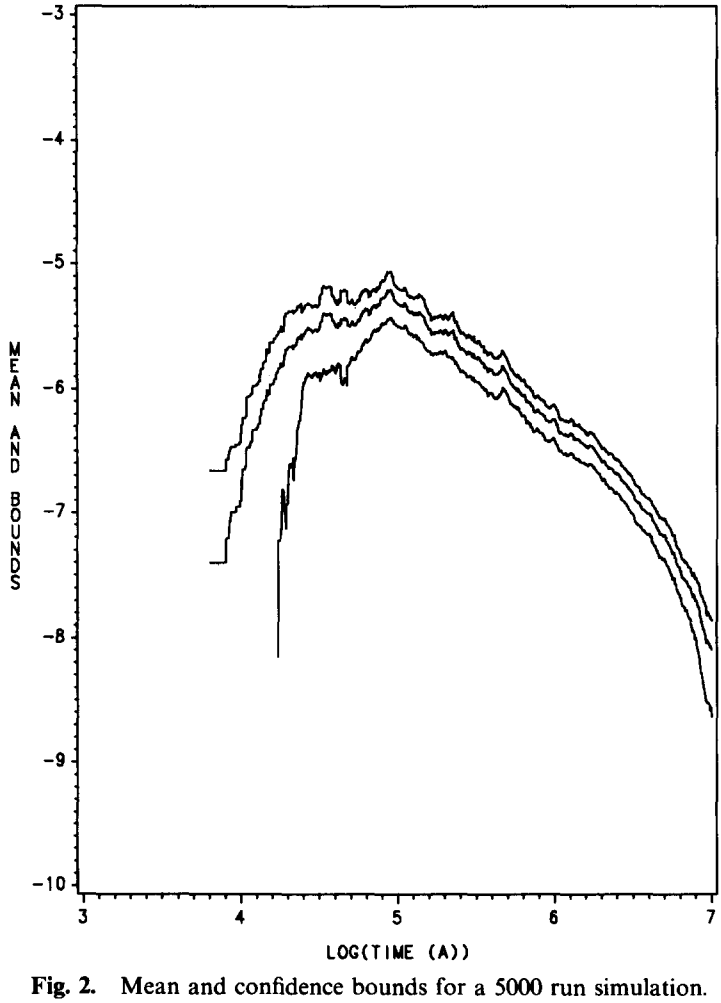

Fig. 2. Mean and confidence bounds for a 5000 run simulation.

equals its input delayed by a time t_B, computed as a function of the buffer characteristics and of the radionuclide dispersion–retention parameters. The geosphere model includes advection and dispersion. The gaussian transfer function corresponding to transport by advection and dispersion is simplified to a rectangular transfer function, the width of which simulates the effect of dispersion. The biosphere model considers water ingestion only, where the water is pumped from a well which constitutes the geo/biosphere interface.

Radioactive decay is considered in each compartment.

Because of the approximate formulae used for diffusion and dispersion, the output mainly consists of sharply peaked or rectangular pulses. The output total dose (all the nuclides) for five typical runs is shown in Fig. 1, where each peak generally represents the contribution from a different nuclide. It can be seen that for many time points there is no dose at all, i.e. $y_i(t) = 0$, so that the output vector $Y(t)$ contains mostly zeroes. This results in a large number of ties when the ranks are taken and constitutes a serious problem for most of the SA techniques employed.

TABLE 2
Input Parameters and Their Distribution for the Test Model

Notation	*Definition*	*Distribution*	*Value*
RLEACH	leach rate	log-uniform	(0·002 69, 12·9) kg/m^2/a
XBFILL	buffer thickness	uniform	(0·5, 5) m
XPATH	geosphere path length	uniform	(1 000, 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^2/a
BD(Cs)	sorpt. const. in the buffer	lognormal	mean = −0·46, std = 0·26, 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

The mean total dose resulting from a simulation of 5000 runs is given in Fig. 2, together with the 95% Tchebycheff confidence bounds. In spite of the large number of runs the curve has not yet converged to a smooth profile.

The model parameters are given in Table 2 together with the characteristics of their distributions. The large range of variability of the parameters is also a source of difficulty for the SA.

4 RESULTS AND DISCUSSION

4.1 Preliminary analysis

One of the findings of the code intercomparison exercise described in Ref. 9 was that the test case used there (and in the present work) was much more difficult for the sensitivity analysis part than it was for the uncertainty analysis part.

This difficulty is illustrated by the analysis of the model coefficient of determination R_y^2, which provides a measure of the effectiveness of the linear regression model based upon the input parameters. Values close to one (in absolute value) indicate a good performance of the regression model (see Appendix). Low R_y^2 values suggest that the output under consideration is poorly reproduced by the linear regression model, and indicate a poor performance of the SA techniques based upon regression.

In Fig. 3, three quantities have been plotted:

(1) the R_y^2 values based upon the *values* of the input parameters for the output variable 'dose rate at the time point' (R_y^2 based upon the SRCs; see Appendix);
(2) the R_y^2 values based upon the *ranks of the input values* (R_y^2 based upon the SRRCs);
(3) the percentage of non-zero output for each time point.

The curves in Fig. 3 have been computed from a simulation of 1000 runs. It can be seen that the percentage of non-zero runs never exceeds 27% and is as low as a few per cent 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 must be stressed that the results given in Fig. 3 are not due to the sample size, i.e. the model coefficient of determination does not increase when increasing the sample size.

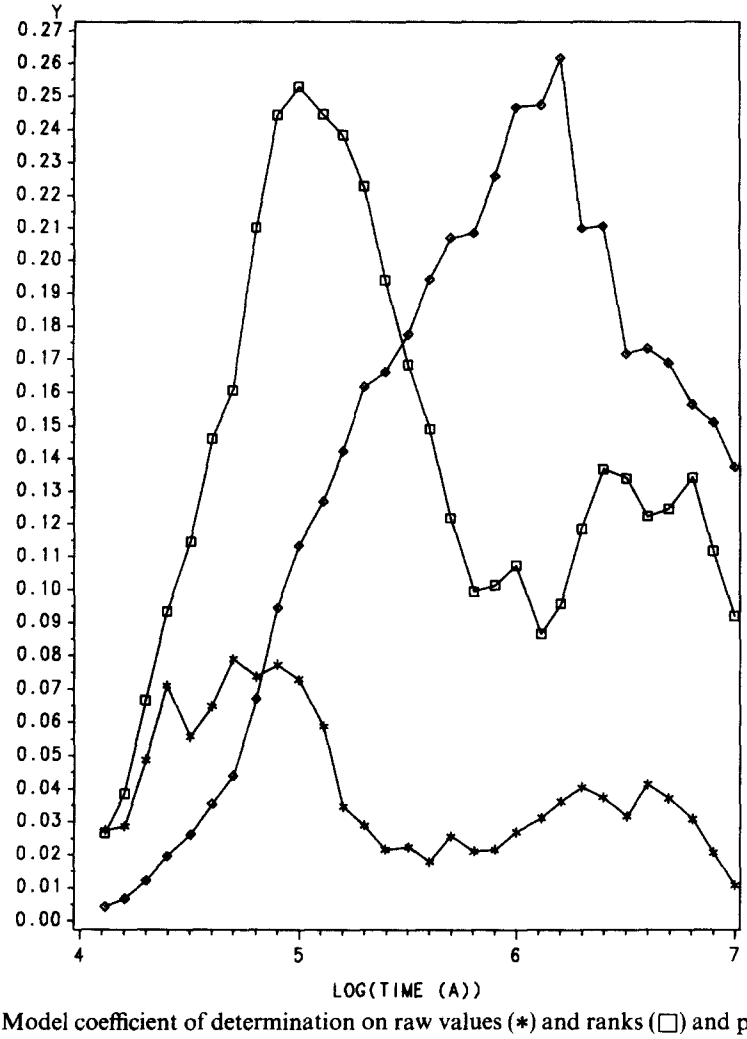

Fig. 3. Model coefficient of determination on raw values (*) and ranks (□) and percentage of non-zero outputs (◇) for output = 'dose at the time point' as a function of time.

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 24% of the data variance it may not be worthwhile to determine how much variance each of them can account for individually. Ranking the parameters on this basis would be questionable.

In Fig. 4 the percentage of non-zero runs and the R_y^2 based on the SRRCs have been plotted, taking as output the 'maximum total dose rate between $t = 0$ and the considered time point'. The percentage runs yielding non-zero output is much higher in this case, as can be expected, and the model

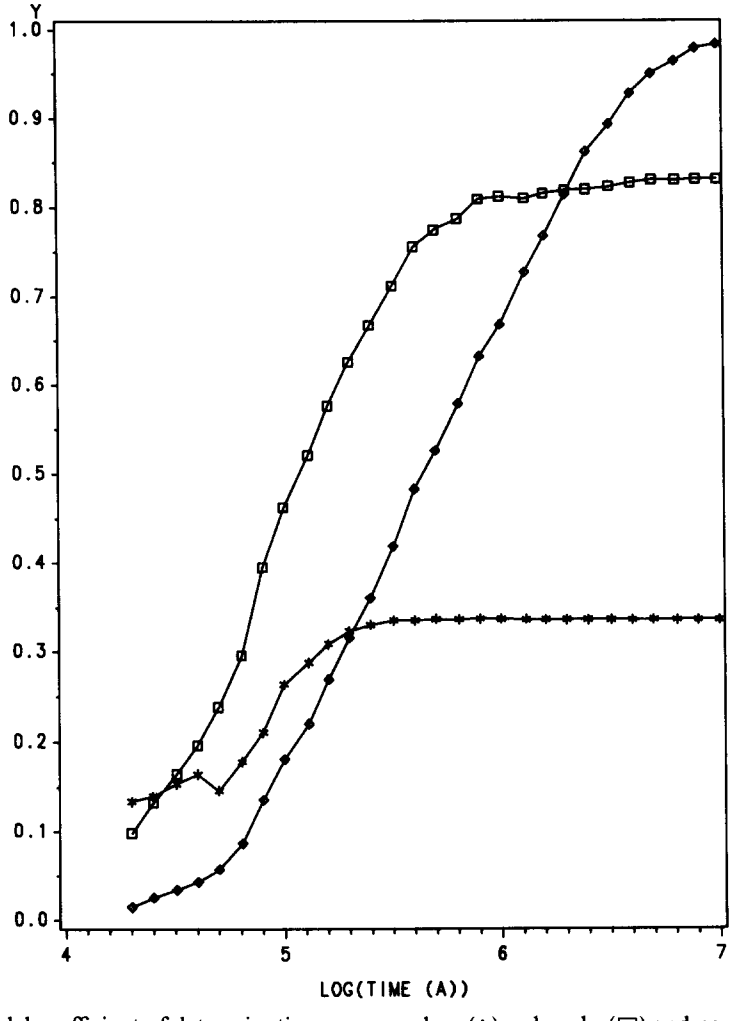

Fig. 4. Model coefficient of determination on raw values (*) and ranks (□) and percentage of non-xero outputs (◇) for output = 'maximum dose up to the time point' as a function of time.

coefficient of determination increases consistently. The strong dependence of R_y^2 upon the fraction of non-zero runs is evident.

It is a peculiar characteristic of this model that the model coefficient of determination increases with time.

In view of the above considerations it was decided to use the maximum dose up to time t, rather than the dose at that time point, as the output quantity. In this way the SRRCs can be used effectively to rank the input parameters and a comparison can be made with the predictions of the other SA estimators.

4.2 Relative performance of SA estimators as a function of sample size

The output from the test model consists of the maximum total doses, at selected points in time between zero and 10 million years.† For the sake of conciseness this quantity shall be simply referred to as 'dose' in the following discussion. In order to investigate the influence of the sample size on the relative performances of the selected SA estimators the following procedure was taken:

(a) A sample of 5000 runs is generated using the LISA.SCK code; for each run the sampled input parameter and the output dose time series are stored.
(b) The sample is then divided into subsamples ranging between 100 to 1000 runs (see table). Each set of subsamples characterised by the same number of runs is named a 'partition', where each partition contains the same 5000 runs of the original sample.

Partition	*Number of simulations in the partition*	*Size of each simulation (= number of runs)*
1st	50	100
2nd	20	250
3rd	10	500
4th	5	1 000

(c) Each of the 85 simulations (50 + 20 + 10 + 5) is taken as the input for a separate SA study, where the sensitivities of dose at different time points (10^5, $10^{5\cdot5}$, 10^6, $10^{6\cdot5}$, 10^7 yr) are analyzed using the SPOP code.
(d) The variance of the estimators' prediction over the various simulations is investigated.

For each simulation the following quantities are computed:

(a) quantiles of the test distribution corresponding to the number of runs employed at the selected significance level (see Appendix);
(b) model coefficient of determination R_y^2 (on raw values and on ranks) at the selected time points;

† The huge time span chosen for the intercomparison is characteristic of current 'radwaste' risk calculations in European countries[13–15] and in Canada.[16,17] A time scale of the same order of magnitude is adopted in the NEA coordinated international feasibility study for the sub-seabed disposal.[18]

(c) percentage of non-zero dose runs at the selected time points;
(d) values of the 10 statistics (e.g. PEAR, SPEA, etc.) for the 22 considered variables at the selected time points;
(e) variable ranking corresponding to each statistic;
(f) statistic–statistic score correlation coefficient.

An example of a quantile table is given in Table 3, which refers to one of the 1000 run simulations contained in the 4th partition. Values of R_y^2 and of the percentage of non-zero runs for the same simulation are given in Fig. 4 discussed previously. A statistics' table is presented in Table 4 for the $t = 10^6$ yr time point, and in Table 5 the corresponding ranks matrix is given, where each entry represents the rank given by each SA estimator to each variable (rank = 1 for the most important variable, rank = 22 for the least important one). Only the 6 most important variables are ranked in Table 5.

It can be seen that different techniques produce different rankings: variable V (water velocity in the geosphere) is identified as the top rank variable by all the tests, while for the second most influential variable there is disagreement even among the more 'reliable' non-parametric tests (*KD*(I), *XPATH* and *ABSR* are selected by SPEA, PRCC and SMIR respectively).

Estimators' prediction also varies from simulation to simulation, and this variation has been taken as the basis for investigating the relative 'stability' of the estimators at different sample sizes.

For each simulation five tables similar to Table 5 were generated, corresponding to the five time points under consideration. These tables are used to analyze the reproducibility of the SA estimators over the various simulations. Let the 1st partition be taken as an example. Here there are 50 different simulations of 100 runs each, and for each simulation and time point a ranking table like Table 5 is produced. Because of the random sampling, different rankings will be produced in different simulations.

Let R_{jk} (test, SIM_i) represent the rank given by the statistic 'test' (e.g.

TABLE 3

Quantiles W of the Test Distribution for a 1000 Run Simulation. Contingency Level ALPHA = 0·05. For the Two Sample Tests (e.g. SMIR) a 10%–90% Partition was used

W(ALPHA/2) for the normal distr. = −1·96		
SPEA	W(ALPHA/2) = −0·062	W(1 − ALPHA/2) = 0·062
SMIR	W(1 − ALPHA) = 0·143	
CRAM	W(1 − ALPHA) = 0·461	
TTST	W(ALPHA/2.) = −1·96 (t distribution)	
TMWT	W(ALPHA/2.) = −1·96 (t distribution)	

TABLE 4
Statistics Value (Non-significant Values Omitted)

Variable	*PEAR*	*SPEA*	*PCC*	*PRCC*	*SRC*	*SRRC*	*SMIR*	*CRAM*	*TTST*	*TMWT*
V	0·48	0·83	0·50	0·88	0·48	0·83	0·66	17·78	−17·27	−12·97
RLEACH	—	—	—	0·11	—	0·05	0·16	0·54	—	—
ABSR	−0·22	−0·15	−0·25	−0·32	−0·21	−0·15	0·34	5·11	7·48	7·14
XPATH	−0·11	−0·19	−0·16	−0·47	−0·13	−0·24	0·18	1·34	3·68	3·65
KD(Zr)	0·15	—	0·14	—	0·12	—	—	—	—	—
KD(Cs)	—	—	—	—	—	—	—	—	—	—
KD(Pd)	—	—	—	—	—	—	—	—	—	—
KD(Se)	—	—	—	—	—	—	—	—	—	—
KD(Sm)	—	—	—	—	—	—	—	—	—	—
KD(Sn)	—	—	—	—	—	—	—	—	—	—
KD(I)	—	−0·20	—	−0·41	—	−0·19	0·19	0·92	—	2·98
BD (Zr)	—	—	—	—	—	—	—	—	—	—
BD (Cs)	—	—	—	—	—	—	—	—	—	—
BD (Pd)	—	—	—	0·08	—	—	—	—	—	—
BD (Se)	—	—	—	—	—	—	—	—	—	—
BD (Sm)	—	0·11	—	—	—	—	—	—	—	—
BD (Sn)	—	—	—	0·06	—	—	—	—	—	—
BD(I)	—	—	—	—	—	—	—	—	—	—
RMW	—	—	—	—	0·06	—	—	—	—	—
XBFILL	—	—	—	—	—	—	—	—	—	—
ADISPG	−0·15	−0·13	−0·19	−0·27	−0·16	−0·12	0·33	4·49	5·37	6·61
DIFFC	—	—	−0·08	−0·06	−0·06	—	—	—	—	—

SPEA) to the variable j ($j = 1, 2, \ldots, 22$) at the time point k ($k = 1, 2, \ldots, 5$) in the simulation SIM_i ($i = 1, 2, \ldots, 50$). The variance of R_{jk} (test, SIM_i) over the 50 simulations could then be used to measure the stability of 'test', once the variances for all the reference times and variables are summed together.

However in this way the top ranks would yield as much weight as the low ranks, i.e. an agreement, between two simulations for the top ranking variable ($R = 1$) would be considered equal to an agreement on a low rank one (say $R = 20$). Instead the degree of agreement between two techniques should be evaluated mostly on the top ranks, where the tests are more significative, giving decreasing weights to decreasing ranks. This can be achieved by replacing the ranks with their Savage scores.[4]

The Savage score S_i of a certain rank value R_i can be computed as

$$S_i = \sum_{m=R_i}^{K} (1/m)$$

TABLE 5
Variable Ranking (only the 6 Most Significant Ranks are Given)

Variable	*PEAR*	*SPEA*	*PCC*	*PRCC*	*SRC*	*SRRC*	*SMIR*	*CRAM*	*TTST*	*TMWT*
V	1	1	1	1	1	1	1	1	1	1
RLEACH	—	—	—	6	—	6	6	6	—	—
ABSR	2	4	2	4	2	4	2	2	2	2
XPATH	5	3	4	2	4	2	5	4	4	4
KD (Zr)	4	—	5	—	5	—	—	—	6	—
KD (Cs)	—	—	—	—	—	—	—	—	—	—
KD (Pd)	—	—	—	—	—	—	—	—	—	—
KD (Se)	—	—	—	—	—	—	—	—	—	—
KD (Sm)	—	—	—	—	—	—	—	—	—	—
KD (Sn)	—	—	—	—	—	—	—	—	—	—
KD (I)	—	2	—	3	—	3	4	5	—	5
BD (Zr)	—	—	—	—	—	—	—	—	—	—
BD (Cs)	—	—	—	—	—	—	—	—	—	—
BD (Pd)	—	—	—	—	—	—	—	—	—	—
BD (Se)	—	—	—	—	—	—	—	—	5	6
BD (Sm)	—	6	—	—	—	—	—	—	—	—
BD (Sn)	—	—	—	—	—	—	—	—	—	—
BD (I)	—	—	—	—	—	—	—	—	—	—
RMW	6	—	—	—	—	—	—	—	—	—
XBFILL	—	—	—	—	—	—	—	—	—	—
ADISPG	3	5	3	5	3	5	3	3	3	3
DIFFC	—	—	6	—	6	—	—	—	—	—

where K = number of variables (this formula must be slightly modified in the case of ties); so the variable with the highest rank ($R_i = 1$) is given a score

$$S_i = 1 + 1/2 + 1/3 \cdots + 1/22 = 3{\cdot}69$$

whereas for $R_i = 22$

$$S_i = 1/22 = 0{\cdot}0455$$

In this way the scores for the 50 simulations were compared, computing for each variable and for each reference time, the variance of the estimators over the 50 simulations. Taking SMIR as an example,

$$\mathrm{var}_{jk}(\mathrm{SMIR}) = (1/49) \sum_{i=1}^{50} (S_{jk}(\mathrm{SMIR}, \mathrm{SIM}_i) - \bar{S}_{jk}(\mathrm{SMIR}))^2$$

where $S_{jk}(\mathrm{SMIR}, \mathrm{SIM}_i)$ = score attributed by the SMIR test to the variable X_j at the reference time k in the simulation SIM_i.

$\bar{S}_{jk}$(SMIR) = average of the above quantity over the 50 simulations. A new statistic D(SMIR) can then be defined as

$$D(SMIR) = \sum_{k=1}^{5} \sum_{j=1}^{22} \mathrm{var}_{jk}(\mathrm{SMIR})$$

where the estimator variances are summed over all the variables and reference times. The above treatment was repeated on all the partitions for all the estimators under consideration.

Values of the D statistics for the various partitions are given in Table 6, where the sample size N is also indicated. This table gives a clear picture of the effect of the sample size on the reproducibility of the estimators' prediction. As a general trend D values decrease with increasing sample size, yet the decrease is more pronounced at small sample sizes. In Fig. 5, values for PCC (= SRC), PRCC (= SRRC) and SMIR (very similar to CRAM and TMWT) have been plotted. This figure clearly shows that PRCC, besides having the smallest D for the entire range of explored sample size N, is also the least affected by N (smallest $\partial D/\partial N$).

The non-parametric two-sample tests are dramatically affected by the sample size, and for $N = 250$ and $N = 100$ their performances become worse than those of the parametric tests, e.g. the scatter in the SMIR predictions from simulation to simulation become more relevant than that of PCC of PEAR.

This is a very interesting result when the sensitivity of PEAR to the distribution outliers is considered. Also interesting is the convergence of D values for PRCC, SRRC, SMIR, CRAM and TMWT at $N = 500$.

TABLE 6
D Statistic for the Four Partitions

1st Partition 50 simulations of 100 runs		*2nd Partition 20 simulations of 250 runs*		*3rd Partition 10 simulations of 500 runs*		*4th Partition 5 simulations of 1000 runs*	
PRCC	0·22	PRCC	0·18	CRAM	0·15	PRCC	0·15
SRRC	0·22	SRRC	0·18	PRCC	0·15	SRRC	0·15
SPEA	0·31	SPEA	0·23	SRRC	0·15	SMIR	0·15
PCC	0·32	CRAM	0·23	SMIR	0·15	CRAM	0·15
SRC	0·33	PCC	0·24	TMWT	0·16	TMWT	0·16
PEAR	0·34	SRC	0·24	SPEA	0·20	SPEA	0·19
TMWT	0·36	TMWT	0·24	SRC	0·21	TTST	0·19
TTST	0·36	SMIR	0·25	PCC	0·21	PCC	0·21
CRAM	0·37	PEAR	0·27	TTST	0·21	SRC	0·21
SMIR	0·38	TTST	0·27	PEAR	0·23	PEAR	0·22

The poor performance of SMIR was further investigated. An analysis of the rank tables (as Table 5) for the fifty 100 run simulations has shown that in effect the PEAR predictions are more reproducible than those of SMIR at these sample sizes. The tables of the $t = 1\,000\,000$ yr time point, for instance, show that the variable V is the most influential variable for both the estimators; for the second most important variable, then, PEAR selects very often *ABSR* (29 times out of 50), whereas the selections of SMIR are more scattered (16 times *ABSR*, 8 times *ADISPG*, 3 times *XPATH*, etc.). This can be due to the fact that PEAR always favours the variables having a strong linear influence on the output (such as *ABSR*, see next section), whereas SMIR tends to select both this variable and those having a non-linear influence (such as *XPATH*).

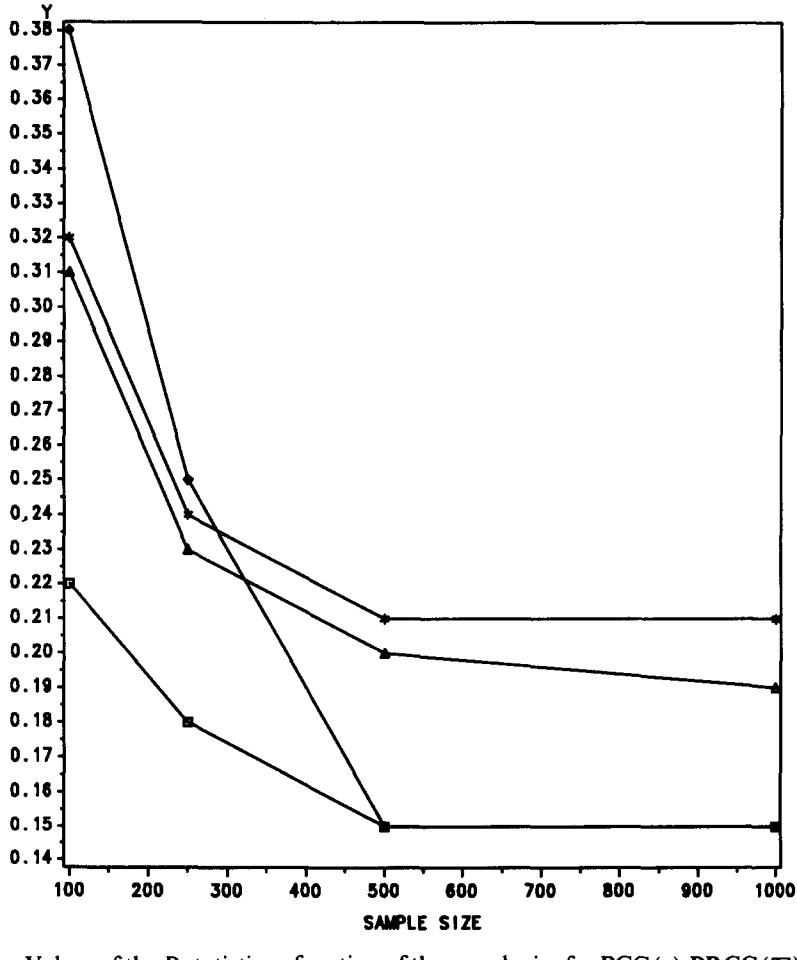

Fig. 5. Values of the D statistic as function of the sample size for PCC (*), PRCC (□) SMIR (◇) and SPEA (△).

4.3 The $t = 10^6$ yr time point. A detailed analysis

Tables 4 and 5, relative to the $t = 10^6$ yr time point, can be used to further investigate the differences between the various tests.

If the ranks in Table 5 are converted in scores—as described in the previous section—the statistic–statistic correlation (Table 7) can be computed.

TABLE 7
Score Correlation Coefficients

Statistic	*PEAR*	*SPEA*	*PCC*	*PRCC*	*SRC*	*SRRC*	*SMIR*	*CRAM*	*TTST*	*TMWT*
PEAR	1·00	—	—	—	—	—	—	—	—	—
SPEA	0·73	1·00	—	—	—	—	—	—	—	—
PCC	0·90	0·71	1·00	—	—	—	—	—	—	—
PRCC	0·69	0·84	0·78	1·00	—	—	—	—	—	—
SRC	0·90	0·71	1·00	0·78	1·00	—	—	—	—	—
SRRC	0·69	0·84	0·78	1·00	0·78	1·00	—	—	—	—
SMIR	0·86	0·80	0·89	0·80	0·89	0·80	1·00	—	—	—
CRAM	0·87	0·79	0·89	0·82	0·89	0·82	0·99	1·00	—	—
TTST	0·92	0·79	0·83	0·70	0·83	0·70	0·86	0·89	1·00	—
TMWT	0·87	0·80	0·89	0·81	0·89	0·81	0·96	0·98	0·92	1·00

It shows that, as expected, non-parametric techniques correlated more with each other than with the parametric ones. PRCC and SRRC (as well as PCC and SRC) have correlation coefficients equal to one. In fact, given the functional relationship existing between standardized regression and partial correlation coefficients, these SA indicators yield identical ranking in all the simulations. Nevertheless it can be shown that the rankings from PRCC and SRRC diverge when significant correlations are involved among the input variables, which is not the case for the present exercise. There is also a high correlation within the group of the three non-parametric two-sample tests SMIR, CRAM, TMWT (correlations between 0·96 and 0·99); these three estimators correlate better with the parametric tests PEAR, PCC, SRC than with the equivalent non-parametric ones (though this is true only for the $t = 10^6$ yr time point). TTST is nonspecific, and correlates equally well with the non-parametric tests and with the other two-sample tests. The group of SPEA, PRCC and SRRC correlates better with the group of SMIR, CRAM, TMWT than with their parametric equivalents PEAR, PCC, SRC.

This does not yet solve the problem of selecting the proper variable ranking among those provided in Table 5. An essential piece of information is provided by the model coefficient R_y^2. The R_y^2 value for the data in Table 4 is:

$$R_y^2 = 0{\cdot}385 \text{ for SRCs (based on the raw values)}$$

$$R_y^2 = 0{\cdot}810 \text{ for the SRRCs (based on the ranks)}$$

Thus, as expected, the SRRCs are to be more trusted than the SRCs, and the PRCC more than the PCC. As could have been easily anticipated, non-parametric techniques based on rank are more adequate for this type of model.

Yet the information provided by the parametric tests (PEAR, etc.) should not be disregarded.

The variable *ABSR*, water extraction rate from the well, provides a good example of the differences between the parametric and non-parametric responses. This variable has no effect on the travel time of the nuclides (governed by V, *XPATH*,...) but it has a strong effect on the height of the dose peak. As *ABSR* varies over one order of magnitude it has a very strong influence on a linear scale. On a logarithmic scale instead, *ABSR* can only vary dose by one unit, against the many units' variation dominated by V and *XPATH*.

For this reason in all the simulations and at all the reference times *ABSR* is given more importance from PEAR, PCC, SRC than from SPEA, PRCC, SRRC. The two-sample tests (SMIR, CRAM, TMWT) also classify *ABSR* as the second most important variable, and this is not surprising as the high dose runs collected in the 10% sub-sample† are likely to be associated with low *ABSR* values (dose and *ABSR* are inversely proportional).

The same can be said for *ADISPG* (rank 3 for SMIR, CRAM, TMWT and only 5 for SPEA, PRCC and SRRC); this variable represents the geosphere dispersivity, and in the present model its effect is to smooth out the dose peaks. Most of the runs in the 10% high dose sub-sample are likely to be associated with low *ADISPG* values, although, by itself, *ADISPG* does not have the capability of producing non-zero dose outputs and is in fact the 5th parameter in the regression model built by the SRRCs.

As a general trend the two-sample tests give more weight to the parameters which influence the high dose outputs, where the SRRC's search for the best model to fit all the outputs. These examples, together with the analysis of the score correlation Table 7, lead to the conclusion that the two sample tests, when applied to the 10%–90% sub-samples, may exhibit some of the peculiar features of the parametric SA estimates (for example PEAR), overestimating a few high outputs (outliers) of the distribution.

† When using the two-sample tests (SMIR through TMWT) the output sample from a given simulation is partitioned into two sub-samples, the first one containing, for example, the 10% runs yielding the highest output, the second one containing the remaining runs (see the Appendix).

5 CONCLUSIONS

A number of interesting results have been obtained from the inter-comparisons that have been made among the selected sensitivity analysis estimators. These are summarized in the following points:

(1) The relative stability of sensitivity analysis indicators depends upon the sample size (Fig. 5). When increasing the sample size the variability (from simulation to simulation) of the non-parametric estimators tend to converge to its lowest asymptote.
(2) The disagreement between estimators, including the non-parametric ones (generally considered as the most reliable) are non-negligible. In particular there are differences among the predictions of the family SPEA, PRCC, SRRC and the two sample tests SMIR, CRAM, TMWT. These latter, when employed on the 10%–90% sub-samples, exhibit some of the negative characteristics of the parametric tests.
(3) The disagreement increases when decreasing the sample size.
(4) The estimators PRCC and SRRC appear to be, in general, the most robust and reliable. In particular they seem much more effective than SMIR at low sample size.

An additional remark can be made. The discussion of the $t = 10^6$ yr time point has shown that, given a sample of size 1000, about 5 variables were successfully ranked (in average) by the SA estimators. If the sample were to be increased indefinitely more variables would be ranked by each estimator. All the test statistics used in this work for hypothesis testing are in fact consistent in the statistical sense of the term, i.e. if the sample size tends to infinity the power of these tests tend to unity, where the power is defined as the probability of rejecting a false hypothesis. Taking again SMIR as an example, an increase in the sample size will result in a lower value of the quantile for the Smirnov test distribution, and SMIR will become significant for more influential variables. In an analogous way, by increasing the sample size, R_y^2 will tend to its asymptote, which corresponds to the degree of linearity between the ranks of Y and X_js. Nevertheless, for sufficiently complex models, the ranking of all the variables might be impossible to achieve, especially when the model involves non-monotonic input–output relationships.

The advantage of using more than just one SA technique lies in the fact that performing a statistical test is generally much cheaper than running a simulation. Furthermore, it has also been seen in the discussion of the role of the variables V, *XPATH*, *ABSR* and *ADISPG*, that sensitivity analysis techniques must be complemented with knowledge of the system. Although the use of Hypothesis Testing can give a certain degree of

confidence in the results of the sensitivity analyses, errors are always possible. It would be very difficult to interpret the results of the analysis wihtout cross-checking between statistics and understanding the model.

ACKNOWLEDGEMENT

The assistance of Drs David Stanners (JRC) and Toshimitsu Homma (JAERI, (J)) in revising this article is gratefully acknowledged.

REFERENCES

1. Iman, R. L., Helton, J. C. & Campbell, J. E., Risk methodology for geological disposal of radioactive waste: sensitivity analysis techniques. Sandia Natl. Laboratories report, SAND 78-0912 (1978).
2. Iman, R. L. & Conover, W. J., Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Commun. Statist. Theor. Meth.*, **A9**(17) (1980) 1749–842.
3. Iman, R. L., Helton, J. C. & Campbell, J. E., An approach to sensitivity analysis of computer models. Parts I and II. *J. Quality Technology*, **13**(3,4) (1981) 174–83 and 232–40.
4. Iman, R. L. & Helton, J. C., A comparison of uncertainty and sensitivity analysis techniques for computer models. Sandia Natl. Laboratories report NUREG/CR-3904, SAND 84-1461 (1985).
5. Conover, W. J., *Practical Non-Parametric Statistics*. 2nd Edition. John Wiley & Sons, New York, 1980.
6. Iman, R. L., Shortencarier, M. J. & Johnson, J. D., 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, SAND 85-0044 (1985).
7. Iman, R. L. & Conover, W. J., A distribution free approach to inducing rank correlation among input variables. *Comm. Statist.* **B11**(3) (1982) 311–34.
8. Iman, R. L. & Helton, J. C., An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis*, **8**(1) (1988) 71–90.
9. Saltelli, A., Sartori, E., Goodwin, B. W. & Carlyle, S. G., PSACOIN Level 0 Intercomparison. An international Code Intercomparison Exercise on a Hypothetical Safety Assessment Case Study for Radioactive Waste Disposal Systems. OECD-NEA publication, Paris (1987).
10. Marivoet, J., The LISA.SCK code. In *PAGIS, Disposal in Clay Formation*, ed. J. Marivoet & A. Bonne. CEC, report EUR 11776 EN, Luxembourg (1988) pp. 337–40.
11. Saltelli, A., Bertozzi, G. & Stanners, D., LISA, A code for safety assessment in nuclear waste disposal. Program description and user guide. Joint Research Centre of Ispra report EUR 9306 EN, Luxembourg (1984).

12. Saltelli, A., PREP and SPOP utilities. Two FORTRAN programs for sample preparation, uncertainty analysis and sensitivity analysis in Montecarlo simulation. Programs description and user's guide. Joint Research Centre of Ispra report EUR 11034 EN, Luxembourg (1987).
13. SKBF/KBS, Final storage of spent nuclear fuel—KBS3, SKBF, Swedish Nuclear Fuel Supply Co./Division KBS report ISSN 0349-6015 (1983).
14. Hill, M. D. & Lawson, G., An assessment of the radiological consequences of disposal of high-level waste in coastal geologic formations. Nuclear Radiation Protection Board, report NRPB-R108, Harwell (UK) 1980.
15. NEA, Long-Term radiation protection objectives for radioactive waste disposal. Nuclear Energy Agency experts report, OECD, Paris (F), 1984.
16. Wuschke, D. M., Mehta, K. K., Dormuth, J. W., Andres, T., Sherman, G. R., Rosinger, E. L. J., Goodwin, B. W., Reid, J. A. K. & Lyon, R. B., Environmental and safety studies for nuclear waste management, volume 3: Post-closure assessment. Atomic Energy of Canada Ltd report AECL TR-1127-3 (1981).
17. Wuschke, D. M., Gillespie, P. A., Mehta, K. K., Heinrich, W. F., Leneveu, D. M., Guvanesen, V. M., Sherman, G. R., Donahue, D. C., Goodwin, B. W., Andres, T. H. & Lyon, R. B., Second interim assessment of the Canadian concept for nuclear fuel waste diposal—Volume 4: Post closure assessment. Atomic Energy of Canada Ltd report AECL-8373-4 (1985).
18. De Marsily, G., Behrendt, V., Ensminger, D. A., Flebus, C., Hutchinson, B. L., Kane, P., Karps, A., Klett, R. D., Mobbs, S., Poulin, M., Stanners, D. A. & Wuschke, D., Radiological assessment of the consequences of the disposal of high level radioactive waste in sub-seabed sediments. Presented at the 1987 winter meeting of the America Nuclear Society, Los Angeles, CA, 15–19 November 1987.

APPENDIX: SENSITIVITY ANALYSIS TECHNIQUES

The Pearson product moment correlation coefficient (PEAR) is the usual linear correlation coefficient computed on the x_{ij}, y_is $(i = 1, 2, \ldots, N)$. For non-linear models the *Spearman 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 instead of the raw values:[5]

$$\mathrm{SPEA}(Y, X_j) = \mathrm{PEAR}(R(Y), R(X_j))$$

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.

The numerical value of SPEA, commonly known as the Spearman 'rho', can also be used for hypothesis testing, to quantify the confidence in the correlation itself.

Partial Correlation Coefficients (PCC) and *Standardized Regression Coefficients* (SRC) are two very useful 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).

A description of how these coefficients are computed is given in Ref. 6. The $\mathrm{SRC}(Y, X_j)$ are the coefficients of the regression model for Y; they may provide an approximation to Y in the form:

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

where X_j^* are the normalized variables:

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

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

When using the SRCs 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 SRCs can reproduce the actual output vector **Y**. In particular:

$$R_y^2 = \sum_{i=1}^{N} (y_i^{\mathrm{m}} - \bar{y})^2 \Big/ \sum_{i=1}^{N} (y_i - \bar{y})^2$$

where $\bar{y}$ is the mean of the output values y_i and the y_i^{m} are the model prediction based on the SRCs, 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 $\mathrm{SRC}(Y, X_j)$ can themselves provide a very effective measure of the relative importance of the input variables. Of course the validity of the SRCs as a measure of sensitivity is conditional to the degree to which the regression models fits the data, i.e. to R_y^2.

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 PCCs 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 PRCCs 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 $\mathrm{SMIR}(Y, X_j)$ and *Cramer–Von Mises* $\mathrm{CRAM}(Y, X_j)$ belong to the same class of non-parametric statistics.

In particular they are 'two-sample' tests originally designed to check the hypothesis that two different samples belong to the same population. The application of such 'two-sample' tests 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_is 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 these statistics to be applicable, a number of basic 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 and CRAM 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.

The Cramer–Von Mises and Smirnov statistics resemble each other very closely; however, the test function for the former is related to the total area enclosed by the two cumulative distributions, and involves the summation of the squared distances between the two curves computed at all x_{ji} points, with $i = 1, 2, \ldots, N$. Because in this statistic the total area of the two distributions is scanned, it may be more appropriate for sensitivity analysis when $Y(X_j)$ is a non-monotonic function.

A description of both the SMIR and the CRAM tests is given in Ref. 5.

As with the two preceding statistics, the *Mann–Whitney test* (TMWT) is also applied to two samples of the same parameter and the hypothesis to be tested is whether or not the two samples come from the same population.

Actually TMWT is a test specially designed for detecting differences in the population location, so that the hypothesis $F_1(X_j) = F_2(X_j)$ can be replaced by an hypothesis stating the equivalence between the two population means.

Under certain circumstances (equal variance of the two samples) the hypothesis can be written as:

$$E(Z_1) = E(Z_2)$$

where E stands for the expectation value and Z_1, Z_2 refer to the X_j values selected in the two sub-samples.

For this statistic the ranks of the parameter values are used and the means of the two sub-samples are compared. The same basic assumptions made for the SMIR test hold for TMWT. The Mann–Whitney statistic and its test distribution are described in Ref. 5.

The *t-test* (TTST) is a widely used parametric statistic on the sample mean. The two sample version of the t-test used here is the parametric equivalent of the Mann–Whitney test; for practical calculations in fact the formula for the t-test can be used for computing the Mann–Whitney test, once the parameter values have been replaced by their ranks.[5]

Because of its parametric nature, linked to the assumption of normality of the samples, the t-test is likely to compare unfavourably with its non-parametric equivalent TMWT for the generally non-normal data under consideration. The test, however, has been included for comparison purposes, in order to have a parametric statistic also in the class of the two sample tests (SMIR, CRAM, TMWT). A description of the two sample t-test (TTST) can be found in Ref. 5.

For all the above statistics *Hypothesis testing* is used to quantify the degree of confidence in the identification of an influential variable. This is exemplified here for the statistic SPEA.

The numerical value of SPEA can be used for hypothesis testing by making first the base hypothesis:

'no correlation exists between Y and X_j'

SPEA(Y, X_j) is then computed from a simulation of a given number of runs N, and its value is compared with the quantiles of the Spearman test distribution. The comparison is made at a certain pre-established level of significance (α), and the hypothesis of no correlation is rejected if SPEA is either lower than $W(\alpha/2)$ or higher than $W(1-\alpha/2)$, where the Ws are the quantiles of the test distribution.

The level of significance α is the probability of erroneously rejecting the hypothesis, i.e. in this context, the probability that the test indicates a correlation when Y and X_j are actually uncorrelated. To apply the test at a 0·05 significance level $W(0{\cdot}025)$ and $W(0{\cdot}975)$ must be computed or read on tables.[5]

Taking, for instance, $N = 500$, the Spearman quantiles are:

$$W(0{\cdot}025) = -0{\cdot}088$$
$$W(0{\cdot}975) = 0{\cdot}088$$

The hypothesis of no correlation is rejected if SPEA, as computed from a 500 run simulation, falls outside the range $(-0{\cdot}088, 0{\cdot}088)$, and the probability of an erroneous rejection, when Y and X_j are actually uncorrelated, is 0·05.