On the Relative Importance of Input Factors in Mathematical Models

Andrea Saltelli & Stefano Tarantola

Andrea Saltelli and Stefano Tarantola are researchers at the Joint Research Centre of the European Commission, Institute for the Protection and Security of the Citizen, Ispra, Italy.

Published online: 31 Dec 2011.

To cite this article: Andrea Saltelli & Stefano Tarantola (2002) On the Relative Importance of Input Factors in Mathematical Models, Journal of the American Statistical Association, 97:459, 702-709, DOI: 10.1198/016214502388618447

To link to this article: http://dx.doi.org/10.1198/016214502388618447

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions
On the Relative Importance of Input Factors in Mathematical Models: Safety Assessment for Nuclear Waste Disposal

Andrea Saltelli and Stefano Tarantola

This article deals with global quantitative sensitivity analysis of the Level E model, a computer code used in safety assessment for nuclear waste disposal. The Level E code has been the subject of two international benchmarks of risk assessment codes and Monte Carlo methods and is well known in the literature. We discuss the Level E model with reference to two different settings. In the first setting, the objective is to find the input factor that drives most of the output variance. In the second setting, we strive to achieve a preestablished reduction in the variance of the model output by fixing the smallest number of factors. The emphasis of this work is on how to define the concept of importance in an unambiguous way and how to assess it in the simultaneous occurrence of correlated input factors and non-additive models.

KEY WORDS: Analysis of variance; Correlated input; Nonadditive model; Sensitivity analysis.

1. INTRODUCTION

The work focuses on data that are the output from a computer code. Statisticians have concerned themselves with these kinds of datasets since the early 1990s (Sacks, Schiller, and Welch 1989; Sacks, Welch, Mitchell, and Wynn 1989b; Welch et al. 1992) addressing the issue from several viewpoints: (a) how to approximate the output from a complex computer code; (b) how to optimally design exploratory points to do (a); (c) how to characterize the empirical distribution of the output given probability distributions of the input; and (d) how to assess the importance of input factors in relation to (c). In this work, the emphasis is substantially on (d).

In recent years, global quantitative sensitivity analysis techniques have received considerable attention in the literature (RESS 1997; JSCS 1997; CPC 1999; JMCDA 1999). Many techniques have been developed that can be applied even to nonlinear, nonmonotonic models. In particular, in an earlier work (Saltelli, Tarantola, and Chan 1999) we proposed a model-free method for efficient global sensitivity analysis for sets of noncorrelated input factors. A review of applications of the same methods was given by Saltelli, Tarantola, and Camponogaro (2000). Saltelli, Chan, and Scott (2000) presented a broad spectrum of techniques.

In the present work we extend the methodology for use in the simultaneous presence of correlated inputs and nonadditive models, and apply it to a well-known test case, the Level E model. The Level E test case was first used as a benchmark to compare the performance of computer codes used in safety assessment for nuclear waste disposal (OECD 1989), and later was used as a test model for sensitivity analysis techniques (OECD 1993). In this latter study, called Level S, the Level E model was run in a Monte Carlo fashion, with the model fed with samples from the distribution of the inputs. The Level S aimed to rank the uncertain input factors in order of importance according to two different criteria:

a. Reduction in the spread of the output at given times for a 5% reduction at each end of the input factor’s range, that is, taking a new distribution for that factor running between the former 5th and 95th percentiles

b. Change in the mean of the output at given times for a 5% increase of the central value of the factor’s distribution, that is, shifting the whole distribution up, so that the bottom end is at the former 5th percentile.

The results showed that the sensitivity analysis methods applied in the Level S exercise were of limited value in ranking the importance of the input factors in the Level E test model. Depending on the formulation of the questions in Level S, the results were in fact dependent on the criterium selected (i.e., shrinking or shifting), and the effect of nonmonotonicity in the input-output relationship was not captured.

The main difficulty in sensitivity analysis at the time of the Level S exercise was on how to interpret its results. Although both criteria (a) and (b) are legitimate in principle, and answerable by straightforward Monte Carlo methods (OECD 1993), one does not know what question is being answered by a ranking of the input factors based on (a) or (b). Should we prioritize research on a given factor if it scores first in importance under (a)? Or under (b)? What if the answer depends on the percentage values previously stipulated? Should the same percentage apply to all factors, or should it include consideration of the likely reduction in uncertainty subsequently achievable in the laboratory, or in the field, for each individual factor? Was neglecting the covariance structure of the problem legitimate, or did it induce too large an error? A number of articles that followed the Level S report attempted to answer these questions (see Sec. 2.2).

In this work, we attempt to define “importance” in a rigorous fashion. The Level S benchmark revealed quite a latitude of different understandings of the concept of “importance” when applied to sensitivity analysis. We revisit the nonlinear, nonadditive Level E case study, complementing it with the input correlation that was omitted in the original exercise. As we show, a sensitivity analysis in the simultaneous occurrence...
of correlation (a property of the input sample) and interaction (a property of nonadditive models) is nontrivial.

2. THE LEVEL E MODEL

2.1 Description of the Model

Level E was used both as a benchmark of Monte Carlo computation (Robinson and Hodgkinson 1987; OECD 1989) and as a benchmark for sensitivity analysis methods (Level S), (OECD 1993). The model predicts the radiologic dose to humans over geologic time scales due to the underground migration of radionuclides from a nuclear waste disposal site. The scenario considered in the model tracks the one-dimensional migration of four radionuclides ($^{129}$I and the chain $^{237}$Np $\rightarrow$ $^{233}$U $\rightarrow$ $^{229}$Th) through two geospheric layers with different hydrogeologic properties. The processes considered in the model are radioactive decay, dispersion, advection, and chemical reaction between the migrating nuclide and the porous medium. The repository is represented as a point source. Sometime after the steel canister containing the waste has lost its integrity (with the time of containment failure indicated by $T$), the release of radionuclides to the geosphere depends on only the leach rates ($k_i$) and the initial inventory ($C_{i0}$). The source term for $^{129}$I is given by

$$\frac{dC_i}{dt} = -\lambda_i C_i, \quad t \leq T,$$

$$\frac{dC_i}{dt} = -\lambda_i C_i - k_i C_i, \quad t > T,$$

where $C_i$ (mols) is the amount of $^{129}$I and $\lambda_i$ (yr$^{-1}$) and $k_i$ (yr$^{-1}$) are the decay rate and the leaching rate for $^{129}$I. The initial condition is $C_i(t = 0) = C_{i0}$; that is, the amount of $^{129}$I at the time of vault closure (Table 1). The source term for $^{237}$Np, the first element of the chain, is also described by (1), in which the parameters $\lambda_{Np}$ and $k_c$ are used. Here $k_c$ is the leaching rate for the radionuclides of the chain. The source term for $^{233}$U is given by

$$\frac{dC_u}{dt} = -\lambda_u C_u + \lambda_{Np} C_{Np}, \quad t \leq T,$$

$$\frac{dC_u}{dt} = -\lambda_u C_u + \lambda_{Np} C_{Np} - k_c C_u, \quad t > T.$$

The source term for $^{229}$Th is described similarly by (2), where $U$ is replaced by $Th$ and $Np$ is replaced by $U$. The migration through the geosphere is the core of the model. For example, the migration of $^{233}$U is governed by

$$R_U \frac{dF_U^{(k)}}{dt} = v^{(k)} d^{(k)} \frac{d^2 F_U^{(k)}}{dx^2} - v^{(k)} \frac{dF_U^{(k)}}{dx} - \lambda_U R_U F_U^{(k)} + \lambda_{Np} R_{Np} F_{Np}^{(k)},$$

where $U$ stands for the isotope $^{233}$U, $Np$ stands for $^{237}$Np, $(k)$ refers to geosphere layer number $k$ (1 or 2), $R_i$ is the retardation coefficient for nuclide $i$ (dimensionless), $F_i(x,t)$ is the flux (amount transported per unit time) of nuclide $i$ in the geosphere at position $x$ and time $t$ (mols/yr), $v^{(k)}$ is the water travel velocity in the $k$th geosphere layer (m/yr), $d^{(k)}$ is the dispersion length in the $k$th geosphere layer (m), and $\lambda_i$ is the decay constant of nuclide $i$ (yr$^{-1}$).

The same equation holds for $^{229}$Th provided that the index $U$ is replaced by $Th$ and the index $Np$ by $U$. To simplify the model structure, the retardation coefficients $R_U$, $R_{Np}$, and $R_{Th}$ were replaced in the Level E exercise by a single parameter $R_c$. The equation for $^{129}$I is

$$R_{I_{129}} \frac{dF_{I_{129}}^{(k)}}{dt} = v^{(k)} d^{(k)} \frac{d^2 F_{I_{129}}^{(k)}}{dx^2} - v^{(k)} \frac{dF_{I_{129}}^{(k)}}{dx} - \lambda_{I_{129}} F_{I_{129}}^{(k)}.$$  

### Table 1. List of Input Factors for the Level E Model

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
<th>Distribution</th>
<th>Range</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>Containment time</td>
<td>Uniform</td>
<td>[100, 1000]</td>
<td>yr</td>
</tr>
<tr>
<td>$k_i$</td>
<td>Leach rate for iodine</td>
<td>Log-uniform</td>
<td>[10$^{-2}$, 10$^{-1}$]</td>
<td>mols/yr</td>
</tr>
<tr>
<td>$k_{Np}$</td>
<td>Leach rate for Np chain nuclides</td>
<td>Log-uniform</td>
<td>[10$^{-6}$, 10$^{-5}$]</td>
<td>mols/yr</td>
</tr>
<tr>
<td>$v^{(1)}$</td>
<td>Water velocity in geosphere’s 1st layer</td>
<td>Log-uniform</td>
<td>[10$^{-3}$, 10$^{-1}$]</td>
<td>m/yr</td>
</tr>
<tr>
<td>$l^{(1)}$</td>
<td>Length of geosphere’s 1st layer</td>
<td>Uniform</td>
<td>[100, 500]</td>
<td>m</td>
</tr>
<tr>
<td>$R_{(1)}$</td>
<td>Retention factor for I (1st layer)</td>
<td>Uniform</td>
<td>[1, 5]</td>
<td>—</td>
</tr>
<tr>
<td>$R_{C_{(1)}}$</td>
<td>Factor to compute retention</td>
<td>Uniform</td>
<td>[3, 30]</td>
<td>—</td>
</tr>
<tr>
<td>$v^{(2)}$</td>
<td>Water velocity in geosphere’s 2nd layer</td>
<td>Log-uniform</td>
<td>[10$^{-2}$, 10$^{-1}$]</td>
<td>m/yr</td>
</tr>
<tr>
<td>$l^{(2)}$</td>
<td>Length of geosphere’s 2nd layer</td>
<td>Uniform</td>
<td>[50, 200]</td>
<td>m</td>
</tr>
<tr>
<td>$R_{(2)}$</td>
<td>Retention factor for I (2nd layer)</td>
<td>Uniform</td>
<td>[1, 5]</td>
<td>—</td>
</tr>
<tr>
<td>$R_{C_{(2)}}$</td>
<td>Factor to compute retention</td>
<td>Uniform</td>
<td>[3, 30]</td>
<td>—</td>
</tr>
<tr>
<td>$W$</td>
<td>Stream flow rate</td>
<td>Log-uniform</td>
<td>[10$^{0}$, 10$^{5}$]</td>
<td>m$^3$/yr</td>
</tr>
<tr>
<td>$C_{0_{129}}$</td>
<td>Initial inventory for $^{129}$I</td>
<td>Constant</td>
<td>100</td>
<td>mols</td>
</tr>
<tr>
<td>$C_{0_{Np}}$</td>
<td>Initial inventory for $^{237}$Np</td>
<td>Constant</td>
<td>1000</td>
<td>mols</td>
</tr>
<tr>
<td>$C_{0_{U}}$</td>
<td>Initial inventory for $^{233}$U</td>
<td>Constant</td>
<td>1000</td>
<td>mols</td>
</tr>
<tr>
<td>$C_{0_{Th}}$</td>
<td>Initial inventory for $^{229}$Th</td>
<td>Constant</td>
<td>1000</td>
<td>mols</td>
</tr>
<tr>
<td>$w$</td>
<td>Water ingestion rate</td>
<td>Constant</td>
<td>.73</td>
<td>m$^3$/yr</td>
</tr>
<tr>
<td>$\beta_{I_{129}}$</td>
<td>Ingestion-dose factor for $^{129}$I</td>
<td>Constant</td>
<td>56</td>
<td>Sv/mols</td>
</tr>
<tr>
<td>$\beta_{Np}$</td>
<td>Ingestion-dose factor for $^{237}$Np</td>
<td>Constant</td>
<td>6.8 x 10$^3$</td>
<td>Sv/mols</td>
</tr>
<tr>
<td>$\beta_{U}$</td>
<td>Ingestion-dose factor for $^{233}$U</td>
<td>Constant</td>
<td>5.9 x 10$^3$</td>
<td>Sv/mols</td>
</tr>
<tr>
<td>$\beta_{Th}$</td>
<td>Ingestion-dose factor for $^{229}$Th</td>
<td>Constant</td>
<td>1.8 x 10$^6$</td>
<td>Sv/mols</td>
</tr>
</tbody>
</table>
The same equation holds for $^{237}Np$ provided that the index $I$ is replaced by $Np$. Modeling of the biosphere is extremely simplified; via an ingestion factor and the water consumption rate, the dose to the most exposed individual of a hypothetical critical group is computed. The radiologic dose (measured in Sv/yr) from nuclide $i$ is given by

$$dose_i(t) = \beta_i \frac{w}{W} F_i^{(2)}(I^{(2)}, t), \quad i = 129I, 237Np, 233U, 229Th,$$

where $\beta_i$ is an ingestion–dose conversion factor and is assumed fixed, $F_i^{(2)}(I^{(2)}, t)$ is the flux at the end of the second layer (the output to the biosphere), $w$ denotes the drinking water requirement for an individual in the most exposed critical group, and $W$ is the stream flow rate. The quantity of interest in this study is the annual radiologic dose due to the four radionuclides,

$$Y(t) = \sum_i dose_i(t).$$

(6)

The overall predictive uncertainty about that dose is due to uncertainties in model parameters, both intrinsic (or stochastic), such as the time of canister failure, and due to our poor knowledge of the system, such as a poorly known kinetic parameter. The simulation model includes 12 uncertain input factors, listed in Table 1 together with a set of parameters that are assumed constant.

The probability distributions for each factor were selected on the basis of expert judgement. Such data refer to the original formulation of Level E (OECD 1989). Although we were aware of the existence of correlations (e.g., between $R_I$ and $R_C$ for each layer of the geosphere), these were omitted in the original system specifications for the sake of simplicity. For a subset of our simulations, we have reinstated these correlations by assuming that a serious fault failure induces high leaching for both iodine and the chain elements, and vice versa. Also, high release coefficients for iodine should be accompanied by high release coefficients for the chain elements, within a given geosphere layer. The geochemical properties in the two layers of the geosphere could hardly be independent from each other, and the water flows in the two layers should show a certain degree of dependency. Besides, the time of containment failure is likely correlated to the flow in the first layer, because corrosion will be faster if the flow is faster. The proposed correlation pattern is given in Table 2. The set of correlation values was defined by consulting with the authors of the benchmark (P. C. Robinson 2000, personal communication).

To replicate the Level E exercise, the system (1)–(6) needs to be solved numerically. We used the Crank–Nicholson method (see Crank 1975; Prado, Homma, and Saltelli 1991). As an alternative, the system can be solved using an algorithm for accurate numerical inversion of the solution in the Laplace space to obtain the solution in the real space (Robinson and Hodgkinson 1987). We tackled below both cases (correlated and noncorrelated) using two alternative definitions of importance.

2.2 Previous Work on Level E

Following the Level S benchmark, a number of authors attempted to address the concept of importance in sensitivity analysis using the Level E as a key test case. Saltelli and Homma (1992) suggested an early version of the first-order conditional variance $V_j$ (Hora and Iman 1989; Ishigami and Homma 1990) as a sensitivity analysis tool for nonmonotonic models. We treat this measure in Section 3. Saltelli, Andres, and Homma (1993) improved the estimation procedure for $V_j$ and applied it to the Level E example. The results are reported in Section 5.1. A version of $V_j$ based on the ranks of the output values, rather than the values themselves, was also suggested to reduce the relative error of the estimates. It was realized later (Saltelli and Sobol’ 1995), using, among others, the Level E case, that the improved reproducibility of the rank-based measure came at the cost of substantial alteration of the sensitivity pattern of the model output $Y$ considered. The sensitivity pattern of “rank of $Y$” was substantially different from the sensitivity pattern of $Y$.

Level E was instrumental in the development of the total sensitivity indices $S_Tj$ (Homma and Saltelli 1996), described in Section 4.1. In brief, the fact that some interaction terms were important for this model led to the development of a sensitivity measure incorporating both the first-order and higher-order effects of a given factor on model output.


3. LOTTERY SETTING 1

How do we judge the relative importance of model input factors? We assume to have information about the factors’ probability distribution, either joint or marginal, with or without correlation, and that this knowledge comes from measurements, estimates, expert opinion, physical bounds, output from simulations, analogy with factors for similar species, and so forth.

We also assume as good practice in model use that from the many output variables generated by the model, we select

Table 2. Configuration for Correlated Input of the Level E Model

<table>
<thead>
<tr>
<th>Pairs of correlated factors</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_I, k_C$</td>
<td>.5</td>
</tr>
<tr>
<td>$R_I^{(1)}, R_C^{(1)}$</td>
<td>.3</td>
</tr>
<tr>
<td>$R_I^{(2)}, R_C^{(2)}$</td>
<td>.3</td>
</tr>
<tr>
<td>$T_I^{(1)}, T_C^{(1)}$</td>
<td>-.7</td>
</tr>
<tr>
<td>$v_I^{(1)}, v_C^{(1)}$</td>
<td>.5</td>
</tr>
<tr>
<td>$R_I^{(1)}, R_C^{(2)}$</td>
<td>.5</td>
</tr>
<tr>
<td>$R_C^{(1)}, R_C^{(2)}$</td>
<td>.5</td>
</tr>
</tbody>
</table>
a single top-level variable, \( Y \), that summarizes the thesis that the model is supposed to prove or disprove (Saltelli, Tarantola, and Campolongo 2000). Examples of \( Y \) include the extent of a geographic area where the concentration of a given pollutant exceeds a given threshold value, the estimated failure probability of a given system in a given time range, and the risk that the return of a given portfolio, falls below a preestablished value within a given time span. As a general rule, the functional of interest is selected so that it is meaningful with respect to the question asked of the model.

Let us assume that we can extract a sample from the probability distribution of the input factors and that we can compute the model output for every sample point. Also assume that each factor has a true, albeit unknown, value. We are hence poised to stipulate the following setting.

**Lottery Setting 1.** We are asked to bet on the factor that, if determined (i.e., fixed to its true value), would lead to the greatest reduction in the variance of the \( Y \).

The factors could be ranked according to \( V(Y|X_i = x_i^*) \) or, equivalently, to \( V(Y|X_i = x_i^*) / V(Y) \), where \( V(Y) \) is the unconditional variance of \( Y \) and \( V(Y|X_i = x_i^*) \) is the variance obtained by fixing \( X_i \) to its true value \( x_i^* \). The problem is that \( x_i^* \) is unknown for each \( X_i \).

It would sound sensible to bet on the factor with the smallest value of \( E[V(Y|X_i)] \), that is, the average of the conditional variance \( V(Y|X_i = x_i^*) \) evaluated over all possible values \( x_i^* \) of \( X_i \). Given that

\[
V = V[E(Y|X_i)] + E[V(Y|X_i)],
\]

betting on the lowest \( E[V(Y|X_i)] \) is equivalent to betting on the highest \( V[E(Y|X_i)] \).

Unsurprisingly, many practitioners of sensitivity analysis have come up with different estimates of \( V[E(Y|X_i)] \) (in short, \( V_i \)) as a measure of sensitivity. Some have called this the **importance measure** (see Chan, Tarantola, Saltelli, and Sobol 2000 for a review). Statisticians, and practitioners of experimental design, call \( V_i \) the first-order effect of \( X_i \) on \( Y \).

### 3.1 Noncorrelated Factors

The factors \( X_i \) are noncorrelated. Assume that \( V_j \) and \( V_i \), related to \( X_j \) and \( X_i \), are both smaller than \( V_j \), associated with \( X_j \). However, \( X_i \) and \( X_j \) may be involved in an interaction effect that is not captured by \( V_i \) or \( V_j \). How can we establish whether \( X_j \) is actually more important than \( X_i \) and \( X_j \) in lottery setting 1? The interaction effect between \( X_i \) and \( X_j \) is given by

\[
V_{ijr} = V[E(Y|X_i, X_j)] - V[E(Y|X_i)] - V[E(Y|X_j)],
\]

where \( V[E(Y|X_i, X_j)] \) is the joint effect of \( X_i \) and \( X_j \) and \( V_{ijr} \) is the pure interaction of \( X_i \) and \( X_j \), known as a second-order (or two-way) effect. Analogous formulas can be written for higher-order terms. If there are no interactions, then

\[
V = \sum_i V_i
\]

and we can establish the relative importance of the factors using the respective \( V_j \) values. If the model contains interactions, then

\[
V = \sum_i V_i + \sum_{i < j} V_{ij} + \sum_{i < j < m} V_{ijm} + \ldots + V_{12\ldots k};
\]

that is, the sum of all the first-order and higher-order terms adds up to the total unconditional variance \( V \) (Sobol’ 1990). \( V_{ijm} \) is the third-order interaction among \( X_i \), \( X_j \), and \( X_m \). \( V_{12\ldots k} \) is the \( k \)th-order interaction among all of the factors. It can be easily shown that, even when higher-order terms in (10) are nonzero, the \( V_i \) are sufficient to make an informed choice on lottery setting 1, based on the very definition of this measure. If \( V_j = V[E(Y|X_j)] \) is greater than \( V_i \) and \( V_j \), then \( V_i \) is the most important factor, regardless of the value of \( V_j \). For \( V_i \) to come into play, one should stipulate a setting in terms of “what couple of factors should be fixed” and so forth.

**Conclusion 1.** In lottery setting 1, for noncorrelated factors, the \( V_i \)’s are the measure to use to make an informed choice, whether or not the factors interact.

### 3.2 Correlated Factors

When the \( X_j \) are correlated, the output variance cannot be decomposed as in (10). Furthermore, the \( V_j \) are more expensive to obtain because the computational shortcuts available for noncorrelated inputs (see, e.g., Sobol 1990) are no longer applicable. However, a rather efficient estimation procedure for the first-order terms due to McKay (1995) uses the replicated Latin hypercube sampling (LHS) design (\( r - \text{LHS} \)). The correlation in the \( r - \text{LHS} \) sample is induced using the permutation procedure proposed by Iman and Conover (1982).

Leaving aside the computational issues, and going back to our lottery setting 1, we observe that \( V_j \) is again the measure to use for making an informed choice, even when the input is correlated. This descends from the definition of the setting, which implies the “fixing” of a single factor when computing the conditional variance. Note that \( V_j \) will be sensitive in general to the existence of correlations. Another observation relevant to the case of correlated input is that although in lottery setting 1, the fact that factors are fixed one at a time would normally prevent the detection of interactions, yet in the presence of correlation, fixing one factor also influences the distribution of the others. This may allow the influence of interactions to emerge, depending on the relative patterns of correlation and interactions. As a trivial example, if the model is \( Y = x_1 + x_2 + x_3 + a_{x3} \cdot x_2 x_3 \) and \( x_1 \) and \( x_2 \) are correlated, then \( V(E(Y|X_i)) \) will depend on \( a_{x3} \) as well as on the distribution parameters of \( x_2 \) and \( x_3 \). In this sense, we say that interactions may be “carried over” by correlation. This effect is a possibility only when the input is correlated, and is absent when the input is not correlated. As a result, an important consequence of adopting lottery setting 1 is that we accept the risk of remaining ignorant about important features of the model that are the objects of the sensitivity analysis: the presence of interactions. We are hence ready to offer our second conclusion.
Conclusion 2. In Lottery Setting 1, the terms \( V_j \) are the measures that guarantee an informed choice in the cases where the factors do not interact and do not correlate, interact but do not correlate (Conclusion 1), and correlate and interact.

4. LOTTERY SETTING 2

This context is relevant to the broad area of risk assessment and management, to which Level E belongs. Here one seeks to achieve a preestablished reduction in the variance of \( Y \).

Lottery Setting 2. We are asked to bet on sets of factors (couples, triplets, and so on). The prize is for obtaining a variance of \( Y \) equal or smaller than a given target variance, \( V_{\text{tar}} < V \), by fixing simultaneously the smallest number of factors.

We introduce first a more compact notation, with the superscript \( c \) standing for “closed” within a subset of factors. For instance, \( V_{jm} = V[E(Y|X_j, X_m)] \) and \( V_{npq} = V[E(Y|X_n, X_p, X_q)] \).

4.1 Noncorrelated Factors

Assume that the model is additive (i.e., there are no interactions) and that the target reduced variance is \( V_{\text{tar}}/V = .1 \), which corresponds to 90% reduction in the variance of \( Y \). Given that \( V = \sum V_j \), the factors can be ranked in order of importance using the \( V_j \) (i.e., \( V_{R_1} \geq V_{R_2} \geq \cdots \geq V_{R_k} \)). Then we select the first \( r \) factors such that \( \sum_{i=1}^{r} V_{R_i} \geq V - V_{\text{tar}} \).

If the model is nonadditive (i.e., there are interactions between factors), then we could use this procedure to select the smallest set of \( V_{t_1, t_2, \ldots, t_i} \) such that their sum just about exceeds \( V - V_{\text{tar}} \):

1. Take the factor with the highest \( V_j \). If \( V_j > V - V_{\text{tar}} \), then end the procedure; otherwise, go to step 2.
2. Evaluate all of the second-order partial variances (i.e., \( k(k-1)/2 \) terms) and take the pair \( X_j \) and \( X_m \) with the highest \( V_{jm} \) (neither \( X_j \) nor \( X_m \) must necessarily coincide with the \( X_i \) considered in step 1). If \( V_{jm} > V - V_{\text{tar}} \), then choose \( X_j \) and \( X_m \); otherwise, go to step 3.
3. Evaluate all of the third-order partial variances (i.e., \( k(k-1)(k-2)/6 \) terms) and take the triplet \( X_n, X_p, X_q \) with the highest \( V_{npq} \). If \( V_{npq} > V - V_{\text{tar}} \), then choose \( X_n, X_p, X_q \), and so on with higher-order terms.

The number of terms \( V_{t_1, t_2, \ldots, t_i} \) required is not fixed a priori, and the procedure might be computationally impracticable even for moderate values of \( k \). We recommend a different strategy for this setting that requires using the total sensitivity index \( S_{R_j} = V_{R_j}/V \), where the \( V_{R_j} \) for a given factor \( X_j \) is the sum of all of the terms in (10) that include the index \( j \). For instance, in a model with three factors, the \( V_{R_j} \) are:

\[
\begin{align*}
V_{R_1} &= V_1 + V_{12} + V_{13} + V_{123}, \\
V_{R_2} &= V_2 + V_{12} + V_{23} + V_{123}, \\
V_{R_3} &= V_3 + V_{13} + V_{23} + V_{123}.
\end{align*}
\]

4.2 Correlated Factors

If the input factors are correlated, then we are not allowed to relate \( V_{\text{ir}} = V[E(Y|X_i, X_j)] \) to the sum of a “purely interaction” term \( V_{\text{ir}} \) and the individual effects \( V_{X_i} \) and \( V_{X_j} \), as in (8), although we can still compute \( V_{\text{ir}} \). Furthermore, if either \( X_j \) or \( X_j \) are correlated with a third factor \( X_j \), then the \( V_{\text{ir}} \) measure will depend on the strength of such correlation as well as on the distribution of \( X_j \).

It should also be recalled that the estimation of the conditional variances of any order is much more expensive computationally in case of correlated input (see Sec. 5.4 for details). For this reason, the number of conditional variances must be kept to a minimum. Our proposed procedure for correlated input is as follows:

1. Rank the factors in order of importance using the \( V_i \) and obtain a sequence \( V_{R_1}, V_{R_2}, \ldots, V_{R_k} \), where \( V_{R_1} \geq V_{R_2} \geq \cdots \geq V_{R_k} \). If \( V_{R_1} \geq V - V_{\text{tar}} \), then \( X_{R_1} \) is the choice; end the procedure. Otherwise, go to step 2.
2. Compute the second-order term \( V_{R_1 j} \), that is, the reduction of the variance of \( Y \) that can be achieved by fixing the pair \( (X_{R_1}, X_j) \). \( X_j \) is the factor that exhibits the highest “figure of merit” \( M_j \) (defined later). If \( V_{R_1 j} > V - V_{\text{tar}} \), then \( X_{R_1} \) is the choice; otherwise, go to step 3.
3. Compute the third-order term \( V_{R_1 j m} \), where \( X_m \) is selected using the figure of merit. If \( V_{R_1 j m} > V - V_{\text{tar}} \), then \( (X_{R_1}, X_j, X_m) \) is the choice, and so on.

The figure of merit \( M_j \) is defined as:

\[
M_j = S(R_j) \left( 1 - \max_{i \neq j} c_{ij} \right) \left( 1 + \frac{V_{NC} - V_{NC}^j}{V_{NC}} \right) \left( \frac{V_{R_j}}{V_{NC}} \right),
\]

where \( S(R_j) = \sum_{i \neq j} c_{ij} \) is a Savage score (Savage 1956). Equation (12) means that the rank \( R_j \) of the candidate factor \( X_j \) provided at step 1 is first converted into a Savage score and is then penalized or prized depending on its correlations and interactions. Here \( u \) is the set of the input factors already
selected, and $c_{ij}$ is the correlation coefficient between $X_i$ and $X_j$, which is known a priori. A candidate factor for inclusion in $u$ that is correlated with some of the factors already in $u$ must be penalized, because part of its effect on the output variance might have already been offset by fixing the set in $u$. This penalization should be proportional to the degree of correlation. The resulting score is then prized if $X_j$ interacts with other factors already in $u$, because this could lead to a higher variance reduction than is implied by its $V_j$. The extent of the interaction is quantified by $V_{ij}^{NC} - V_j^{NC}$, where the superscript “NC” means that the measures are computed for the noncorrelated problem. Empirical tests have shown that the figures of merit are effective when the prize for interaction is squared.

5. THE LEVEL E IN SETTINGS 1 AND 2

5.1 Setting 1 With Noncorrelated Factors

The output of interest is the total radiologic dose, $Y(t)$, from (6), calculated at $t = 10^4$ yr in the future. This time point is special because it corresponds to the lowest value for the model coefficient of determination on ranks $R^2_y$ (Fig. 1), where the model shows a strong nonadditive, nonmonotonic behavior.

We express computational costs in terms of model evaluations, that is, the number of times the model must be evaluated. Because $Y$ may be the output of an expensive computer code, it is clear that effective sensitivity analysis methods must strive to keep this number low. The method of Sobol’ (Sobol’1990), improved by Saltelli (2002), is used to estimate all of the $S_j$. Circles refer to the values of $R^2_y$ across time is almost identical to that in Figure 1 for the noncorrelated case. We have estimated all of the $V_j$, and hence the first-order sensitivity indices $S_j = V_j / V$.

A base sample of 1,024 points is considered. The total cost needed to estimate all the $V_j$ is $1,024 \times (12 + 2) = 14,336$ model evaluations, where two extra samples of 1,024 points each are taken to obtain two estimates of the output mean. (Details on sample size and computational cost have been given in Saltelli 2002.)

The analysis shows that the most important factors are $v^{(1)}$ and $R^{(1)}$, with $S_j$ equal to .18 and .09. The sum of all the $S_j$ is .38, meaning that the first-order effects explain 38% of the uncertainty in the total dose, with the remaining portion explained by interactions.

The ideal use for the lottery setting 1 is for the prioritization of research, one of the most common uses of sensitivity analysis. Under the hypothesis that all uncertain factors are susceptible for determination (at the same cost per factor), lottery setting 1 allows identification of the factor most deserving an experimental measurement. If we are in a forecast (prognostic) setting and want to reduce the overall uncertainty in our predicted dose, then $v^{(1)}$ is the factor whose uncertainty should be reduced first. If we are in a diagnostic or calibration setting, then again $v^{(1)}$ is the factor that has the best chance of being calibrated. The total cost of this exercise is 14,336 model runs.

5.2 Setting 1 With Correlated Factors

Let us consider Level E with the input correlation structure given in Table 2. The output of interest is again the total radiologic dose, $Y(t)$, at $t = 10^4$ yr. The pattern of $R^2_y$ and $R^2$ across time is almost identical to that in Figure 1 for the noncorrelated case.

We have estimated all of the $V_j$ using $r = \text{LHS}$. An $r = \text{LHS}$ consists of $r$ replicates of the LHS obtained as suitable permutations of all the columns of a base LHS sample of size $n$. We used $n = 500$ and $r = 20$. The results are listed in Table 3. With respect to the noncorrelated case, $v^{(1)}$ has lost its leadership due the high correlation with $T$ and $v^{(2)}$, which are less influential. This result shows that the omission of correlations in the original specification of the Level E exercise was legitimate as far as uncertainty analysis is concerned (the output dose does not change much), but was an oversimplification as far as the identification of the influential factors is concerned.

Note that for the noncorrelated case of Level E, $\sum V_j = 2 \times 10^{-15}$ and $V = 5.3 \times 10^{-15}$, the difference being attributable

<table>
<thead>
<tr>
<th>Factor</th>
<th>$V_j \cdot 10^{15}$</th>
<th>$S_j = V_j / V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>91.0</td>
<td>.08</td>
</tr>
<tr>
<td>$v^{(1)}$</td>
<td>38.4</td>
<td>.03</td>
</tr>
<tr>
<td>$R^{(1)}$</td>
<td>11.0</td>
<td>.01</td>
</tr>
<tr>
<td>$T$</td>
<td>8.7</td>
<td>.01</td>
</tr>
<tr>
<td>$R^{(2)}$</td>
<td>6.8</td>
<td>.01</td>
</tr>
<tr>
<td>$v^{(2)}$</td>
<td>6.6</td>
<td>.01</td>
</tr>
<tr>
<td>$j^{(1)}$</td>
<td>6.4</td>
<td>.01</td>
</tr>
<tr>
<td>$j^{(2)}$</td>
<td>6.1</td>
<td>.01</td>
</tr>
<tr>
<td>$R^{(1)}$</td>
<td>5.9</td>
<td>.00</td>
</tr>
<tr>
<td>$R^{(2)}$</td>
<td>2.1</td>
<td>0</td>
</tr>
<tr>
<td>$k_c$</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>$k_y$</td>
<td>1.0</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: The factors are ranked in decreasing order of importance with respect to the $S_j$. 

Figure 1. Model Coefficient of Determination for the Level E Case Study as a Function of Time (yr). Circles refer to the values of $R^2_y$; diamonds, to the values of $R^2$ (i.e., obtained using the ranks of the model output values). The output variable is the total annual dose, $V$. The poor performance of the regression model is highlighted by the very small value of $R^2_y$ at every time point. At $t = 10^4$ yr, $R^2$ also is very low, indicating that even rank-based statistics are inappropriate for the purpose of sensitivity analysis. $R^2_y$ and $R^2$ have been estimated using the sequence of output values obtained by running the Level E model on a Latin hypercube sample of size 1,000 generated over the space of the inputs.
to interactions. For the correlated case $\sum_j V_j = 1.85 \times 10^{-15}$ and $V = 1.18 \times 10^{-14}$. When input is correlated, we cannot conclude that the missing fraction of variance is that due to interaction, because (10) does not hold. The fraction due to interaction could in fact be higher than the difference $V - \sum_j V_j$. The total cost of this exercise is 10,000 model runs.

5.3 Setting 2 With Noncorrelated Factors

We now apply the procedure of Section 4.1 to Level E with noncorrelated input. We assume that the targeted variance reduction in the model output is 40%. The technique proposed by Saltelli (2002) enables us to estimate all of the $V_j$ and the $V_{Tj}$ (step 1 of the procedure) for the 12 factors at the cost of 14,336 model runs, as in Section 5.1. The model is strongly nonadditive given that, as mentioned in Section 5.1, $\sum_j V_j^{NC}/V = .38$. Sizeable variance reductions can be achieved only by fixing more factors simultaneously. The top-ranked factor is $v^{(1)}$ with the highest $V_{Tj}^{NC}$ (step 2), but $V_{v^{(1)}}^{NC}/V = .18$ is below the target imposed (step 3). $W$ is the factor with the second-highest $V_{Tj}^{NC}$; we estimate $V_{v^{(1)W}}$ with the method of Sobol’ at the cost of 1,024 model evaluations (step 4). The target variance reduction is still not reached (28%). We need to enter the third-ranked factor, $l^{(1)}$, and estimate $V_{v^{(1)Wl}}$ using the method of Sobol’ with another 1,024 simulations. The variance reduction obtained by simultaneously fixing $v^{(1)}$, $W$, and $l^{(1)}$ is 51%, which enables us to meet the target. This result is remarkable and novel. It shows that we can achieve a considerable overall variance reduction by judiciously fixing a set of interacting factors (a triplet in this case). To a certain extent, we can generalize these results to problems of mass transfer with chemical reaction. We know that factors that are downstream in the modeling chain and that influence the output linearly ($W$) tend to interact with those that govern the transit time, $v^{(1)}$ and $l^{(1)}$. The total cost of this exercise is $14,336 + 1.024 + 1.024 = 16,384$ model runs.

5.4 Setting 2 With Correlated Factors

We now apply the procedure of Section 4.2 to Level E with correlated input, with the same target variance reduction (40%). The ranking provided by the $V_j$ (step 1) yields $W$ as the top-ranked factor with $S_i = 8%$ (see Table 3). The cost of estimating the $V_j$ is 10,000 runs, as explained in Section 5.2. We compute the figures of merit $M_j$ as from (12) (Table 4). This requires other 14,336 runs, as explained in Section 5.3. The largest $M_j$ is that for $v^{(1)}$. According to step 2, we compute $V_{v^{(1)W}}$. To obtain this estimate, we must compute the inner conditional expectation $E(Y|W, v^{(1)})$ through a numerical integration over a base sample of, say, $N_b = 100$ points. The same conditional expectation has to be evaluated over a set of different values of the pair $(W, v^{(1)})$ (say, $N_v = 100$), to compute the variance of the conditional expectations. Therefore, this estimate requires $N = N_b \times N_v = 10,000$ model simulations. The target is not reached, because a variance reduction of 32% is obtained by fixing the pair $(W, v^{(1)})$. The factor $R_l^{(1)}$ (with $M_l = 2.43$) is included in the analysis (step 3), and the term $V_{v^{(1)Wl}}$ is estimated (requiring another 10,000 runs); the target is reached because the variance reduction rises to about 48%. (The variance reduction grows rapidly if we are able to fix the “right” factors.) The result of the correlation between factors, with respect to the uncorrelated case is that $R_l^{(1)}$ is now a better partner for $W$ and $v^{(1)}$ for obtaining the maximum output variance reduction. If $l^{(1)}$ instead of $R_l^{(1)}$ were fixed, then the output variance reduction would be less, that is, 43%. The total cost of this exercise is $10,000 + 14,336 + 10,000 + 10,000 = 44,336$ model runs.

6. CONCLUSIONS

The motivation of this article is to devise a strategy for sensitivity analysis that could work even in the simultaneous presence of correlated input and nonadditive models. To establish objective criteria for the evaluation of the proposed strategy, we have tried to define in a rigorous fashion the questions posed by the sensitivity analysis. The literature shows that many different approaches and framing assumptions result in different definitions of importance and sensitivity. Without a rigorous problem setting, alternative definitions of importance are equally plausible (see OECD 1993).

The two lottery settings proposed in this article cover important areas of application for sensitivity analysis. The first setting can be used for priority setting, to identify which factor has the greatest potential for output uncertainty reduction through additional research. The setting also applies to calibration analysis; it would help to design an experiment in such a way that the factors to be measured, in the presence of other nonreducible uncertainties, are the most important in lottery setting 1. For this setting, the $V_j$’s represent the proper measure to use whatever the correlation and interaction structure of the model.

The second setting looks at the compounded effects of factors in a model and could represent the setting in which “risk” is involved. Extreme events, those of interest to risk analysts, often occur when interacting factors assume their most unfavorable values concurrently. Lottery setting 2 aims to identify the smallest number of factors to be fixed to obtain the maximum reduction in the variance of the output. Control theory is another area in which this setting might be of use.

For the two settings, we have proposed procedures to fight the curse of dimensionality, for example, the fact that the number of conditional variances in a model grows exponentially.
with problem dimensionality. It is clear that importance assessments are more cumbersome for correlated inputs than for noncorrelated inputs, as many more evaluations of the model are needed. In the procedure suggested in Section 4.2, the number of conditional variances to be estimated remains on the order of $k$. Better (although much more expensive) results can be obtained by computing all possible conditional variances, $O(2^k)$, or using the procedure proposed by McKay, Morrison, and Upton (1999), $O(k(k-1)/2)$.

The high computational cost of quantifying importance of input factors makes the approach described in this article unfeasible for computationally expensive models (1 hour or more of CPU time per simulation, say). The present approach is instead reasonable for models in the seconds-to-minute range, which represent the vast majority of models. The test case addressed in this work offered two major practical advantages. On the one hand, it is very well known to practitioners in risk assessment and sensitivity analysis; on the other hand, it is representative of a large class of environmental or risk analysis models that deal substantially with a problem of mass transfer with chemical reaction. In our experience, for instance, the pattern of interaction displayed by Level E is very representative of situations found in hydrology, air pollution, and nuclear safety studies.

References


CPC, Computer Physics Communications, (1999), Special Issue on Sensitivity Analysis, 117(1–2), 1–188.


RESS, Reliability Engineering and System Safety, (1997), Special Issue on Sensitivity Analysis, 57(1).


