

Sequential screening with elementary effects

Hugo Maruri-A.

joint with Alexis Boukouvalas* and John Paul Gosling†

School of Mathematical Sciences, Queen Mary, University of London,

*Aston University and †Food and Environment Research Agency



Accelerating Industrial Productivity via Deterministic Computer
Experiments and Stochastic Simulation Experiments
Isaac Newton Institute, 6th September 2011

Sequential screening with elementary effects

The Elementary Effects (EE) method (Morris, 1991) is a simple but effective screening strategy. Starting from a number of initial points, the method creates random trajectories to then estimate factor effects. In turn, those estimates are used for factor screening. Recent research advances (Campolongo *et al.*, 2004, 2006) have enhanced the performance of the elementary effects method and the projections of the resulting design (Pujol, 2008).

The presentation concentrates on a proposal (Boukouvalas *et al.*, 2011 [2]) which turns the elementary effects method into a sequential design strategy. After describing the methodology, some examples are given and compared against the traditional EE method.

Analysis of complex computer simulators

- Rapidly evolving field, with numerous new applications springing, such as climate projections [6], investigation of biological processes [14] and estimation of national carbon balances [8]
- Emulators built as surrogate alternatives to expensive computer simulators [11].
- Emulators used in sensitivity analyses, prediction of response at new points.
- The analysis is still restricted to the number of input dimensions, and screening of input factors is required.
- Screening alternatives: Elementary Effects method [10], Fourier Amplitude Sensitivity Test (FAST), Sobol' variance decomposition, Sequential bifurcation (SB) [9].
- We look for a sequential procedure for screening to separate input factors with linear/non-linear effects.

Elementary Effects method for input screening

Simulator $Y(\cdot)$ is assumed smooth function in k inputs; design region is $[0, 1]^k$. The elementary effect for the i -th input variable at point $\mathbf{x} \in [0, 1]^k$ is defined as

$$EE_i(\mathbf{x}) = \frac{Y(\mathbf{x} + \Delta \mathbf{e}_i) - Y(\mathbf{x})}{\Delta}. \quad (1)$$

Divisor Δ is a fixed step size, and \mathbf{e}_i is the i -th unit vector. Each elementary effect is computed with observations at the pair of points \mathbf{x} , $\mathbf{x} + \Delta \mathbf{e}_i$.

The classic elementary effects approach [10] starts in a point \mathbf{x} , from which a trajectory is constructed with k random moves of size Δ , each movement in the direction of a coordinate axis, to end in the point $\mathbf{x} + \Delta(\mathbf{e}_1 + \dots + \mathbf{e}_k)$. Thus $k + 1$ evaluations of simulator $Y(\cdot)$ are performed, ending with k elementary effects $EE_1(\mathbf{x}), \dots, EE_k(\mathbf{x})$.

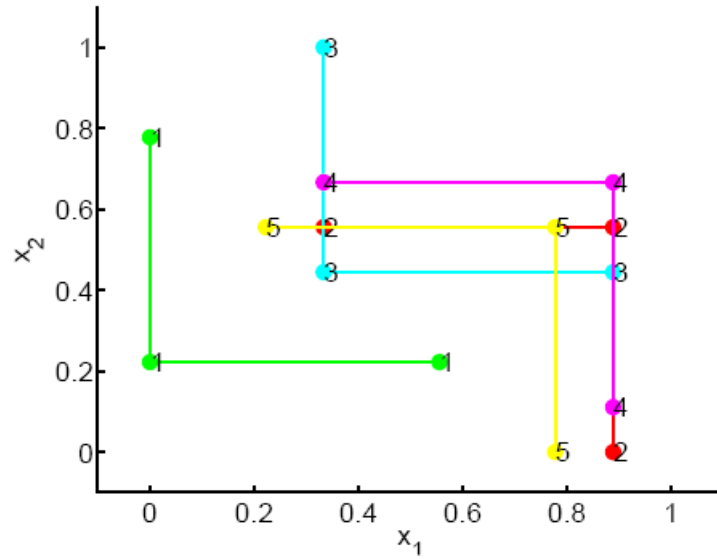
Elementary Effects method for input screening 2

Now consider a set of R points $\mathbf{x}_1, \dots, \mathbf{x}_R$ in the input space. At each point \mathbf{x}_r , we perform k one-at-a-time (OAT) runs and compute elementary effects $EE_i(\mathbf{x}_r)$ for every input factor and define sample moments:

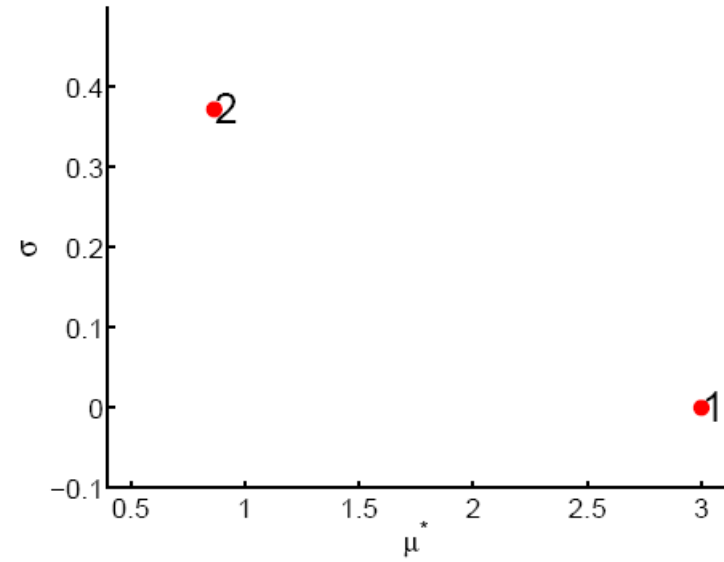
$$\mu_i = \frac{1}{R} \sum_{r=1}^R EE_i(\mathbf{x}_r) \text{ and } \sigma_i = \sqrt{\sum_{r=1}^R \frac{(EE_i(\mathbf{x}_r) - \mu_i)^2}{R - 1}}. \quad (2)$$

The sample moment μ_i is an average effect measure, and a high value suggests dominant contribution of the i -th input factor in positive or negative response values. Non-linear and interaction effects are estimated with σ_i .

Elementary Effects method for input screening 3



Trajectories



Moments

Elementary Effects method for input screening 4

The sample moment

$$\mu_i^* = \frac{1}{R} \sum_{r=1}^R |E E_i(\mathbf{x}_r)|$$

is a main effect measure; a high value indicates large influence of the corresponding input factor. The moment μ_i^* was proposed in [3] since μ_i may prove misleading due to cancelation of effects.

The total number of model runs needed in the Morris's method is $(k + 1) \times R$. Usually small values of R are used; for instance, [10] used $R = 3$ and $R = 4$ in his examples. A value of R between 10 and 50 is mentioned in the more recent literature, see [3,4].

Large values of R will improve the quality of the estimations, but at the price of extra runs.

Elementary Effects method for input screening 5

Step size Δ selected such that

- all the runs lie in the input space, and
- the elementary effects are computed within reasonable precision.

Usual choice based on a k dimensional grid constructed with p uniformly spaced values per input.

The number p is recommended to be even; Δ to be a multiple of $1/(p - 1)$, for example $\Delta = p/(2(p - 1))$, see [10,3]. Step Δ is usually kept at the same value for all the inputs, but method can be generalised to use different values of Δ and p for every input.

Points $\mathbf{x}_1, \dots, \mathbf{x}_R$ were taken at random from the input space grid [10]. In [4], runs are spread over design space by generating a large number of trajectories and selecting a subset that maximises the minimum distance between them.

Elementary Effects method for input screening 6

However, design points on the EE method fall on top of each other when projected into lower dimensions. This may prevent reusing runs after screening process.

An alternative is to put a randomly rotated simplex at every point \mathbf{x}_r , from which elementary effects are computed [12].

Computation of μ_i, μ_i^*, σ_i and further analysis is similar to the EE method, with the advantage that projections of the resulting design do not fall on top of existing points, and all observations can be reused in a later stage.

A potential disadvantage of this approach is the loss of efficiency in the computation of elementary effects.

Sequential screening for linear effects

We propose to separate between factors with linear effect and those with a non-linear effect.

If σ_i is small, we investigate if it remains so over other areas of the design region.

At the end of experimentation, those factors for which σ_i remained small are considered to have linear effect, and factors for which σ_i was bigger than a threshold have a non-linear effect on the output.

At a later stage, the emulator may be simplified with the results. If $A \subseteq \{1, \dots, k\}$ indexes linear factors, the GP may be

$$Y(x) = \beta_0 + \sum_{i \in A} \beta_i x_i + Z^*$$

where Z^* is a stochastic process whose covariance structure depends on variables with non linear effects, i.e. those x_i with $i \in \{1, \dots, k\} \setminus A$.

Sequential screening for linear effects 2

Algorithm

Input: Simulator $Y(\cdot)$ with k inputs; total number of one-at-a-time experiments M ; step size Δ ; threshold σ_0 .

Output: Moments μ_i, σ_i, μ_i^* ; lists of factors with linear (C) and with non-linear effect (A).

A Build (maximin) space filling design, order points using maximum distance between them: $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(M)}$.

B1 Starting with two most extreme points, build one-at-a-time runs and compute EEs and moments.

B2 Compare σ_i with σ_0 . If bigger, separate factor, update C , A .

B3 Continue adding points, updating EEs and moments. Repeat from B2 until runs exhausted or all factors screened.

Selecting the threshold σ_0

The variance threshold σ_0 is an input, but it may be quite hard in certain cases to suggest a value. We propose a very simple approach to this, in sharp contrast with full probabilistic elicitation about the beliefs about the parameters of a linear model ([7] and [5]).

A linear (or near-linear) effect of the variable x_i is represented by an additive noise model:

$$Y(x_i) = ax_i + b + \varepsilon_i, \quad (3)$$

where the ε_i are independent normal random variables with zero mean and variance γ . In other words, the marginal effect due to the factor x_i is modeled with a simple regression line.

Given the variance γ , the sampling distribution of the variance of the elementary effects can be calculated according to the following lemma.

Selecting the threshold σ_0 2

Lemma 1 Let x_1, \dots, x_R be univariate design points, at each of which trajectories are constructed. Assume that observations (taken at design points and trajectories) follow the model given in Equation (3). Let Elementary effects and moments be as before and let $\sigma_{\Phi}^2 = \frac{2\gamma}{\Delta^2}$. Then

$$\sigma^2 \sim \frac{\sigma_{\Phi}^2}{R-1} \chi_{R-1}^2.$$

We propose to use the 99% quantile of the distribution of σ^2 to determine the threshold $\sigma_0 = \chi_{0.99, R-1}^2 \sigma_{\Phi}^2 / (R-1)$.

Lemma 1 applies directly in a multivariate setting, in which case the comparison is performed separately for each input variable.

Threshold σ_0 may be fixed for computations rather than adapting to the number of trajectories. If so, the method becomes more conservative.

Rabies Model (FERA [13])

Wildlife rabies has been eradicated historically from large parts of Europe, but there is a remaining risk of disease re-introduction. The situation is aggravated by an invasive raccoon dog (*Nyctereutes procyonoides*) that can act as a second rabies vector in addition to the red fox (*Vulpes vulpes*).

The purpose is to analyse the risk of rabies spread in this new type of vector community [13].

Individual-based, non-spatial, time-discrete model that incorporates population and disease dynamical processes (host reproduction, mortality rates, disease transmission).

These processes are modelled stochastically to reflect natural variability (e.g. demographic stochasticity). Thus model analysis has to contend with stochastic, indeed heteroscedastic, model output [1].

Rabies Model



Rabies Model 2

The (Java) model consists of an input generation phase, the actual calculation of the model and two types of output: time series and summary statistics.

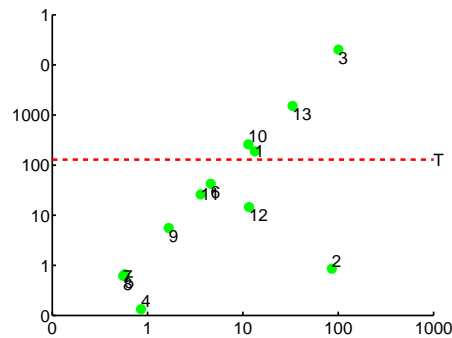
The simulator has 16 input factors. For each input factor, FERA has specified upper and lower bound values. Three parameters were kept fixed: the number of steps was 400; the cross infection input was 0.002 and area size was fixed at 5400 km^2 .

Although the simulator output is stochastic, the output of interest was the probability of extinction after five years.

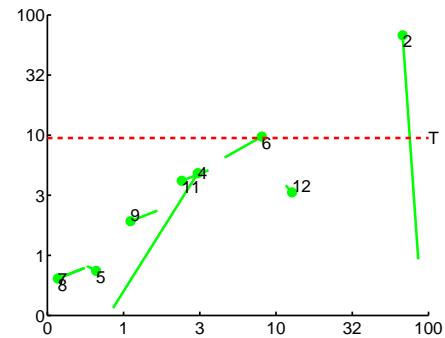
The output is hard bounded in the range $[0, 100]$ and a factor has near-linear effect if the output varies no more than 5.6% from linear. We thus set $\gamma = 3.5$ reflecting this requirement, i.e. $\pm 3\sqrt{\gamma} = \pm 5.6$.

A maximum of twenty design points were considered for the study.

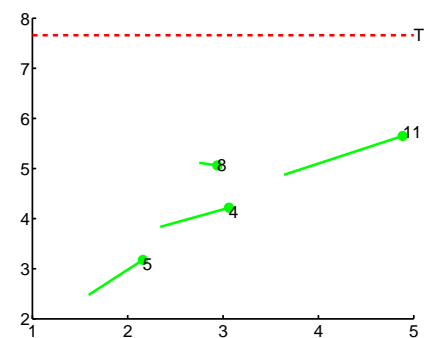
Rabies Model 3: Evolution of the procedure



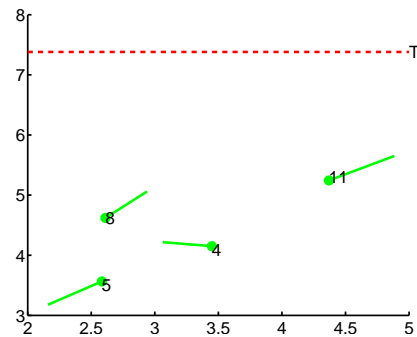
Step 1



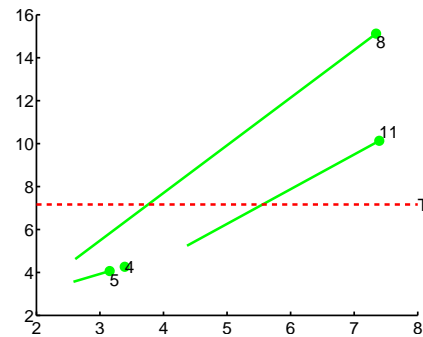
Step 2



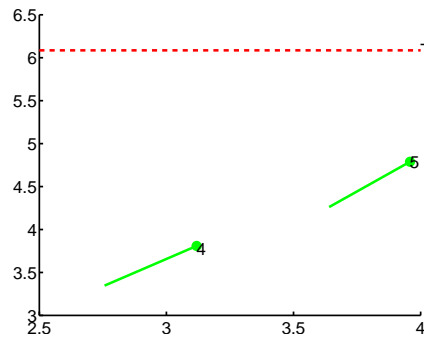
Step 5



Step 6



Step 7



Step 19

Two factors were found linear after 20 steps (102 simulator evaluations). Similar results were obtained with full Morris (280 evaluations).

Simulated high dimensional example

Twenty-dimensional test function [10, 12]

$$y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{ij} w_i w_j + \sum_{i<j<l}^{20} \beta_{ijl} w_i w_j w_l \\ + \sum_{i<j<l<s}^{20} \beta_{ijls} w_i w_j w_l w_s,$$

The w_i are functions of x_i . Factors x_1, \dots, x_7 have non-linear effect while x_8, x_9, x_{10} have linear and the rest negligible effect.

Set design with maximum of 20 points and repeated 100 times.

Identified x_1, \dots, x_7 as non-linear 99% of time; x_8, x_9, x_{20} as linear 95% of time with an average of 153 runs (210 total runs under EE method).

Discussion

- The proposed methodology aims to sequentially screen among factors with non-linear and linear effect. In the worst case, the total number of simulator runs is $(k + 1)M$ (same as full EE method-Morris).
- Good alternative to EE method, Sobol' and FAST.
- Tried in several simulators (Rabies, Rainfall-runoff) and simulated high dimensional case.
- It may be possible to sequentially run simplexes [12] in the current set of input factors, thus improving projections.
- Comparison with other alternatives. For instance, Sequential bifurcation (SB) assumes monotonicity of the model output and it might be very efficient, requiring only a very reduced number of runs [9].

7. References

1. Boukouvalas *et al.* (2009). *Technical report SPWS*.
2. Boukouvalas *et al.* (2011). An efficient screening method for computer experiments. Technical report NCRG (Aston), submitted to *JSPI*.
3. Campolongo *et al.* (2004). *SAMO*, 369-379.
4. Campolongo *et al.* (2007). *Envir. Model. Softw.* 1509-1518.
5. Garthwaite *et al.* (1988). *JRSSB*, 462-474.
6. Hargeaves *et al.* (2004). *Climate Dynamics*, 745-760.
7. Kadane *et al.* (1980). *JASA*, 845-854.
8. Kennedy *et al.* (2008). *JRSSA* 109-135.
9. Kleijnen (2009). Review of Sequential Bifurcation. *ISORMS* 153-167.
10. Morris (1991). *Techno.* 161-174.
11. O'Hagan (2006). *Rel. Eng. Sys. Saf.* 1290-1300.

12. Pujol (2009). *Rel. Eng. Sys. Saf.* 1156-1160.
13. Singer (2008). *Dev. Biol. Basel*, 213-222.
14. Wedge *et al.* (2009). *Jour. Theo. Biol.* 131-141.

Thank you!

Contact details:

H.Maruri-Aguilar@qmul.ac.uk