*11th International Conference on Sensitivity Analysis of Model Output, April 23–25 2025, Grenoble, France*

# Optimized clustering of model input samples based on sensitivity indices

Sébastien Roux
*MISTEA, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*

Patrice Loisel
*MISTEA, Univ Montpellier, INRAE, Institut Agro, Montpellier, France*

Samuel Buis
*EMMAH, INRAE, Avignon Université, Avignon, France*

The present work is motivated by the Sensitivity Analysis (SA) of models having multivariate (MV) inputs among their input factors. SA in this context is challenging because of dependency issues within the MV input components, which prevents to find and characterize easily the sensitive ones.

We investigate the use of clustering in order to provide more insights on the sensitive components of MV sensitive inputs. More precisely, we propose to use clustering to find groups of MV inputs samples such that group characteristics explains as best as possible the influence of the MV inputs. When successful, this strategy means that group characteristics are good summaries of the MV inputs influence on the model outputs.

In order to apply this strategy, two questions must be answered: i) how to define quantitatively the influence of groups on the output variability and ii) how to find clustering that maximize the associated criteria.

**Notations:**
We study $y = f(\boldsymbol{w}, z)$, where $\boldsymbol{w}$ is a complex input (typically a vector of weather variables in environmental models) and $z$ an independent input (possibly a large vector grouping all other inputs of interest). Using a labeling approach [2] based on samples $\boldsymbol{w}_1, ..., \boldsymbol{w}_L$, we now study $y = g(l, z) = f(\boldsymbol{w}_l, z)$. The Sobol' decomposition on $g$ writes simply: $S_l + S_z + S_{lz} = 1$.

We introduce a general clustering function $\mathcal{C}$ such that $\mathcal{C}(l) = c \in 1, .., K$ is the cluster label of the input with label $l$. We introduce also a 'within-cluster selection factor' $u \in [0, 1[$ that is used to choose elements within a cluster.
Let us note $(l_1^c, .., l_{N_c}^c)$ the $N_c$ elements in cluster $c$. We denote as $h$ the model having cluster labels and selection factors (along with the co-variable $z$) as inputs: $h(c, z, u) = g(l^c_{\lfloor u.N_c \rfloor + 1}, z)$, where $\lfloor x \rfloor$ is the integer part of $x$.

**Sensitivity analysis with selection factor $u$:**
Our central idea to define clustering criteria is to use the sensitivity indices associated to model $h$, where the cluster label $c$ has a discrete distribution with values $c_1, .., c_K$ and probabilities $p_1, .., p_K$ (probabilities of clusters according to their size), where $u$ has an uniform distribution within $[0, 1[$ and $z$ its (unchanged) uncertainty distribution. Writing the Sobol' decomposition on $h$, we have: $S_c + S_z + S_{cz} + S_u^T = 1$,where $S^T$ denotes a total Sobol' index.

**First clustering problem:** $\max_{\mathcal{C}(.)} S_c$
This optimization problem will allow to find clustering that maximize the main effect of the cluster type, which is at best equal to $S_l$. More precisely, we show that $S_c = S_l - \frac{1}{V}\sum_{c=1}^{K} p_c \widetilde{V}_c$, where $\widetilde{V}_c = \mathbb{V}_{l \in c}\mathbb{E}_z[g(l, z)]$. We show that solutions of this problem are defined using quantiles of the distributions $\mathbb{E}_z[g(l, z)]$, leading to an efficient numerical algorithm to find solutions of the global optimization problem. However a drawback of this criterion is that it does not take into account the variability of model responses along direction $z$.

**Second clustering problem:** $\min_{\mathcal{C}(.)} S_u^T$

Using this criterion, we try to minimize the effect of the within-cluster selection factor $u$, thus to minimize the effect (this time including interaction effects) of the within-cluster variability. We show that $S_u^T = \frac{1}{V}\mathbb{E}_z[\sum_{c=1}^{K} p_c V_c(z)]$, with $V_c(z) = \mathbb{V}_{l \in c}[g(l,z)]$. We show that numerical solutions of this problem can be found using a K-means like algorithm. Compared to a classical K-means problem, our algorithm uses distances in the space of outputs, i.e not in the space of the variable to be clustered.

**Numerical examples**

We implemented the algorithms for solving the two previous problems and tested them firstly on a simple function at the level of $g_l$ (i.e. on functions having label $l$ and co-variable $z$ as inputs). The model output $y$ has no variability along $z$ for $l \leq 75$, where $y = 1 + 0.005\ l$. For $l > 75$, $\mathbb{E}_z[y]$ is also equal to $1 + 0.005\ l$, but $y$ can take two values depending on $z$, which are inverted if $l > 87$. We can see in Figure 1 that the $S_u^T$-based criterion takes into account the variability along $z$ and creates clusters in the region of variability in $z$, which was not the case of the $S^c$-based criterion.

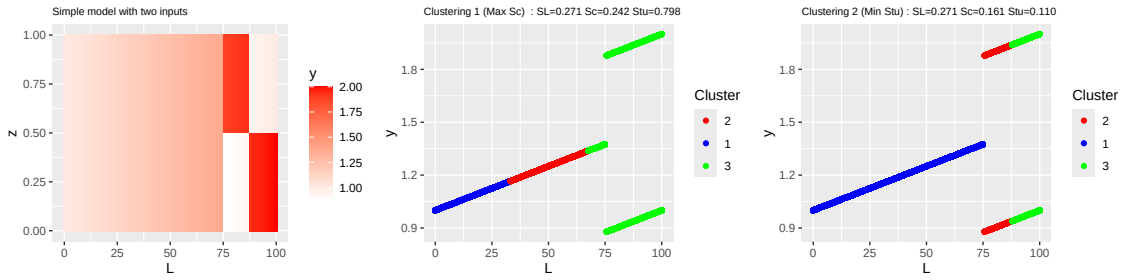


Figure 1: Clustering result of a simple model for the two criteria. Left: model definition; Middle: clustering based on $S_c$; Right: clustering based on $S_u^T$.

We will also present during the conference clustering on a crop model [1] having vector of weather variables among its inputs. We will be particularly interested in looking at the influence of the number of clusters and in showing how the produced clusters can help to better understand the influence of weather inputs.

**References:**

[1] Nadine Brisson et al. "An overview of the crop model STICS". In: *European Journal of agronomy* 18.3-4 (2003), pp. 309–332.

[2] Linda Lilburne and Stefano Tarantola. "Sensitivity analysis of spatial models". In: *International Journal of Geographical Information Science* 23.2 (2009), pp. 151–168.

[ Presenting author's name; affiliation and regular mailing address ]
[ `Sébastien Roux, INRAE,sebastien.roux@inrae.fr` – ]