ESAIM: PS 30 (2026) 342–353
https://doi.org/10.1051/ps/2026008

ESAIM: Probability and Statistics
www.esaim-ps.org

CLUSTERING MODEL INPUT SAMPLES: LINKING K-MEANS TO SOBOL' INDICES

Sébastien Roux[1,*], Patrice Loisel[1] and Samuel Buis[2]

Abstract. This work is devoted to the problem of clustering a set of samples according to the effect they have as one of the many varying inputs of a model. An example is the problem of clustering weather series according to the effect they have on the yield simulated with a crop model when also other inputs such as soil or plant parameters vary. We introduce both simple solutions based on the K-Means algorithms and the derivation of two possible formulations of the clustering problem in a sensitivity analysis framework. We show that these formulations coincide in the sense that their criteria are closely related to each other, leading to a possibility to use the K-Means algorithm with an improved expression of clustering performance through the use of Sobol' indices.

Mathematics Subject Classification. 65C60, 65C20.

Received September 30, 2025. Accepted May 28, 2026.

1. Introduction

Clustering methods are powerful tools when dealing with large set of data with a complex structure (typically multivariate), either to reveal the few main interpretable features underlying the large initial set, or to provide a non-interpretable but statistically relevant summary of the initial data. In the present work, we consider the case where the data to be clustered are one of the inputs of a simulation model. In this context, we address the problem of clustering these data i) according to the effect they have on simulated outputs, ii) when some other inputs vary in the same time. This issue can typically arise when seeking to summarize the impact of weather data on an agro-environmental model in a setting that involves the variation of several other model inputs. Such objective implies that the clustering has to link the input space (*i.e.* the space of models inputs) to the output space (*i.e.* the space of model outputs), meaning that, for the agro-environmental model example, the objective and method will be different from clustering the weather inputs based on their similarities, but also different from clustering the set of simulated yields. One goal of the present work is to link clustering approaches that link model inputs and model outputs with the framework of sensitivity analysis [1], and more precisely with the variance based sensitivity analysis [2], which provides a framework for quantifying the effect of model inputs variability on model outputs.

In the context of sensitivity analysis, clustering can be applied on model inputs in order to investigate the effect of cluster type on the outputs variability. In this case, the clustering is a pre-processing step applied

Keywords and phrases: Clustering, sensitivity indices, regional sensitivity analysis.

[1] MISTEA, Univ Montpellier, INRAE, Institut Agro, Montpellier, France.
[2] EMMAH, INRAE, Avignon Université, Avignon, France.

* Corresponding author: sebastien.roux@inrae.fr

on an input sample with a clustering criterion independent of the model outputs. Clustering approaches have been applied in relation to sensitivity analysis for analysing results obtained on repeated analysis [3–5]. In these examples, the clustering is performed as a post-processing step independent of the sensitivity analysis itself. In [6], the clustering is applied on model outputs and sensitivity indices are computed on the cluster membership functions in order to characterize the influence of parameters variability on the occurrence of clusters, in the same spirit as Regional Sensitivity Analysis [7] or Target Sensitivity Analysis [8].

The present work is inspired by [9], which investigates one step further the link between clustering and sensitivity in the context of model outputs by posing an optimization problem aiming at finding a clustering that maximizes a sensitivity-based criterion on the model outputs. We propose here to investigate a similar idea but on the model inputs, which to our knowledge has not been done before. We precisely raise the question of finding clusters of model inputs samples in relation to sensitivity indices. As this problem deals with model inputs, the cluster-based sensitivity indices proposed in [6] and defined using clusters of model outputs, are not suitable. Different sensitivity indices in relation to clusters of model inputs have to be proposed, and other optimization procedures than the one proposed in [9] are also required.

The article is organized as follows: first the clustering problem is presented in Section 2 along with several intuitive solutions in Section 3, then the sensitivity based framework is introduced in Section 4. We then establish the link between the intuitive and sensitivity-based formulations in Section 5 and finally present some illustrations in Section 6.

2. Problem statement: clustering model inputs samples

We consider a simulation model $\mathcal{M}$ having scalar outputs. The model is written with a focus on an input named W and has p other inputs $Z_1, .., Z_p$ that are supposed to vary independently of W.

We suppose that input W may have a complex structure (for instance a vector or a graph of variables), so that its uncertainty is not modeled with a continuous or model-based random input but instead using a discrete random variable uniformly distributed over a finite empirical set of N_L samples $\mathcal{W} = \{w^{(1)}, .., w^{(N_L)}\}$. For the sake of simplicity, we suppose that all Z_i have an uniform distribution in $[0, 1]$:

- $Y = \mathcal{M}(W, Z_1, .., Z_p)$
- $W \sim \mathcal{U}(\{w^{(1)}, .., w^{(N_L)}\})$
- $Z_i \sim \mathcal{U}([0, 1])$

In this context, finding a clustering of $\mathcal{W}$ can be written as the problem of finding K clusters $(c_1, .., c_K)$. Each cluster of index c_i and cardinal N_{c_i} is made of samples indices noted $\{l_1^{c_i}, .., l_{N_{c_i}}^{c_i}\}$.

3. Simple solution: using K-Means algorithm

A first simple solution of the clustering problem is to consider the application of a K-Means algorithm [10] but this requires to define completely the sample under study.

3.1. Data generation

For the numerical application of a clustering procedure on the set $\mathcal{W}$, we need to sample the outputs according to the distributions given in the previous section. We suppose that we are in favorable setting for this purpose and that we have the simulated outputs on a full factorial design built from a sample $\{w^{(1)}, .., w^{(N_L)}\}$ and a sample $\{z^{(1)}, .., z^{(N_z)}\}$, where $\{z^{(k)}\}$ is a sample of the random vector $(Z_1, .., Z_p)$. The data available for clustering thus consists in a matrix $\boldsymbol{Y} = (y_{ij})_{i,j}$, with $y_{ij} = f(w^{(i)}, z_1^{(j)}, .., z_p^{(j)})$. A clustering is characterized by a function $\mathcal{C}$ of a label l: the fact that label l belongs to the cluster of index c is express by $\mathcal{C}(l) = c$.

In the following we abusively use $l \in c$ for $l \in \mathcal{C}^{-1}(c)$ *i.e.* $\mathcal{C}(l) = c$.

3.2. Two K-Means algorithms

In order to obtain a clustering of the set $\{w^{(1)}, ..., w^{(N_L)}\}$ using matrix $\boldsymbol{Y}$, two options can be straightforwardly applied:

- $(K1)$: clustering the set $\mathcal{W}$ after averaging over z, *i.e.* clustering the mean of $\boldsymbol{Y}$ columns. This amounts to cluster the set $\{\bar{y}_1, .., \bar{y}_{N_L}\}$, with $\bar{y}_i = \dfrac{1}{N_z}\sum_j y_{ij}$,
- $(K2)$: clustering the set $\mathcal{W}$ based directly on all values in matrix $\boldsymbol{Y}$, *i.e.* clustering the columns of $\boldsymbol{Y}$. This amounts to cluster the set $\{y_{1.}, .., y_{N_L.}\}$, with $y_{i.} = (y_{i1}, .., y_{iN_z})$.

In order to perform these two clusterings, an intuitive approach is to use a standard K-Means algorithm [10], which we recall, seeks to minimize the sum of intra-cluster sums of squares, denoted respectively K_1^{WSS} and K_2^{WSS} for the two problems and which write as follows:

- $(K1) : K_1^{WSS} = \min_{\mathcal{C}(.)} \sum_{l=1}^{N_L} (\bar{y}_l - \bar{\bar{y}}_{\mathcal{C}(l)})^2$
- $(K2) : K_2^{WSS} = \min_{\mathcal{C}(.)} \sum_{l=1}^{N_L}\sum_{j=1}^{N_z} (y_{lj} - \bar{y}_{\mathcal{C}(l)j})^2$

where: $\bar{\bar{y}}_c = E_{l\in c}\ \bar{y}_l$ and $\bar{y}_{cj} = E_{l\in c}\ y_{lj}$.

We recall that these two clustering problems, $(K1)$ which is in dimension 1 and $(K2)$ which is in dimension N_z, can be solved efficiently with the K-Means algorithm, which is an alternate optimization procedure that is ensured to converge to a local minimum of the within sum of square cost function (see for instance [11]).

4. Sensitivity-based formulations

As the core of this work, we propose another point of view of the presented clustering problem with a sensitivity analysis angle. We will express the clustering problem as the minimization of sensitivity indices corresponding to a particular model.

4.1. Models with cluster-related factors

We first introduce model g whose inputs are an index $l \in \{1, ..., N_L\}$ for selecting an object in set $\mathcal{W}$, and a vector $z = (z_1, ..., z_p)$ of the co-variables:

$$g(l, z) = f(w^{(l)}, z)$$

We also introduce model h that is associated to a clustering into K clusters. The inputs of h are: a cluster index $c \in \{c_1, .., c_K\}$, a within-cluster selection factor that will be denoted as $u \in [0, 1[$ and the vector $z = (z_1, ..., z_p)$ of co-variables. Recalling that the elements of a cluster of index c_i are $\{l_1^{c_i}, .., l_{N_{c_i}}^{c_i}\}$, model h is defined from these inputs factors as follows:

$$h(c, z, u) = g(l^c_{\lfloor u.N_c \rfloor + 1}, z), \text{ where } \lfloor x \rfloor \text{ is the integer part of } x$$

Keeping in mind that samples of W and Z are supposed given, the new inputs are associated to random variables L, Z, C and U having the following distributions:

- $L \sim \mathcal{U}(\{l_1, .., l_{N_L}\})$
- $C \sim \mathcal{D}(\{(c_1, \frac{N_{c_1}}{N_L}), .., (c_K, \frac{N_{c_K}}{N_L})\})$
- $U \sim \mathcal{U}([0, 1[)$

- $Z \sim \mathcal{U}(\{(z_1^1, .., z_p^1), .., (z_1^{N_z}, ..., z_p^{N_z})\})$

where $\mathcal{U}$ (resp. $\mathcal{D}$) denote the continuous (resp. discrete) uniform distribution.

4.2. Variance decomposition and sensitivity indices

Now that model g and h have been defined with appropriate input factors and associated uncertainty distributions, we can write the decomposition of variance of these models.

We first note that $\mathbb{V}[g(L, Z)] = \mathbb{V}[h(C, Z, U)] = \mathbb{V}[Y] = V$.

The decomposition of variance associated to model g can be written as:

$$S_l + S_z + S_{lz} = 1$$

where:

- $S_l = \frac{1}{V}\mathbb{V}[\mathbb{E}[g(L, Z)|L]]$ (first order Sobol' index for input L)
- $S_z = \frac{1}{V}\mathbb{V}[\mathbb{E}[g(L, Z)|Z]]$ (first order Sobol' index for input Z)
- $S_{lz} = 1 - S_l - S_z$ (interaction effect between inputs C and Z)

The decomposition of variance associated to model h can be written as:

$$S_c + S_z + S_{cz} + S_u^T = 1$$

where:

- $S_c = \frac{1}{V}\mathbb{V}[\mathbb{E}[h(C, Z, U)|C]]$ (first order Sobol' index for input C)
- $S_u^T = \frac{1}{V}\mathbb{E}[\mathbb{V}[h(C, Z, U)|C, Z]]$ (total order Sobol' index for input U)
- $S_{cz} = 1 - S_c - S_z - S_u^T$ (interaction effect between inputs C and Z)

Note that as g and h have the same variance, the term S_z of the two decompositions is the same.

4.3. Two sensitivity-related clustering problems

The interesting feature of model h is that its factors c and u decompose the intra-clusters and inter-clusters variabilities. Using these factors and their sensitivity indices, it is then possible to define two clustering problems that can easily be interpreted.

4.3.1. Clustering based on S_c

First we consider the problem $(P1)$ of finding a clustering that maximizes S_c, *i.e.* the Sobol' first order index of the clustering label c.

$$(P1) : \max_{\mathcal{C}(.)} S_c$$

The sensitivity index S_c can also be written as follows:

$$S_c = S_l - \frac{1}{V}\sum_{k=1}^{K}\frac{N_{c_k}}{N_L}\mathbb{V}_{l\in c_k}\ \bar{g}(l) \tag{4.1}$$

with $\bar{g}(l) = \mathbb{E}_z\ g(l,z)$. Indeed, we have:

$$V.S_l = \mathbb{V}_l[\mathbb{E}_z\ g(l,z)] = \mathbb{V}_l\ \bar{g}(l)$$
$$V.S_c = \mathbb{V}_c[\mathbb{E}_{u,z}\ h(c,z,u)] = \mathbb{V}_c[\mathbb{E}_{l\in c}\ \mathbb{E}_z\ g(l,z)] = \mathbb{V}_c[\mathbb{E}_{l\in c}\ \bar{g}(l)]$$

Therefore:

$$\begin{aligned} V(S_l - S_c) &= \mathbb{E}_l[\bar{g}(l) - \mathbb{E}_l\ \bar{g}(l)]^2 - \mathbb{E}_c[\mathbb{E}_{l\in c}\ \bar{g}(l) - \mathbb{E}_c\ \mathbb{E}_{l\in c}\ \bar{g}(l)]^2 \\ &= \mathbb{E}_c[\mathbb{E}_{l\in c}[\bar{g}(l) - \mathbb{E}_l\ \bar{g}(l)]^2] - \mathbb{E}_c[\mathbb{E}_{l\in c}\ \bar{g}(l) - \mathbb{E}_l\ \bar{g}(l)]^2 \\ &= \mathbb{E}_c[\mathbb{V}_{l\in c}\ \bar{g}(l)] = \sum_{k=1}^{K}\frac{N_{c_k}}{N_L}\mathbb{V}_{l\in c_k}\ \bar{g}(l) \end{aligned}$$

From this expression, we deduce several properties of the optimal clustering S_c^* according to criterion S_c:

- $S_c^* \leq S_l$
- $S_c^* = S_l$ if (and only if) $\forall k\ \forall (l,l' \in c_k),\ \bar{g}(l) = \bar{g}(l')$

4.3.2. Clustering based on S_u^T

Secondly we consider the problem $(P2)$ of finding a clustering that minimizes S_u^T, *i.e.* the total Sobol' index of the within-cluster selection factor u.

$$(P2) : \min_{\mathcal{C}(.)} S_u^T$$

The sensitivity index S_u^T can also be written as follows:

$$V.S_u^T = \mathbb{E}[\mathbb{V}[Y|X_{-u}]] = \mathbb{E}_{c,z}[\mathbb{V}_u\ h(c,z,u)] = \sum_k \frac{N_{c_k}}{N_L}\mathbb{E}_z[\mathbb{V}_{l\in c_k}\ g(l,z)] \tag{4.2}$$

From this expression, we can deduce a property of the clustering minimizing criterion S_u^T: $S_u^T = 0$ if (and only if) $\forall k, \forall z,\ \mathbb{V}_{l\in c_k}\ g(l,z) = 0$ (*i.e.* for any z, no variability along l inside a cluster).

5. Linking formulations

5.1. Notations

We note $y_{lj} = g(l, z^{(j)}),\ \bar{y}_{cj} = \mathbb{E}_{l\in c}\ y_{lj},\ \bar{y}_l = \mathbb{E}_z\ g(l,z)\ (= \bar{g}(l)),\ \bar{\bar{y}}_c = \mathbb{E}_{l\in c}\ \bar{y}_l$.

5.2. Link between $(K1)$ and $(P1)$

From Eq. 4.1, we see that the clustering maximizing S_c minimizes $\displaystyle\sum_{k=1}^{K}\frac{N_{c_k}}{N_L}\mathbb{V}_{l\in c_k}\ \bar{g}(l)$.

We note that: $\sum_{k=1}^{K} N_{c_k} \mathbb{V}_{l \in c_k}\ \bar{g}(l) = \sum_{k=1}^{K} \sum_{l \in c_k} (\bar{y}_l - \bar{\bar{y}}_{c_k})^2 = \sum_{l} (\bar{y}_l - \bar{\bar{y}}_{\mathcal{C}(l)})^2$.

Therefore, providing that the optimum value of the within-cluster sum of square has been globally minimized by the K-Means algorithm $(K1)$, we have the following link between K_1^{WSS} and the maximum value S_c^* of S_c over all clusterings:

$$S_c^* = S_l - \frac{1}{V.N_L} K_1^{WSS} \tag{5.1}$$

5.3. Link between $(K2)$ and $(P2)$

From the definition of S_u^T, we have:

$$\begin{aligned} V.S_u^T &= \sum_k \frac{N_{c_k}}{N_L} \mathbb{E}_z[\mathbb{V}_{l \in c_k}\ g(l,z)] \\ &= \frac{1}{N_L.N_Z} \sum_k \sum_j \sum_{l \in c_k} (g(l, z^{(j)}) - \mathbb{E}_{l \in c_k}\ g(l, z^{(j)})))^2 \\ &= \frac{1}{N_L.N_Z} \sum_k \sum_j \sum_{l \in c_k} (y_{lj} - \bar{y}_{c_k j})^2 = \frac{1}{N_L.N_z} \sum_{l=1}^{N_L} \sum_{j=1}^{N_z} (y_{lj} - \bar{y}_{\mathcal{C}(l)j})^2 \end{aligned}$$

We therefore deduce that, providing that the optimum value of the within-cluster sum of square has been globally minimized by the K-Means algorithm $(K2)$, we have the following link between K_2^{WSS} and the maximum value S_u^{T*} of S_u^T over all clusterings:

$$S_u^{T*} = \frac{1}{V.N_L.N_z} K_2^{WSS} \tag{5.2}$$

5.4. Consequences on clustering algorithms

The sensitivity-based formulations of the clustering problem are satisfying in the sense that they lead to clustering criteria based on sensitivity indices that are normalized and directly interpretable. From the results of the previous section, we conclude that $(P1)$ can be numerically solved using the K-Means algorithm $(K1)$, and that $(P2)$ can be numerically solved using the K-Means algorithm $(K2)$. The optimized values of sensitivity indices S_c^* and S_u^{T*} are deduced respectively from K_1^{WSS} or K_2^{WSS} in Eqs. 5.1 and 5.2.

Note that K-Means being an algorithm that only ensures convergence to a local minimum, it is not guaranteed that the global optima of problems $(P1)$ and $(P2)$ can be found using this approach. However, for problem $(P1)$, which becomes a one-dimensional problem after averaging along direction z, it is possible to derive a global algorithm which can be used either directly or as a starting point for $(K2)$. More precisely, a global algorithm can be written based on the property that solutions of $(P1)$ can be defined using quantiles of the distribution $\mathbb{E}_z[g(l,z)]$ (see Appendix A). As a consequence, an efficient numerical algorithm based on a search of quantiles can be used to find solutions. Note however that this is not possible for problem $(P2)$ for which the K-Means approach is the more efficient.

6. Examples

We present in this section two examples. The first one is analytical and illustrates the differences between criteria based on S_c or S_u^T. The second, based on a simplified crop model, is presented to illustrate how the clustering applies to a problem involving inputs with a complex structure and several co-variables.

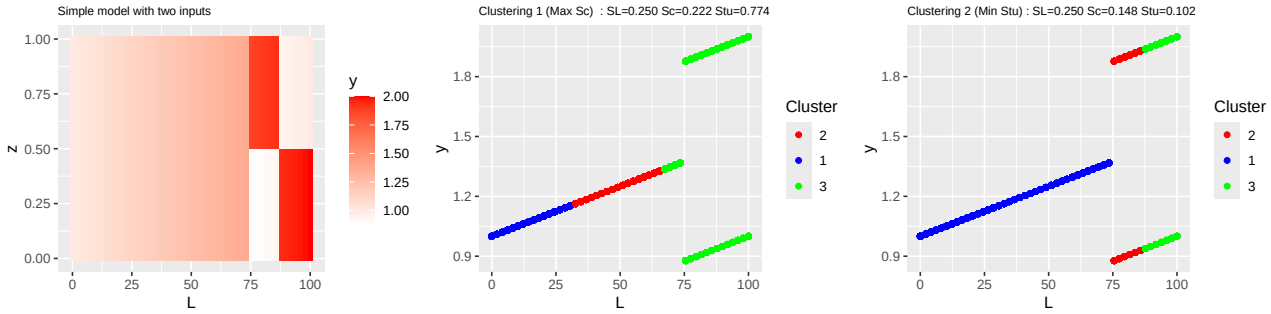

FIGURE 1. Clustering results on a analytical model for the two criteria. *Left*: heatmap of model outputs. *Middle*: clustering based on S_c. *Right*: clustering based on S_u^T.

6.1. Analytical toy model

6.1.1. Model definition

We compared the two sensitivity-based criteria on a simple function at the level functions having label l and 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 $l > 87$. More precisely, the definition of the model and of the input distributions are the following:

- $g(l, z) = 1 + 0.005\ l + [\mathbb{1}_{(87-l).(z-0.5)\geq 0} - 0.5]\mathbb{1}_{l>75}$
- $L \sim \mathcal{U}(\{1, .., 100\})$
- $Z \sim \mathcal{U}([0, 1])$

Note that this definition implies that $\mathbb{E}_z[g(l, z)] = 1 + 0.005\ l$.

6.1.2. Comparison of the two sensitivity-based criteria

As can be seen in Figure 1, and as expected, the clustering associated to the maximization of S_c leads to a higher value (0.222, to be compared to $S_l = 0.250$) than the value of S_c (0.148) obtained when minimizing S_u^T. Conversely, the clustering optimized according to criterion $\max S_c$ can be associated to high S_u^T (0.774 is this example), where the minimization of this criterion gives a value of 0.102. As shown in Section 6.1.1, the clustering optimized according to criterion $\max S_c$ relies only on $\mathbb{E}_z[g(l, z)] = 1 + 0.005\ l$, which is a linear expression in l. That is why the clustering found is just a partition of the interval into continuous and equi-sized clusters, as shown in Figure 1. On the other hand, the clustering optimized according to criterion $\max S_u^T$ takes into account the variability in l of $g(l, z)$ within a cluster, as can be seen in Eq. 4.2, which depends on $\mathbb{E}_z[\mathbb{V}_{l\in c_k} g(l, z)]$. This variability prevents the optimal clustering to gather together the regions $75 < l \leq 87$ and $l > 87$. This can be seen in Figure 1 : the clustering optimized according to criterion $\max S_u^T$ created two clusters for $l > 75$. This example shows that, as opposed to criterion $\max S_c$, criterion $\max S_u^T$ takes into account interacting effects between l and z to find clusters.

6.2. Simplified crop model

6.2.1. Model definition

In the second example, we illustrate the ability of the method to solve the clustering problem with a model having one input with a complex structure and several co-variables. To this aim, we use a parsimonious crop model (close to the one used in [12]) that simulates the yield of a crop influenced by weather variables, soil properties and management practices. Note that this model is simplified in term of crop processes but shares the generic structure of inputs of classical crop models: $Y = \mathcal{M}^{crop}(W,\ Z_1,\ Z_2,\ Z_3)$, with:

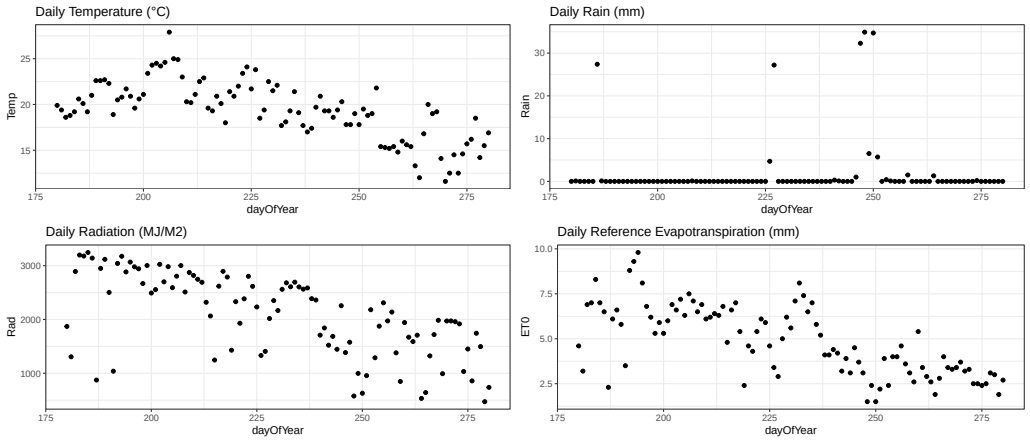

FIGURE 2. Example of sample $w^{(l)}$ made of the discretization of four weather variables (rain, temperature, radiation and reference evapotranspiration) during a crop growing season.

- $W \sim \mathcal{U}(\{w^{(1)}, .., w^{(N_L)}\})$
- $Z_i \sim \mathcal{U}([0,1])$ for $i = 1,\ 2,\ 3$

In this formulation, each $w^{(i)}$ is made of the discretization of four weather variables x_{rain} (rain), x_{ET_0} (reference evapotranspiration), x_{temp} (air temperature), x_{rad} (radiation) over the i^{th} cropping season. The sample $\{w^{(1)}, .., w^{(N_L)}\}$ contains $N_L = 42$ years of observed weather variables (a sample for one cropping season is presented in Fig. 2). The co-variables $(Z_1,\ Z_2,\ Z_3)$ are linked to three crop model inputs: total available soil water, temperature stress threshold, use of irrigation (see Appendix B for details on the system of differential equations that defines model $\mathcal{M}^{crop}$). The K-Means algorithm was applied with multiple starts in order to try to find a global optimum. On this example, using 40 multiple starts led to convergence toward the same solution.

6.2.2. Comparison of the two sensitivity-based criteria

Examples of clustering using criteria S_c and S_u^T are presented in Figure 3 with $K = 5$ clusters. In this example, we have $S_l = 0.253$. The optimisation according to $(P1)$ gives $S_c = 0.238$, $S_u^T = 0.238$ while the optimisation according to $(P2)$ gives $S_c = 0.226$, $S_u^T = 0.086$. We can see in Figure 3 that the clusterings obtained using the two criteria are again not the same, which is expected as the criteria are different, with criterion S_u^T better taking into account interactions. But compared to the previous example, the results are closer (31 out of 42 samples are clustered in the same way using the two approaches). We can also see that the samples not clustered in the same way are close to the S_c-based clusters boundaries, except for $l = 25$ (which becomes grouped with other samples with low variances) and $l = 34$ (which becomes grouped with other samples with bi-modality). For these two samples, the differences are linked to interactions between l and z, which introduce more contrast between the samples than the expectation along z alone.

DISCUSSION

The goal of the illustrating examples was to show the differences between the criteria and that the clustering framework can be applied to the input structure of crop models. That is why we used the same sampling design for the two examples. But it shall be noticed that the number of samples N_Z chosen to generate the factorial design can have an impact on the results, as well as the random sampling for a given N_Z. To address this issue for the crop model example, we computed the number of different clusterings obtained from 50 repetitions of a sampling design of size N_Z, for $N_Z = 10, 20, 50, 100, 500$. The results are shown in Figure 4. We can see

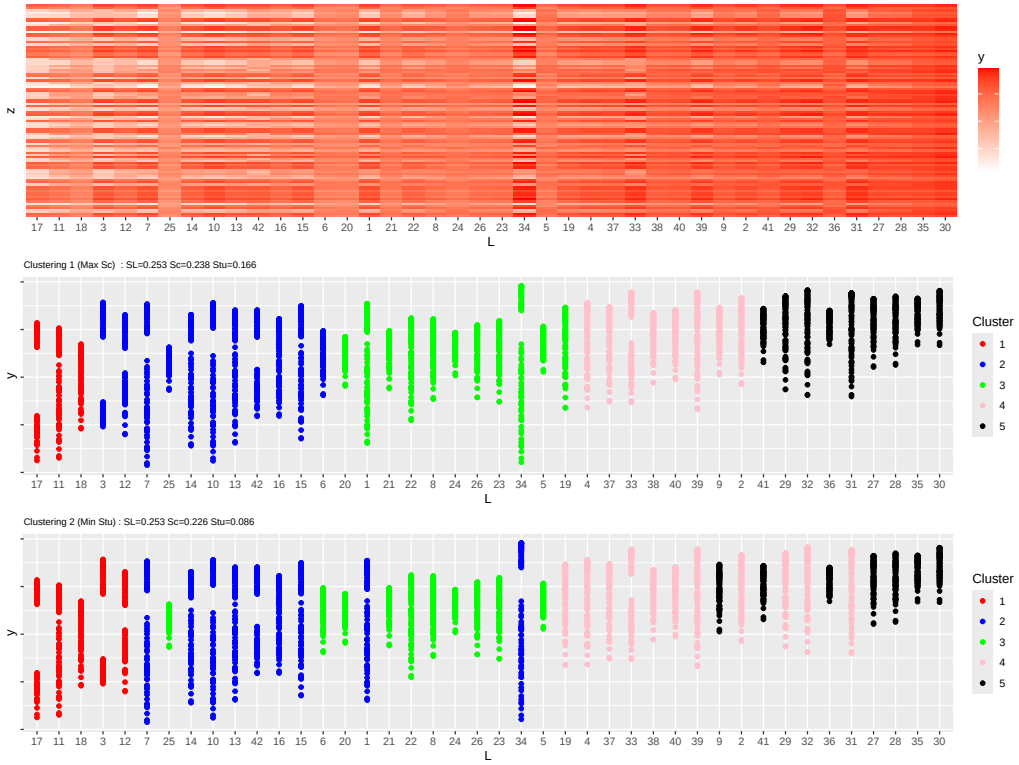

FIGURE 3. Clustering results on a toy model sharing the complex input structure of crop models model for the two criteria. *Top*: heatmap of model outputs. *Middle*: clustering based on S_c. *Bottom*: clustering based on S_u^T. Note that the samples have been sorted according to their mean value in direction z.

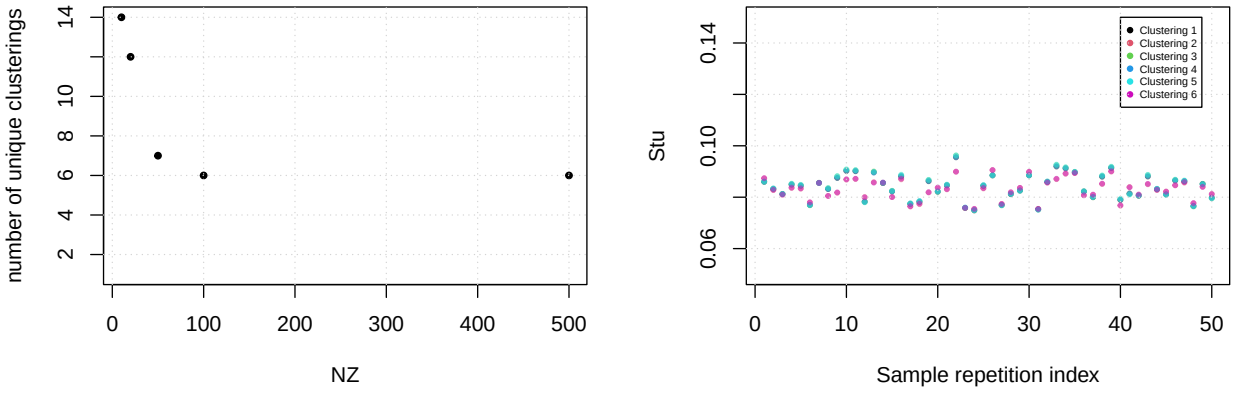

FIGURE 4. Analysis of the impact of N_Z on the clustering results. *Left*: number of unique clusterings found as a function of N_Z, when repeating 50 times the sampling for each N_Z. *Right*: for $N_Z = 100$, comparison of the criteria S_u^T obtained for each 6 clusterings on all sampling designs.

that N_Z has an impact on the results, and that there is a stabilization of the number of different clusterings for sufficiently large N_Z ($N_Z \geq 100$), but that 6 different clusterings are still found among the repetitions. These clusterings appear quantitatively equally good, in the sense that each one is the best clustering for some sampling designs. While we expect that the number of clusterings should converge to a single 'best' clustering for very large N_Z, for reasonable computing time and intermediate value of N_Z, there is an equifinality issue: several clusterings perform equally well, and all of them can be considered as candidates for further processing of the clusters obtained with this approach.

Since the optimization problem can be solved using the K-Means algorithm, the method is scalable for the clustering step and can support much larger problems than those presented in the examples. But before applying the clustering algorithm, the matrix obtained from the factorial design must be built with a sufficiently large N_Z to ensure stability of the results. This requires many more simulations for large data set (*i.e.* for larger libraries of functional inputs or higher-dimensional co-variables) and may be impractical for models with long simulation times. For such cases, building a metamodel can provide a feasible way to apply the clustering method.

The setting addressed in this paper does not allow every clustering approach designed for complex inputs to be used. The reason is that the clustering depends more on the output space of the model (scalars in our settings), than on the input space (vectors of time series in the second example) and actually depends on both. We showed that the input data can be reformulated using the matrix obtained by simulating the model over the factorial design (LxZ). Other clustering methods (such as hierarchical clustering, or spectral clustering) could be applied to cluster the columns of this matrix. However it is the K-Means criterion that is directly related to the optimization of variance-based sensitivity indices.

Clustering model input samples can bring insights on the sensitive characteristics of model inputs. However this is only possible if two conditions are fulfilled. First, the clustering must perform well, *i.e.* the effect of intra-cluster variability on the outputs, quantified here using a total Sobol' index, must be low. In this case, model understanding can be explored by examining interpretable characteristics that explain the different clusters. For instance, for crop models, if the clusters of weather data can be characterized by the presence or absence of drought event during spring, we can conclude that this characteristic is the main driver of yield under the given simulation setting. If no clear characterization is found (which is possible), the clusters can still be of interest as a statistical summary of the model input sample. This can be useful for visualization purpose, as in [13], where a modified Lloyd'd algorithm is applied on model outputs to find representative samples of a rare event in the context of flooding visualization, or to get a building block for constructing efficient metamodels, as in [14] where different Gaussian processes models are built in relation to clusters in the input space.

Conclusion

In this work, we addressed the problem of clustering samples according to their effects on a model when co-variables are also involved. We showed that intuitive approaches coincide with a formalisation based on Sobol' sensitivity indices. These links have two advantages: justifying the use of efficient clustering algorithms in this context and providing interpretable performance metrics (using sensitivity indices) when doing the clustering. The current work was developed under the hypothesis of a favorable setting for sampling the model. Future work will address more general sampling constraints.

Data availability statement

The data used in the current study are available from the corresponding author on request.

References

[1] S. Da Veiga, F. Gamboa, B. Iooss and C. Prieur, *Basics and Trends in Sensitivity Analysis: Theory and Practice in R.* SIAM (2021).

[2] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models. *Math. Model. Computat. Exp.* **1** (1993) 407–414.

[3] P. Casadebaig, B. Zheng, S. Chapman, N. Huth, R. Faivre and K. Chenu, Assessment of the Potential Impacts of Wheat Plant Traits across Environments by Combining Crop Modeling and Global Sensitivity Analysis. *PLoS One* **11** (2016) e0146385.
[4] J.F. Saval, P. Barbillon, C. Benhamou, P. Durand, M.-L. Taupin, H. Monod and J.-L. Drouet, Spatial and dynamic sensitivity analysis of a biophysical model of nitrogen transfers and transformations at the landscape scale. In : Proceedings of the 8th International Congress on Environmental Modelling and Software (iEMSs2016). iEMSs, (2016).
[5] L. Wang, Y.-P. Xu, J. Xu, H. Gu, Z. Bai, P. Zhou, H. Yu and Y. Guo, Increasing parameter identifiability through clustered time-varying sensitivity analysis. *Environ. Model. Softw.* **181** (2024) 106189.
[6] S. Roux, S. Buis, F. Lafolie and M. Lamboni, Cluster-based GSA: globaL sensitivity analysis of models with temporal or spatial outputs using clustering. *Environ. Model. Softw.* **140** (2021) 105046.
[7] R.C. Spear, T.M. Grieb and N. Shang, Parameter uncertainty and interaction in complex environmental models. *Water Resources Res.* **30** (1994) 3159–3169.
[8] A. Marrel and V. Chabridon, Statistical developments for target and conditional sensitivity analysis: Application on safety studies for nuclear reactor. *Reliabil. Eng. Syst. Saf.* **214** (2021) 107711.
[9] S. Roux, P. Loisel and S. Buis, Maximizing regional sensitivity analysis indices to find sensitive model behaviors. *Int. J. Uncertain. Quantif.* **15** (2025) 47–60.
[10] J.B. McQueen, Some methods of classification and analysis of multivariate observations, in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967) 281–297.
[11] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. SIAM (2020).
[12] S. Roux, P. Loisel and S. Buis, A filter-based approach for global sensitivity analysis of models with functional inputs. *Reliabil. Eng. Syst. Saf.* **187** (2019) 119–128.
[13] C. Sire, R.L. Riche, D. Rullière, J. Rohmer, L. Pheulpin and Y. Richet, Quantizing rare random maps: Application to flooding visualization. *J. Computat. Graph. Statist.* **32** (2023) 1556–1571.
[14] C.-L. Sung, B. Haaland, Y. Hwang and S. Lu, A clustered Gaussian process model for computer experiments. *Statist. Sinica* **33** (2023) 893–918.

Appendix A. Global Optimization for problem ($P1$)

We show that the solutions of ($P1$) are defined using quantiles of $\mathbb{E}_z[g(l,z)]$, leading to an efficient numerical algorithm to find solutions to the global optimization problem.

To this end, we will show that the optimal Sobol' index is driven by ordered values of $\overline{y}$. In other words, the optimal clustering satisfies: $\overline{\overline{y}}_I < \overline{\overline{y}}_J$ implies that for all $i,\ j$ such that $\mathcal{C}(i) = I$ and $\mathcal{C}(j) = J$ then $\overline{y}_i < \overline{y}_j$.

Assume the contrary, the optimal Sobol' index not driven by ordered values of $\overline{y}$, hence for the optimal clustering, it exists $\overline{y}_i > \overline{y}_j$ and $\overline{\overline{y}}_I < \overline{\overline{y}}_J$ with $\mathcal{C}(i) = I$ and $\mathcal{C}(j) = J$.

We will show that such clustering is not optimal because the swapping of $\overline{y}_i$ and $\overline{y}_j$ increases the Sobol' index. Let $\mathcal{C}'$ the associated labelling function, with $\mathcal{C}'(j) = I,\ \mathcal{C}'(i) = J$ and $\mathcal{C}'(k) = \mathcal{C}(k)$ for all $k \neq i,\ j$. Cluster sizes are preserved, hence the expectations $\overline{\overline{y}}_K$ for $K \neq I, J$ are unchanged. Moreover $\overline{\overline{y}}'_I - \overline{\overline{y}}_I = -\dfrac{\overline{y}_i - \overline{y}_j}{N_I},\ \overline{\overline{y}}'_J - \overline{\overline{y}}_J = \dfrac{\overline{y}_i - \overline{y}_j}{N_J}$ and:

$$V N_L(S'_c - S_c) = N_I \left(\mathbb{V}_{\mathcal{C}(l)=I}\ \overline{y}_l - \mathbb{V}_{\mathcal{C}'(l)=I}\ \overline{y}_l\right) + N_J \left(\mathbb{V}_{\mathcal{C}(l)=J}\ \overline{y}_l - \mathbb{V}_{\mathcal{C}'(l)=J}\ \overline{y}_l\right)$$

where:

$$\begin{aligned} N_I \ \left(\mathbb{V}_{\mathcal{C}(l)=I}\overline{y}_l - \mathbb{V}_{\mathcal{C}'(l)=I}\ \overline{y}_l\right) &= \overline{y}_i^2 - N_I\overline{\overline{y}}_I{}^2 - (\overline{y}_j^2 - N_I\overline{\overline{y}}_I'^2) \\ &= \overline{y}_i^2 - \overline{y}_j^2 + N_I\left(\overline{\overline{y}}_I + \frac{\overline{y}_j - \overline{y}_i}{N_I}\right)^2 - N_I\overline{\overline{y}}_I{}^2 \\ &= \overline{y}_i^2 - \overline{y}_j^2 + 2(\overline{y}_j - \overline{y}_i)\overline{\overline{y}}_I + \frac{(\overline{y}_j - \overline{y}_i)^2}{N_I} \\ N_J \ \left(\mathbb{V}_{\mathcal{C}(l)=J}\overline{y}_l - \mathbb{V}_{\mathcal{C}'(l)=J}\ \overline{y}_l\right) &= \overline{y}_j^2 - \overline{y}_i^2 + 2(\overline{y}_i - \overline{y}_j)\overline{\overline{y}}_J + \frac{(\overline{y}_j - \overline{y}_i)^2}{N_J} \end{aligned}$$

From the hypothesis on $\overline{y}_i,\ \overline{y}_j$ and $\overline{\overline{y}}_I,\ \overline{\overline{y}}_J$:

$$V.\ N_L(S_c' - S_c) = 2(\overline{y}_i - \overline{y}_j)(\overline{\overline{y}}_J - \overline{\overline{y}}_I) + \frac{(\overline{y}_j - \overline{y}_i)^2}{N_I} + \frac{(\overline{y}_j - \overline{y}_i)^2}{N_J} > 0$$

Therefore S_c' is better than S_c, which contradicts the optimality of S_c.

Appendix B. details on the simplified crop model

We detail here the formulation of the simplified crop model $\mathcal{M}^{crop}(W, Z_1, .., Z_p)$ used in Section 6.2. $\mathcal{M}^{crop}$ computes the biomass when harvesting an annual crop based on a system of two differential equations: one for the soil humidity S, one for the crop biomass B. This is a simplified formulation of the model used in [12], where the run-off process is skipped and where irrigation is added.

$$\begin{aligned} \frac{dS(t)}{dt} &= \frac{1}{V}\left(x_{rain}(t) - 0.5\phi(S(t))\ x_{ET_0}(t)\right) \text{ if } S(t) \in [0,1],\ 0 \text{ otherwise} \\ B(T) &= B(0) + \int_0^T \psi(x_{temp}(t))\ \phi(S(t))\ x_{rad}(t)\ dt \end{aligned}$$

with $\psi(x) = \left(1 - \left(\frac{x-\tau_{temp}}{\tau_{temp}}\right)^2\right)_+$ and $\phi(S) = \max(\delta_{irr}, \min(1, \frac{S}{0.6}))$.

In this model:

- $x_{rain}(t)$ is the amount of rainfall at time (or day) t ,
- $x_{ET_0}(t)$ is the amount of evapotranspiration at time (day) t,
- $x_{temp}(t)$ is the mean temperature at time (or day) t,
- x_{rad} is the mean incident radiation at time (or day) t ,
- V is the amount of water available in the soil,
- δ_{irr} is a parameter that triggers irrigation (which removes water stress),
- τ_{temp} is a parameter that triggers stress induced by high temperatures .

The scalar output is defined as $B(T)$ where T is the harvest time. The model output can be written in the form $\mathcal{M}^{crop}(W, Z_1, Z_2, Z_3)$ by expressing the model variables and parameters from (W, Z_1, Z_2, Z_3) as follows:

- $x_{rain(.)} = W_1$
- $x_{ET_0(.)} = W_2$
- $x_{temp(.)} = W_3$
- $x_{rad(.)} = W_4$
- $\delta_{irr} = \mathbb{1}_{Z_1<0.5}$
- $V = 100 + 250.Z_2$
- $\tau_{temp} = 10 + 20.Z_3$