Approach for Meta-Modeling and Sensitivity Analysis of Computational Codes in the Presence of Dependent Random Variables. Application to a design tool for vegetated strips

Guerlain Lambert, Céline Helbert, Claire Lauvernet

Approach for Meta-Modeling and Sensitivity Analysis of Computational Codes in the Presence of Dependent Random Variables.

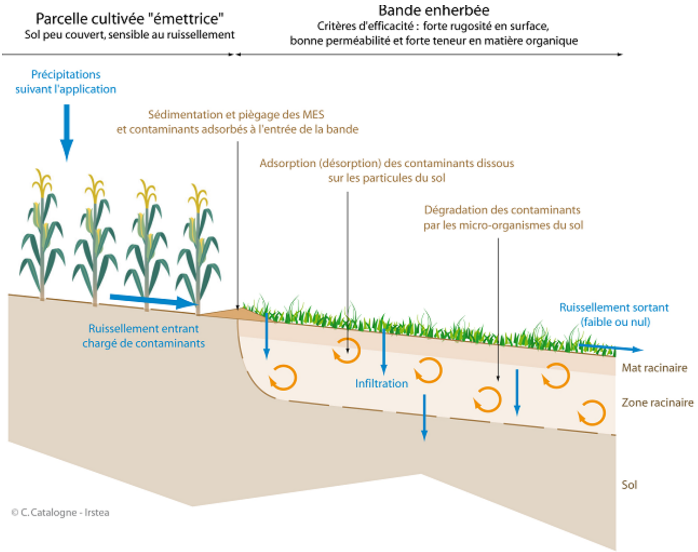
Application to a design tool for vegetated strips.

Guerlain Lambert[1]

Céline Helbert[1] & Claire Lauvernet[2]

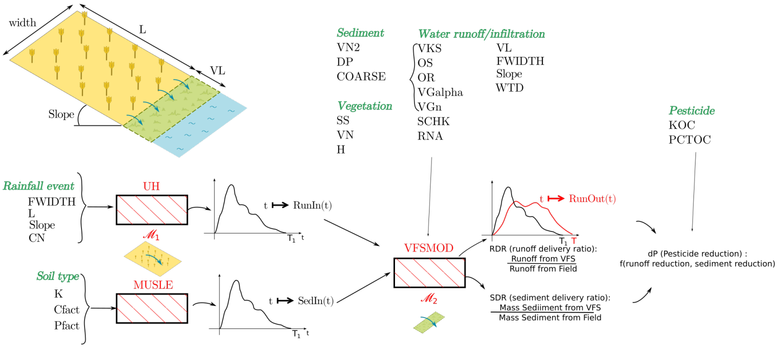
[1] Univ. Lyon, UMR CNRS 5208 École Centrale de Lyon ICJ, [2]INRAE, RIVERLY, Villeurbanne

4 décembre 2023



Réf : [Veillon et al., 2022]

Schematic diagram of BUVARD_MES

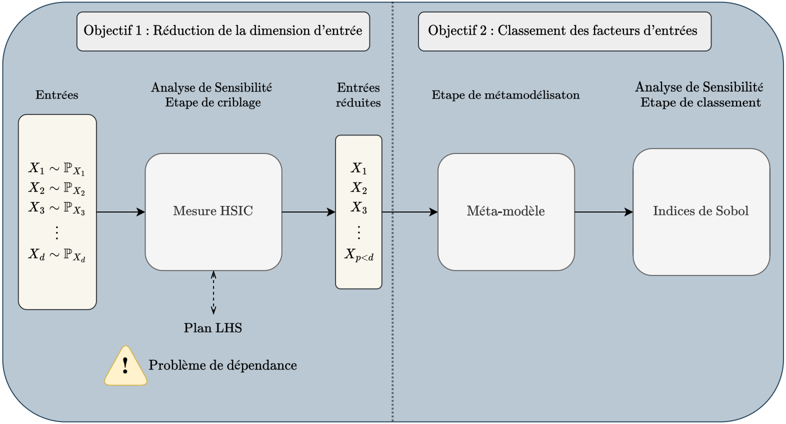


Réf : [Catalogne et al., 2018, Veillon et al., 2022]

General purpose

Goal of this work

Suggesting a global sensitivity analysis of a computationally expensive code: BUVARD_MES applied to Morcille (Beaujolais).



Purpose of this presentation

Significance of using a tailored experimental design for a sensitivity analysis approach in the presence of dependent random variables.

- Using vector quantization to create a design of experiment.
- Comparison with an existing technique : Latin Hypercube Sampling (LHS).

1. Context and motivation
2. Vectorial Quantization
   - Definition
   - In practice, how can such quantification be achieved?
   - Design of Experiment
   - Application to the BUVARD_MES group of dependent inputs
3. Screening step : HSIC measure
4. Ranking step : Sobol' indices

Vector quantification in a nutshell : [Pagès, 2018]

Quantization ? Approach a random vector X by a N-quantizer $\Gamma = \{x_1, \dots, x_N\} \in \left(\mathbb{R}^d\right)^N$:

$$\hat{X} = \mathrm{Proj}(X, \Gamma) = \sum_{i=1}^{N} x_i 1_{\{X \in C_i(\Gamma)\}} \text{ with } C_i \text{ : Voronoï partition}$$

Distorsion and optimal quantization

Let $\hat{X}$ associated to the N-quantizer $\Gamma = \{x_1, \dots, x_N\}$, $\mathbb{N} \in \mathbb{N}^*$.

- We call **distorsion** at level N, the mapping from $\left(\mathbb{R}^d\right)^N \to \mathbb{R}$:

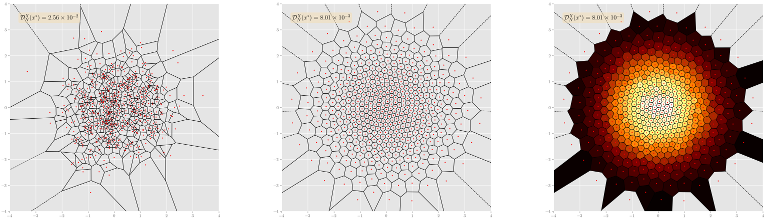
$$\mathcal{D}_N^X(\Gamma) = \|X - \hat{X}\|_2^2 = \mathbb{E}\left[\min_{1 \leq i \leq N} |X - x_i|^2\right]$$

- Any N-quantizer satisfying the following minimization problem is said to be **optimal** :

$$\inf_{\Gamma, \#\Gamma \leq N} \mathcal{D}_N^X(\Gamma) = \inf \left\{\|X - \hat{X}\|_2^2, \Gamma \subset \mathbb{R}^d, \#\Gamma \leq N\right\}$$

An example of an optimal quantification application

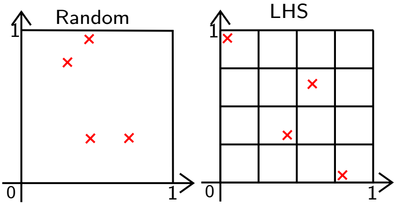
- **Context :** Application to the normal distribution $\mathcal{N}(0, I_2)$
- **Optimal quantization :** difficult to obtain in practice
- Use of a stochastic algorithm : Competitive Learning Vector Quantization (**CLVQ**)



(a) Initial (b) Sans Poids (c) Avec Poids

Réf : [Pagès, 2015]

LHS: the classic



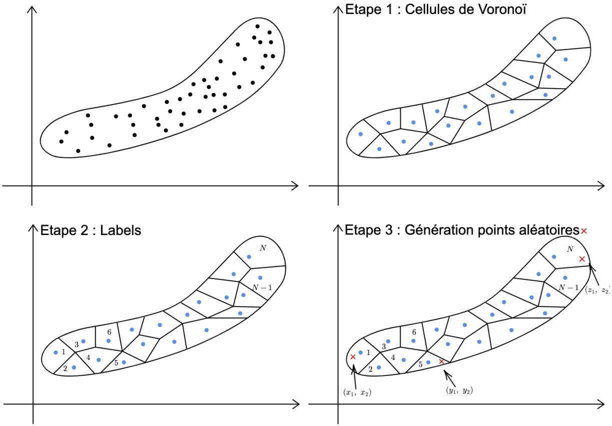
1. Divide $[0,1]^2$ into n^2 squares of equal area.
2. Randomly choose n squares so that no pair of squares is equal.
3. A point is randomly generated in each selected square.

Formally : $\mathbf{x}_j = \left(\frac{\pi_1^j - u_1^j}{n}, \dots, \frac{\pi_n^j - u_1^j}{n}\right)$

- $(u_i^1, \dots, u_i^j)_{1 \leq i \leq n}$ i.i.d $\mathcal{U}([0,1])$
- $\pi^1, \dots, \pi^d$ permutations of $\{1, \dots, n\}$

Remark : A possible extension for dependence subject to knowledge of the copula: LHSD, see : [Packham and Schmidt, 2008]

RandomQuantif : DoE using vectorial quantization

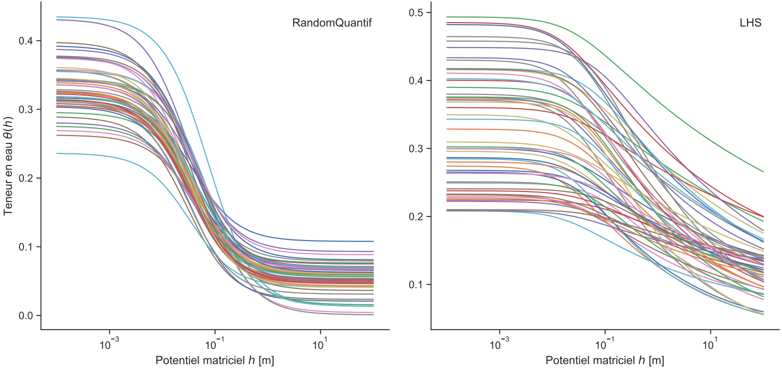


1. Generate d permutations $\pi_1, \dots, \pi_d$ of $\{1, \dots, N\}$
2. Create the design matrix X from these points.

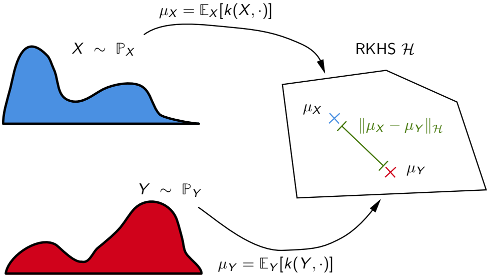
$\pi = (1, 5, \dots, N)$ $\quad\quad X = ((x_1, x_2), (y_1, y_2), \dots, (z_1, z_2))$

Interest of RandomQuantif on the Van Genuchten group

Water retention curve : $\theta(h) = \theta_s + \frac{\theta_s - \theta_r}{(1+(\alpha|h|)^n)^{1-1/n}}$



HSIC measure



HSIC measure (Réf : [Gretton et al., 2005])

Let $\mathcal{H}_i$ (resp. $\mathcal{F}$) be an RKHS associated with a kernel k_{X_i} (resp. k_y). The HSIC measure between two random variables X_i and Y is defined by :

$$\mathrm{HSIC}(X_i, Y) = \|\mu_{P_{X_i,Y}} - \mu_{P_{X_i}} \otimes \mu_{P_Y}\|^2_{\mathcal{H}_i \times \mathcal{F}}$$

with

$$\mu_{\mathbb{P}_{X,Y}} = \int_{\mathcal{X}_i \times \mathcal{Y}} k_{X_i}(u,\cdot) k_Y(v,\cdot) d\mathbb{P}_{X,Y}(u,v) = \mathbb{E}_{X,Y}\left[k_{X_i}(X,\cdot) k_Y(Y,\cdot)\right]$$

Estimation

If k_i and k_Y are characteristic (i.e. *mu* injective), then the following equivalence is true:

$$X_i \perp\!\!\!\perp Y \Leftrightarrow \mathrm{HSIC}(X_i, Y) = 0$$

Fundamental property of HSIC measurements and estimation

Given an i.i.d. copy (X_i', Y') of (X_i, Y) such that $\mathbb{E}_{X_i,X_i'}\left[k_{X_i}(X_i, X_i')\right] < +\infty$ and $\mathbb{E}_{Y,Y'}\left[k_Y(Y, Y')\right] < +\infty$, we obtain that :

$$\mathrm{HSIC}(X_i, Y) = E\left[k_{X_i}(X_i, X_i')k_{\mathcal{Y}}(Y, Y')\right] + E\left[k_{X_i}(X_i, X_i')\right] E\left[k_Y(Y, Y')\right] - 2\mathbb{E}\left[\mathbb{E}\left[k_{X_i}(X_i, X_i')\right]\mathbb{E}\left[k_Y(Y, Y')\right]\right]$$

$$\widehat{\mathrm{HSIC}}(X_i, Y) = \frac{1}{n^2}\mathrm{tr}\left(L_i H L H\right)$$

with

- L_i and L are Gram matrices.
- $H = \left(\delta_{ij} - \frac{1}{n}\right)_{1 \le i,j \le n}$ where δ_{ij} is the Kronecker symbol.

Independence test based on HSIC measures

Test of independence :

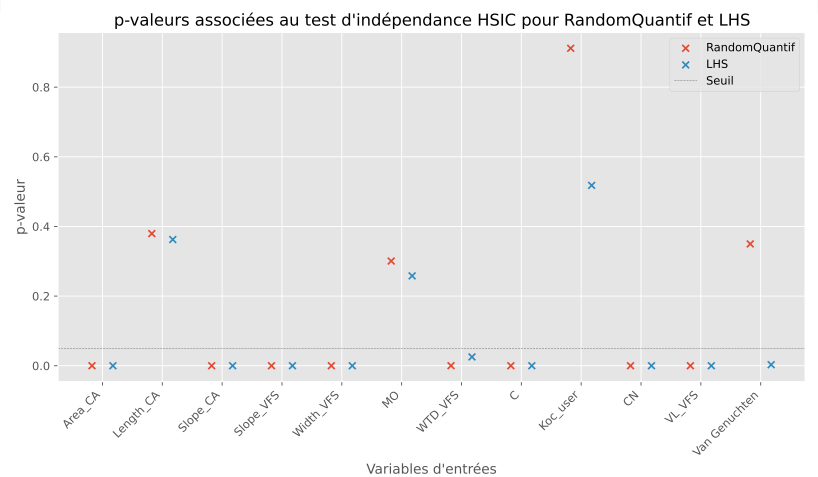
$\mathcal{H}_0$: "X_i and Y are independent"
$\mathcal{H}_1$: "X_i and Y are dependent"
equivalent to
$\mathcal{H}_0$: $\mathrm{HSIC}(X_i, Y) = 0$
$\mathcal{H}_1$: $\mathrm{HSIC}(X_i, Y) > 0$

Estimating the test statistic : $\widehat{\mathcal{S}} = n\widehat{\mathrm{HSIC}}(X_i, Y) \underset{n \to +\infty}{\sim} \gamma(\alpha, \beta)$

HSIC advantage : Requires a single sample of size $n \in \mathbb{N}^*$. !

Réf : [Gretton et al., 2007]

Application to the BUVARD_MES model



Results : LHS : Independence of the VG Group ; RandomQuantif : Dependence on the VG Group

***Expert* knowledge :** Sed_Ratio depends on VG Group

Sobol indices and estimation

Sobol' indices, [Sobol, 1993]

Let $\mathcal{M} \in \mathrm{L}^2(P_X)$, $\mathbf{u} \subset \{1, \dots, d\}$ and $\mathbf{X} = (X_1, \dots, X_d)$ with X_i independent random variables.

1. Sobol indices of the first closed order associated with $\mathbf{u}$:

$$S_{\mathbf{u}} = \sum_{B \subset \mathbf{u}} S_B = \frac{Var\left(E\left[\mathcal{M}(\mathbf{X}) \,|\, \mathbf{X}_{\mathbf{u}}\right]\right)}{V}$$

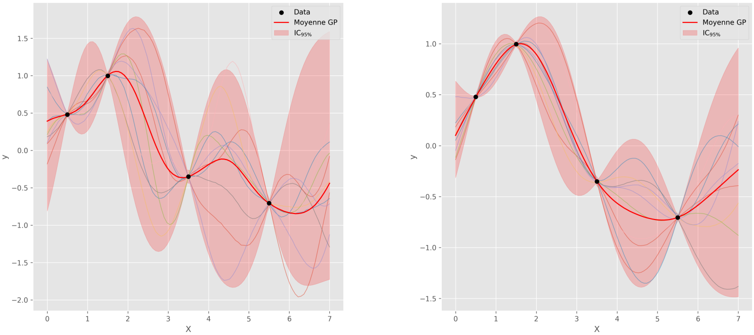
2. Sobol indices of total order associated with $\mathbf{u}$:

$$S_{\mathbf{u}}^T = 1 - S_{\bar{\mathbf{u}}}$$

Estimation : For Pick-Freeze, see for example [Saltelli et al., 2010]

Watch out! Numerous calls to the calculation code: $N \times (d+2)$ with N in the range 100 to 1000 (Réf : [Saltelli et al., 2008]) !

- $\{\eta(\mathbf{x})\}_{x \in R^d} \sim \mathcal{GP}(m, k) \Leftrightarrow$ for any finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ de $\mathbb{R}^d$, $(\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n))$ is a gaussian vector.
- Optimisation of hyperparameters by max-likelihood.



RDR_Star	Sed_Ratio	dP_100
0.990	0.980	0.985

Réf : [Rasmussen and Williams, 2005]

Ranking of inputs

Rang	RDR* (RandomQuantif)	RDR* (LHS)
1	WTD_VFS	WTD_VFS
2	CN	CNA
3	Area_CA	Area_CA
4	Width_VFS	Width_VFS
5	Van Genuchten	VL_VFS
6	VL_VFS	Slope_VFS
7	Slope_VFS	Van Genuchten
8	Slope_CA	Slope_CA
9	Length_CA	C
10	MO	MO
11	Koc_user	Koc_user
12	C	Length_CA

Table: Ranking of variables in descending order (from most influential to least influential) of total order index for RDR$*$.

Conclusion and outlook

What has been done

1. Implementation of an adapted metamodelling strategy and sensitivity analysis in the presence of dependent variables.
2. The benefits of quantization compared with an LHS plan.

An outlook

- Convergence study for HSIC estimation using vector quantization.

Catalogne, C., Lauvernet, C., and Carluer, N. (2018).
Guide d'utilisation de l'outil buvard pour le dimensionnement des bandes tampons végétalisées destinées à limiter les transferts de pesticides par ruissellement.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring Statistical Dependence with Hilbert-Schmidt Norms.
In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, volume 3734, pages 63–77. Springer Berlin Heidelberg, Berlin, Heidelberg.
Series Title: Lecture Notes in Computer Science.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007).
A Kernel Statistical Test of Independence.
In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Packham, N. and Schmidt, W. M. (2008).

Latin hypercube sampling with dependence and applications in finance.

Pagès, G. (2015).
Introduction to vector quantization and its applications for numerics.
ESAIM: Proceedings and Surveys, 48:29–79.

Pagès, G. (2018).
Numerical Probability: An Introduction with Applications to Finance.
Universitext. Springer International Publishing, Cham.

Rasmussen, C. E. and Williams, C. K. I. (2005).
Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).
The MIT Press.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010).
Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index.
Computer Physics Communications, 181(2):259–270.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008).
Global Sensitivity Analysis. The Primer, volume 304.
Journal Abbreviation: Global Sensitivity Analysis. The Primer Publication Title: Global Sensitivity Analysis. The Primer.

Sobol, I. (1993).
Sensitivity Estimates for Nonlinear Mathematical Models.

Veillon, F., Carluer, N., Lauvernet, C., and Rabotin, M. (2022).
BUVARD-MES : Un outil en ligne pour dimensionner les zones tampons enherbées afin de limiter les transferts de pesticides vers les eaux de surface.
In *50e congrès du Groupe Français des Pesticides*.