Quantization-Based Latin Hypercube Sampling for Dependent Inputs With an Application to Sensitivity Analysis of Environmental Models

Guerlain Lambert, Céline Helbert, Claire Lauvernet

Quantization-based Latin hypercube sampling for dependent inputs with an application to sensitivity analysis of environmental models.

Guerlain Lambert*[1,2], Céline Helbert†[1], and Claire Lauvernet‡[2]

[1]Institut Camille Jordan, CNRS UMR 5208, École Centrale de Lyon, Écully, France
[2]INRAE, RiverLy, 69625 Villeurbanne, France

December 19, 2024

Abstract

Numerical models are essential for comprehending intricate physical phenomena in different domains. To handle their complexity, sensitivity analysis, particularly screening, is crucial for identifying influential input parameters. Kernel-based methods, such as the Hilbert Schmidt Independence Criterion (HSIC), are valuable for analyzing dependencies between inputs and outputs. Implementing HSIC requires data from the original model, which leads to the need of efficient sampling strategies to limit the number of costly numerical simulations. While, for independent input variables, existing sampling methods like Latin Hypercube Sampling (LHS) are effective to estimate HSIC with reduced variance, incorporating dependence is challenging. This paper introduces a novel LHS variant, quantization-based LHS (QLHS), which leverages Voronoi vector quantization to address dependent inputs. The method provides good coverage of the range of variations in the input variables. The paper outlines expectation estimators based on QLHS in various dependency settings, demonstrating their unbiasedness. The method is applied on several models of growing complexities, first on simple examples to illustrate the theory, then on more complex environmental hydrological models, when the dependence is known or not, and with more and more interactive processes and factors. The last application is on the digital twin of a French vineyard catchment (Beaujolais region) to design a vegetative filter strip and reduce water, sediment and pesticide transfers from the fields to the river. QLHS is used to compute HSIC measures and independence tests, demonstrating its usefulness, especially in the context of complex models.

Keywords: design of computer experiments, vector quantization, Latin hypercube sampling (LHS), HSIC

*Email: guerlain.lambert@ec-lyon.fr
†Email: celine.helbert@ec-lyon.fr
‡Email: claire.lauvernet@inrae.fr

1 Introduction

Numerical models are used to represent and understand complex physical phenomena fields such as in biology, geophysics, or hydrology, where processes are highly interactive. Such models can be complicated (e.g., black-box models), expensive, and difficult to use when, for example, the goal is to derive a digital twin for a specific real-world context. In order to address these challenges, it may be beneficial to replace the model with a metamodel, or surrogate model. A surrogate model can be defined as a statistical model of the process-based model, and it is less costly to run than the original model. Recent studies have demonstrated the value of surrogate models in the domain of digital twins, particularly in terms of enhancing efficiency and real-time performance. White box models are subject to limitations due to incomplete knowledge and computational constraints, surrogate models offer a viable alternative see, e.g., the review of [1]. Recent studies highlight how surrogate models can improve the efficiency and applicability of digital twins: in a first application on proton exchange membrane fuel cells (PEMFCs), a hybrid surrogate model that combines a three-dimensional physical model with data-driven methods has been shown to significantly reduce computation time while maintaining high accuracy [2]. Furthermore, as digital twins are often tasked with data processing, the use of surrogate models, such as Gaussian process regressors, can prove advantageous to deal with interactive dynamics as demonstrated in [3]. These insights underscore how surrogate models can improve both the efficiency and effectiveness of digital twins. When a model deals with numerous input parameters, before learning a surrogate model it is helpful to perform a global sensitivity analysis [4] to eliminate a large set of inputs that are poorly influential on the studied outputs. For this screening step, Kernel-based methods, such as HSIC [5] which measures the dependence between inputs and outputs are very useful and have been implemented for a wide range of applications, with scalar data, vector, functional [6], or even sets [7]. The implementation of HSIC requires a sample of runs of the costly computational code, which will later be used to fit a surrogate model in a reduced-dimensional space. Therefore, it is essential to build a design of computer experiments that is both small and fills as much space as possible. This is commonly referred to as a space-filling design. Yet, the physical reality of models often imposes dependence between input variables. One example is hydrology, where soil moisture is governed by the Van Genuchten equations [8]. The parameters of the Van Genuchten model are influenced by soil type, leading to a group of interdependent variables. From a design space-filling point of view, it's therefore essential to offer a solution that takes dependence into account. In the case of independent variables several space-filling designs, such as LHS [9], and low-discrepancy sequences, in particular Sobol sequences [10] are commonly used. LHS are often preferred [11] since they ensure space-filling properties projected onto subspaces, allowing accurate estimation of metamodels and good marginal covering stable after dimension reduction. Besides, they provide good properties for estimation of expectation, as required for HSIC measure computation [6]. Extensions of LHS to account for dependency have been made by [12, 13, 14], using methods based on ranks and copulas, the latter requiring knowledge of copulas and quantile functions not always available in practice. Alternatively, kernel-based methods such as kernel herding [15] consist in minimizing a squared Maximum Mean Discrepancy (MMD) between an iteratively-built sequence of points and the target correlated joint distribution. These deterministic approaches strongly depend on the choice of the kernels and introduce a bias in estimation of expectations, such as HSIC. There is therefore an interest in providing a ready-to-use method that requires a minimum of assumptions.

In this paper, we introduce a new LHS method based on Voronoi vector quantization (VQ) to take into account dependent inputs. In [16], a design of experiments based on VQ is proposed, similar to LHS, with a Latinization procedure involving ranking Later, the approach in [17] applies

VQ to stratification by randomly drawing M points per Voronoi stratum for application to functional data. The main difficulties with these methods are that they do not take dependency into account. Our new numerical sampling strategy, called quantization-based LHS, is a direct extension of Latin Hypercube sampling based on Voronoi quantization to take into account dependence within a group of input variables. It has good properties because it swaps the bias resulting from the use of Voronoi centroids for variance by randomly drawing a point in the Voronoi cells. This ensures complete coverage of the group of dependent variables being stratified. Combined with random permutations, we get a well-distributed design across all the marginals: the independent part and the dependent part. All that is required is how to simulate its distribution (and thus conditionally on the Voronoi cells).

To this end, in Section 2, existing LHS technics and Voronoi vector quantization are briefly recalled, with special reference to optimal quantization. Then, in Section 3, the contribution of this paper is presented. That is, expectation estimation using LHS in the context of dependent random variables based on vector quantization. Different estimators are proposed to cover three different settings:

- A unique group of dependent inputs.
- A joint distribution between a group of dependent inputs and another group of independent inputs, the two groups being independent of each other.
- A joint distribution between two independent groups of dependent variables.

In particular, it is shown that the proposed estimators are unbiased. Finally, quantization-based LHS is applied to the computation of HSIC measures and independence tests in Section 4. To illustrate the relevance of these developments, they are tested and compared to other existing methods on two operational environmental models in Section 5: (i) a flood risk model where the dependency structure between inputs and the marginal laws are perfectly known and (ii) a chain of models that simulates water, sediment and pesticide transfers, where dependency is unknown. This last application is implemented on the digital twin of a real vineyard catchment in France (Morcille experimental site), where pollution of agricultural origin by pesticides is a recognized public health problem [18].

2 Existing tools : Latin hypercube sampling and Voronoi vector quantization

2.1 Latin hypercube sampling and Latin hypercube sampling with dependence

The objective of a LHS of size N, as introduced by [9], is to sample N points uniformly in $[0,1]^d$ such that marginally the projected points are well spread on $[0,1]$. The support of each coordinate, i.e. $[0,1]$, is partitioned into N sub-intervals of equal size $\frac{1}{N}$. The points are drawn in $[0,1]^d$ by associating each coordinate to one sub-interval and by uniformly sampling in this sub-interval. The LHS procedure is as follows :

We aim to estimate $\mathbb{E}[f(X)]$ where $f \in \mathrm{L}^2(\mathbb{R}^d)$ and $X \sim \mathcal{U}([0,1]^d)$, $d \in \mathbb{N}^*$. Given an LHS

Algorithm 1 LHS

1: Generate N independent samples $(U_{i1}, \ldots, U_{id})_{i=1,\ldots,N}$, where U_{ij} is i.i.d $\mathcal{U}([0,1])$.
2: Generate d independent equiprobable permutations $\pi_1, \ldots, \pi_d$ of $\{1, \ldots, N\}$. $\pi_j(i)$ is the value to which i is mapped by the j^{th} permutation.
3: An LHS is given by:

$$\begin{cases} V_{ij} & = & \frac{\pi_j(i)-1}{N} & + & \frac{U_{ij}}{N} \\ j & = & 1, \ldots, d & , & i & = & 1, \ldots, & N \end{cases}$$

$(V_{i1}, \ldots, V_{id})_{1 \leq i \leq N}$ of $\mathcal{U}([0,1]^d)$, the LHS estimator of the expected value is :

$$\mu_{LHS} = \frac{1}{N} \sum_{i=1}^{N} f(V_{i1}, \ldots, V_{id})$$

It is unbiased, and the variance satisfies $Var(\mu_{LHS}) \leq \frac{N \times Var(\mu_{MC})}{N-1}$, implying that using an LHS of size N ensures a variance that is less than or equal to that of a Monte Carlo (MC) sample of the same size, see [19]. Furthermore, a central limit theorem for the estimator is provided in [20]. It stratifies the marginal distribution to maximize coverage of the range of each variable. However, the use of LHS imposes the independence of random variables, which raises an issue if the model inputs are correlated. Therefore, it is crucial to use an appropriate sampling strategy to take dependence into account. Over the years, several modifications to LHS have been proposed to incorporate dependence. For instance, a rank-based approach was introduced in [12] and further improved by [13], though this method results in a biased estimator.

Recently, a copula-based Latin Hypercube Sampling method, Latin Hypercube Sampling with Dependence (LHSD), was introduced in [14], allowing the incorporation of dependence into the experimental design while retaining the properties of LHS. The LHSD procedure proposed by [14] is based on the copula (and conditional copulas) construction summarized in Algorithm 2. Consider a random vector $X = (X_1, ..., X_d)$ of joint c.d.f F where the marginal c.d.f are denoted F_j for all $j = 1, \ldots, d$ and C a copula. In particular, C is a distribution with uniform marginals. Sklar's theorem [21] allows a copula to be associated with any multidimensional distribution. The copula makes it possible to model the dependence between the marginals, if F is continuous and let $F_j(X_j) =: U_j$, then :

$$C(U_1, \ldots, U_d) = F\left(F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d)\right)$$

The j-th conditional copula is defined by :

$$C_j(U_j \mid U_1, \ldots, U_{j-1}) = \varphi(U_1, \ldots, U_{j-1}, U_j)$$

where $\varphi(u_1, \ldots, u_{j-1}, u_j) = \frac{\partial^{j-1} C(u_1, \ldots, u_j, 1, \ldots, 1)}{\partial u_1 \ldots \partial u_{j-1}}$. The inverse of the j-th conditional copula, denoted C_j^{-1} , at some probability $0 \leq q \leq 1$, solves the equation:

$$C_j(u_j \mid U_1 = u_1, \ldots, U_{j-1} = u_{j-1}) = q$$

for $u_j \in [0,1]$. Explicitly, C_j^{-1} is defined as:

$$C_j^{-1}(q \mid u_1, \ldots, u_{j-1}) = u_j,$$

Algorithm 2 LHSD from [14]

1: Generate an LHS sample $(Z_{i1}, \ldots, Z_{id})_{i=1,\ldots,N}$ of $[0,1]^d$ using Algorithm 1.
2: Construct sequentially $(U_{i1}, \ldots, U_{id})_{i=1,\ldots,N}$ from the joint copula using inverse conditional copula functions:

$$U_{ij} = C_j^{-1}(Z_{ij}|U_{i1}, \ldots, U_{ij-1})$$

for $j = 1 \ldots d$.
3: Construct the final sample : $(X_{i1}, \ldots, X_{id})_{i=1,\ldots,N}$ using the inverse distribution function $X_{ij} = F_j^{-1}(U_{ij})$ for $j = 1, \ldots, d$.

The properties of the LHSD estimator are similar to those of the LHS as justified in [14]. Specifically, when estimating an expected value, the variance of LHSD is lower than the variance from a simple Monte Carlo sample.

The aim is to reconstruct the dependence of the marginals sequentially by constructing a sample on $[0,1]$ from the inverse of the conditional copula, starting from an LHS sample. To obtain a LHSD, apply the quantile transformation to this sample.

The LHSD method involves selecting a copula from a model to describe the dependency structure (e.g, Gaussian, Clayton), estimating its parameters, and accessing the inverses of the conditional copulas as well as the inverse c.d.f. of marginals. This process can be complex from a numerical perspective, leading to estimation errors and suboptimal copula selection. Ultimately, this can result in a less efficient expectation estimate.

In contrast, our proposed method based on vector quantization overcomes these limitations by only requiring the ability to simulate the joint distribution. This data-driven approach eliminates the need for copula selection and estimation, thereby reducing implementation complexity and the risk of errors. The few requirements of our method are directly linked to its reliance solely on the joint distribution points, making it a more practical and robust alternative.

2.2 Background on vector quantization

Vector quantization was first introduced in signal processing during the 1950s as a method of discretizing continuous signals. It is now widely used in various applications, including speech recognition [22], image compression [23], and numerical probability [24]. The latter is of particular interest to us. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Let X be a random vector with values in $\mathbb{R}^d$ with $\mathbb{P}_X$ its distribution and $N \in \mathbb{N}^*$. Let $\Gamma = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$, the Voronoi partition associated to Γ is defined as $(C_i)_{i=1\ldots N}$ such that

$$C_i(\Gamma) \subset \left\{ y \in \mathbb{R}^d : |y - x_i| \leq \min_{1 \leq j \leq N} |y - x_j| \right\}$$

The Voronoi quantizer can be defined as follows:

$$q_{vor}(X) = \sum_{i=1}^{N} x_i \mathbb{1}_{C_i(\Gamma)}(X)$$

The quantized variable $q_{vor}(X)$ is often denoted as $\widehat{X}$. It is the discrete version of X with the N support points $x_1, \ldots, x_N$. Similarly, the quantized probability distribution $\mathbb{P}_{\widehat{X}}$ of $\mathbb{P}_X$ induced by Γ is given by :

$$\mathbb{P}_{\widehat{X}} = \sum_{i=1}^{N} \mathbb{P}_X \left(C_i(\Gamma) \right) \delta_{x_i}$$

where δ_{x_i} is the Dirac mass centered on x_i.

To assess the performance of the $\widehat{X}$ quantizer, we introduce the quadratic distortion function associated to $\Gamma = \{x_1, \ldots, x_N\}$:

$$\mathcal{D}_N^X(\Gamma) = \|d(X, \Gamma)\|^2_{L^2(\mathbb{P})} = \mathbb{E}\left[\min_{1 \leq i \leq N} |X - x_i|^2 \right] = \int_{\mathbb{R}^d} \min_{1 \leq i \leq N} |x_i - y|^2 \mathbb{P}_X(\mathrm{d}y)$$

Any quantizer that minimizes distortion is called an optimal quantizer. If $X \in L^2(\mathbb{P})$, then the existence of such quantifiers is assured. However, uniqueness is not systematically obtained (1D case of unimodal distributions) [25, 24]. Furthermore, any N-optimal quantizer $\widehat{X}$ is a stationary quantizer i.e :

$$\widehat{X} = \mathbb{E}\left[X \,|\, \widehat{X} \right]$$

In practice, this property enables the construction of fixed-point algorithms to obtain optimal (or at least suboptimal) quantization. The most well-known of these algorithms is Lloyd's algorithm [26], also known as k-means. It is very popular and easy to implement, especially if the distribution of X is known from a large sample of simulated points. An example of optimal quantization is given in figure 1, based on the centered bivariate normal distribution $X = (X_1, X_2)$ with $cov(X_1, X_2) = 0.8$.

In the context of vector quantization, cubature formulas are available to compute $\mathbb{E}[f(X)]$ where X is a random vector on $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ a function [24]:

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(\widehat{X})] = \sum_{i=1}^{N} f(x_i) \mathbb{P}\left[\widehat{X} = x_i \right] = \sum_{i=1}^{N} f(x_i) \mathbb{P}\left[X \in C_i \right]$$

The accuracy of this estimate depends on the regularity of f. If f is L-Lipschitz, $L > 0$, then

$$\left| \mathbb{E}\left[f(X) \right] - \mathbb{E}\left[f\left(\widehat{X} \right) \right] \right| \leq L \sqrt{\mathcal{D}_N^X(\Gamma)}$$

If $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, ∇f is L-Lipschitz, and $\widehat{X}$ is stationary, then :

$$\left| \mathbb{E}\left[f(X) \right] - \mathbb{E}\left[f\left(\widehat{X} \right) \right] \right| \leq \frac{L}{2} \mathcal{D}_N^X(\Gamma)$$

Unlike Monte Carlo estimation, vector quantization estimation is deterministic. The estimator is therefore variance-free, but biased, whereas Monte Carlo estimation is unbiased, but with variance. This leads to modification of estimation through vector quantization with the addition of controlled randomness, such as in stratification. Thus, one contribution of the paper is to introduce a stochastic version of vector quantization. This procedure will be used for the group of dependent inputs and associated to independent inputs in an LHS-style. This new methodology called quantization-based LHS is discussed in the following section.

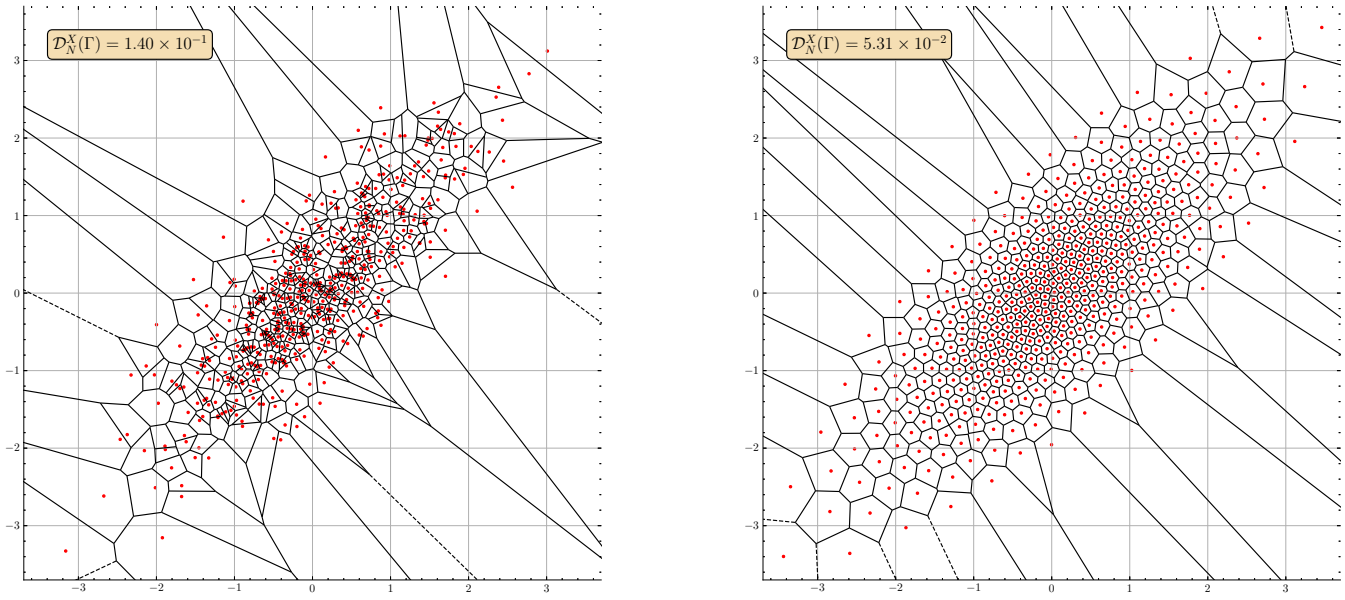

Figure 1: Voronoi quantization of the centered bivariate, with covariance of 0.8 between marginals encoded into $N = 500$ centroids. On the left : Voronoi tessellation of 500 random points from the joint distribution. On the right, Voronoi tessellation of 500 centroids obtained via optimal quantization (kmeans).

3 Quantization-based Latin hypercube sampling

In this section, we introduce different sampling strategies which account for dependency and such that the space-filling property is maintained after dimension reduction, so they can be used for screening purposes. It should be noted that we do not guarantee a strict space-filling property in the multidimensional space. However, we do achieve a well-distributed sample by ensuring a good spread in each marginal dimension and in each group of dependent inputs. For each sampling strategy, we study the estimation of $m = \mathbb{E}[f(X)]$, for $f : \mathbb{R}^d \to \mathbb{R}$, $f \in \mathrm{L}^2(\mathbb{R}^d)$, where

- case 1 : $X = (X_1, \ldots, X_d)$ with d dependent components.
- case 2 : $X = (X_{\text{dep}}, X_{\text{indep}})$ with X_{dep} composed of s dependent components, and $X_{\text{indep}} = (X_{s+1}, \ldots, X_d)$ composed of $d - s$ independent components and X_{dep} and X_{indep} are independent.
- case 3 : $X = (X_{G_1}, X_{G_2})$ with $X_{G_1} = (X_1, \ldots, X_s)$ composed of s dependent components, $X_{G_2} = (X_{s+1}, \ldots, X_d)$ composed of $d - s$ dependent components and X_{G_1} and X_{G_2} are independent.

3.1 Random quantization (case 1)

In this section, X is composed of a unique group of d dependent components $X_1, \ldots, X_d$. We introduce a stochastic version of vector quantization, called Random Quantization (RQ) to account for dependency. RQ is a stratification technique. The strata are the Voronoi cells obtained after quantization. A point is randomly drawn in each cell according to the probability distribution of X conditional on the cell. RQ is summarized in Algorithm 3.

Algorithm 3 RQ

Let $X \in \mathbb{R}^d$ be a random vector composed of d dependent components.
Let $(x_{i1}, \ldots, x_{id})_{i=1,\ldots,N}$ be a N-optimal quantizer of X. Let $(C_i)_{i=1,\ldots,N}$ be the associated Voronoi partition of $\mathbb{R}^d$.
for $i = 1$ to N **do**
Generate one random point $U_i = (U_{i1}, \ldots, U_{id})$ in the cell C_i according to the probability distribution of X conditioned on C_i, i.e. $U_i \sim \mathcal{L}(X \mid X \in C_i)$.
end for
return $U = (U_1, \ldots, U_N)$

The use of randomized vector quantization for design of computer experiment has the advantage that it only requires knowledge of how to simulate in Voronoi cells. No knowledge of copulas or quantile functions is necessary. This allows for the entire distribution of X to be explained using a finite number of support points, while perfectly accounting for input dependence.

Definition 3.1. Let $(U_i)_{i=1\ldots N}$ a sample provided by Algorithm 3. We define the following RQ estimator :

$$\mu_{RQ} := \sum_{i=1}^{N} f(U_i)\mathbb{P}[X \in C_i]. \tag{1}$$

Proposition 3.1. μ_{RQ} is an unbiased estimator of m, and its variance is given by :

$$\mathrm{Var}\left(\mu_{RQ}\right) = \sum_{i=1}^{N} \mathbb{P}[X \in C_i]^2 \mathrm{Var}\left(f(U_i)\right)$$

The proof of Proposition 3.1 can be found in Appendix A. We illustrate the behavior of μ_{RQ} on two toy examples. We first consider the function $f : \mathbb{R} \to \mathbb{R}$ defined for all $x \in \mathbb{R}$ by $f(x) = x^2$ and look for $\mathbb{E}[f(X)]$ with $X \sim \mathcal{N}(0, 1)$. In the context of using a costly numerical model, the number of model evaluations is limited, therefore we choose low values of N (10, 20, 50, and 100) and the behavior of μ_{RQ} is studied through 1000 repetitions. The results, summarized in Figure 2, show that the estimator is unbiased and that its variance decreases as N increases. μ_{RQ} is compared to Monte Carlo and LHS estimation. We recall that to obtain an LHS sample we first apply Algorithm 1 to produce a sample in $[0, 1]$ to which the inverse of standard Gaussian c.d.f is applied. μ_{RQ} exhibits lower variance at low N, making it the most efficient method in this case. It should be noted that, in this example, there is no dependency, yet RQ utilization is still feasible. This straightforward example offers an opportunity to examine the characteristics of the proposed μ_{RQ} weight estimator, which has been demonstrated to be effective. We consider a second 2D example with correlation between marginals. Let $X = (X_1, X_2)$ be a centered Gaussian vector such that $cov(X_1, X_2) = 0.8$. To estimate $\mathbb{E}(X_1 X_2)$, Monte Carlo, LHSD and RQ were used. The results are shown in Figure 3. A Gaussian copula with a correlation of 0.8 was used, and an analytical expression for the inverse of the conditional copula is known. All three estimators are unbiased, and the proposed estimator, μ_{RQ}, has minimal variance.

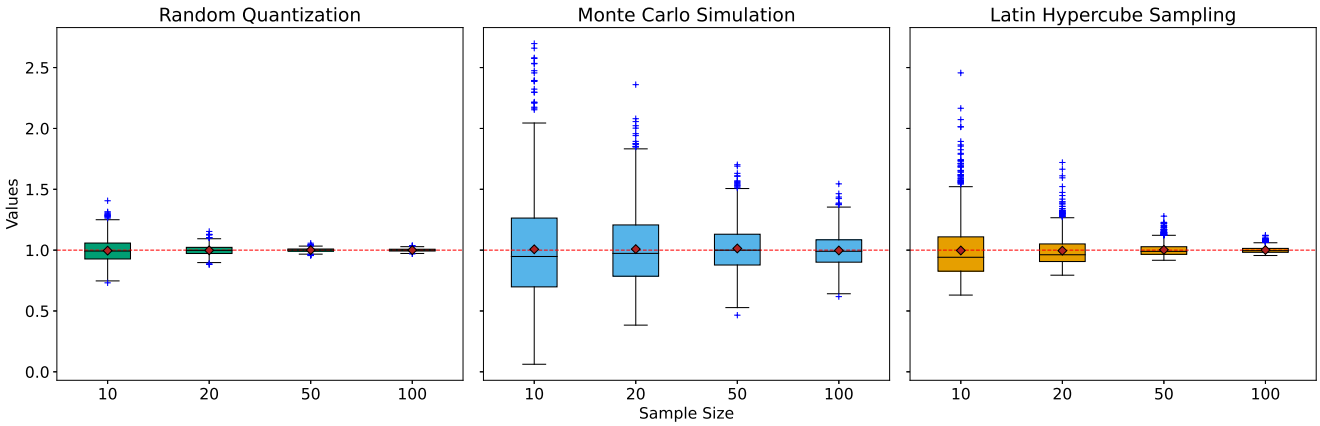

Figure 2: Estimation of $\mathbb{E}(X^2)$ where $X \sim \mathcal{N}(0,1)$ with sample size $N = 10, 20, 50, 100$ and 1000 repetitions per N. The red dashed line is the theoretical value.

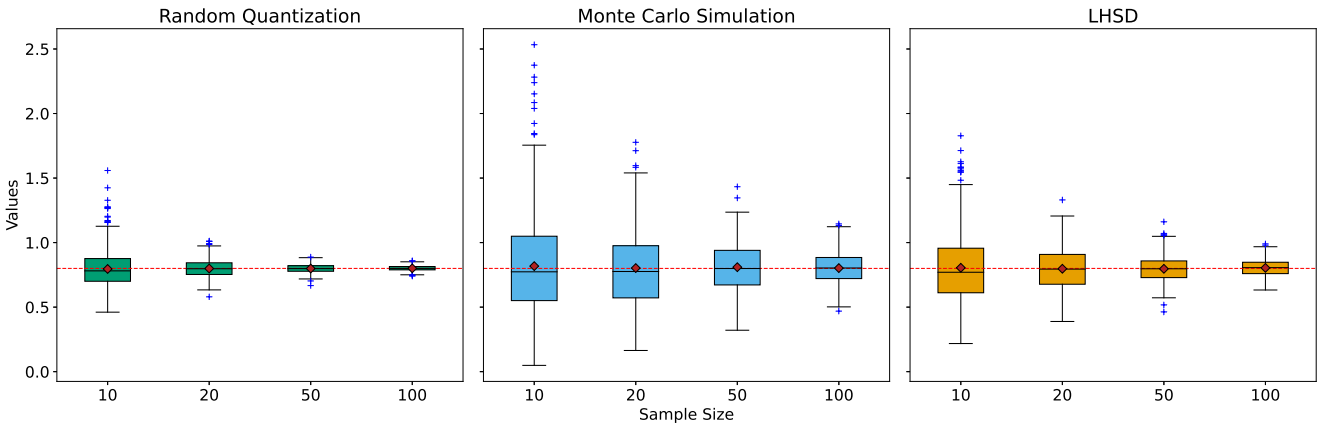

Figure 3: Estimation of $\mathbb{E}(X_1 X_2)$ where $X = (X_1, X_2)$ is a centered Gaussian vector with covariance 0.8 with sample size $N \in \{10, 20, 50, 100\}$ and 500 repetitions per N. The red dashed line is the theoretical value of 0.8.

3.2 Quantization-based Latin hypercube sampling (case 2)

Consider $X_{\text{dep}} = (X_1, \ldots, X_s)$, $s \in \mathbb{N}^*$, a random vector with dependent components and $X_{\text{indep}} = (X_{s+1}, \ldots, X_d)$, a group of independent random variables. X_{dep} and X_{indep} are independent. In the dependent group of inputs X_{dep}, the sole assumption made regarding the dependence structure is

$$\mathbb{P}_{(X_1,\ldots,X_s)} \neq \bigotimes_{i=1}^{s} \mathbb{P}_{X_i}$$

Then, it can be either linear (correlation) or more general. We want to summarize the theoretical distribution of $(X_{\text{dep}}, X_{\text{indep}})$ though an empirical sample of N points in $\mathbb{R}^d$ which preserves space-filling properties after dimension reduction. To do so, we stratify X_{dep} using the Random Quantification procedure introduced in section 3.1. In the same way, we stratify X_{indep} using an LHS sample. The idea is then to associate to each stratum of X_{dep} a stratum of X_{indep} using a random permutation π. The sampling scheme is described in Algorithm 4.

Algorithm 4 Quantization-based LHS

Let $X_{\text{dep}} \in \mathbb{R}^s$ be a random vector composed of s dependent components.
Apply Algorithm 3 to provide a RQ sample $U = (U_1, \ldots, U_N)$ of X_{dep}.
Let $X_{\text{indep}} \in \mathbb{R}^{d-s}$ be a random vector composed of $d - s$ independent components.
Apply Algorithm 1 to provide an LHS sample $V = (V_1, \ldots, V_N)$ of X_{indep}.
Let π be a random permutation of $\{1, \ldots, N\}$ in $\{1, \ldots, N\}$.
return $((U_1, V_{\pi(1)}), \ldots, (U_N, V_{\pi(N)}))$

Definition 3.2. Let $((U_i, V_{\pi(i)}))_{i=1\ldots N}$ a sample provided by Algorithm 4 where $U_i \sim \mathcal{L}(X_{\text{dep}} | X_{\text{dep}} \in C_i)$ and $(C_i)_{i=1,\ldots,N}$ is the Voronoi tessellation associated to the quantization of X_{dep}. We define the following Quantization-based LHS estimator :

$$\mu_{QLHS} := \sum_{i=1}^{N} \mathbb{P}[X_{\text{dep}} \in C_i] f(U_i, V_{\pi(i)}) \tag{2}$$

Proposition 3.2. μ_{QLHS} is an unbiased estimator of m.

The proof of Proposition 3.2 can be found in Appendix A. To illustrate the behavior of the μ_{QLHS} estimator, we begin by considering an initial case. Similar to RQ, this case assumes no dependencies. However, it allows for testing the previously introduced μ_{QLHS} estimator, which combines a stratified variable with weights varying from one state to another and another variable with identical weights. We study the behavior of μ_{QLHS} through the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined for all $(x, y) \in \mathbb{R}^2$ by $f(x, y) = x^2 y$. The problem is to estimate $\mathbb{E}[f(X_{\text{dep}}, X_{\text{indep}})]$ where $X_{\text{dep}} \sim \mathcal{N}(0, 1)$ and $X_{\text{indep}} \sim \mathcal{U}([0, 1])$. The results are summarized in Figure 4, leading to the same conclusion as for μ_{RQ}. Specifically, the estimator is unbiased and has decreasing variance, resulting in better performance than the typical Monte Carlo and LHS techniques. By estimating $\mathbb{E}((X_1 + X_2)^2 X_{\text{indep}})$ through the addition of dependence in $X_{\text{dep}} = (X_1, X_2)$ from a centered bivariate Gaussian vector, we have obtained results shown in Figure 5. The LHSD method is parameterized based on the Gaussian copula with parameter $\rho = 0.8$. It is observed that μ_{QLHS} is unbiased with lower variance than the Monte Carlo and LHSD methods for all sample sizes, and therefore offers the best performance.

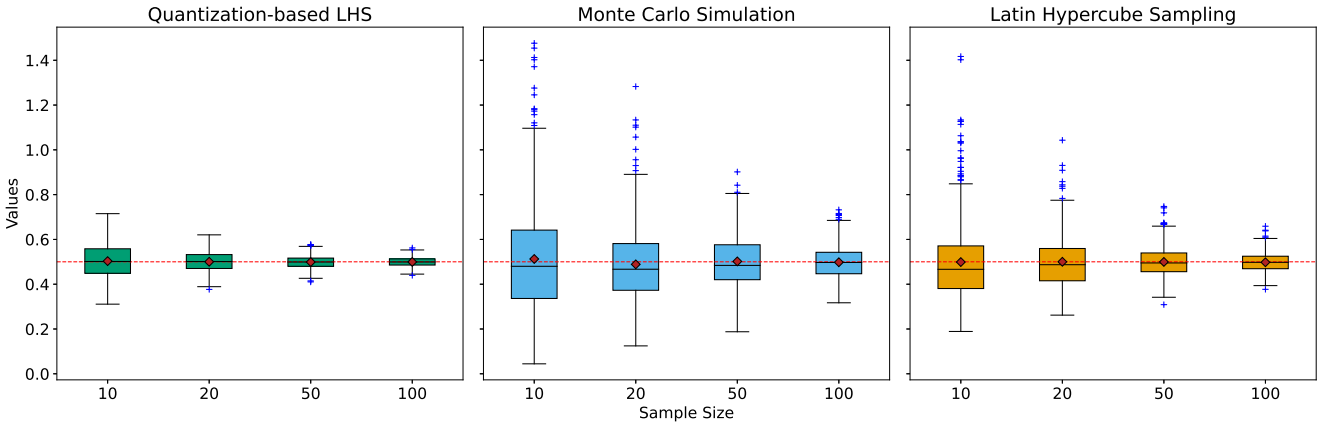

Figure 4: Estimation of $\mathbb{E}(X_{\text{dep}}^2 X_{\text{indep}})$ where $X_{\text{dep}} \sim \mathcal{N}(0,1)$ and $X_{\text{indep}} \sim \mathcal{U}([0,1])$ with sample size $N \in \{10, 20, 50, 100\}$ and 1000 repetitions per N. The red dashed line is the theoretical value.

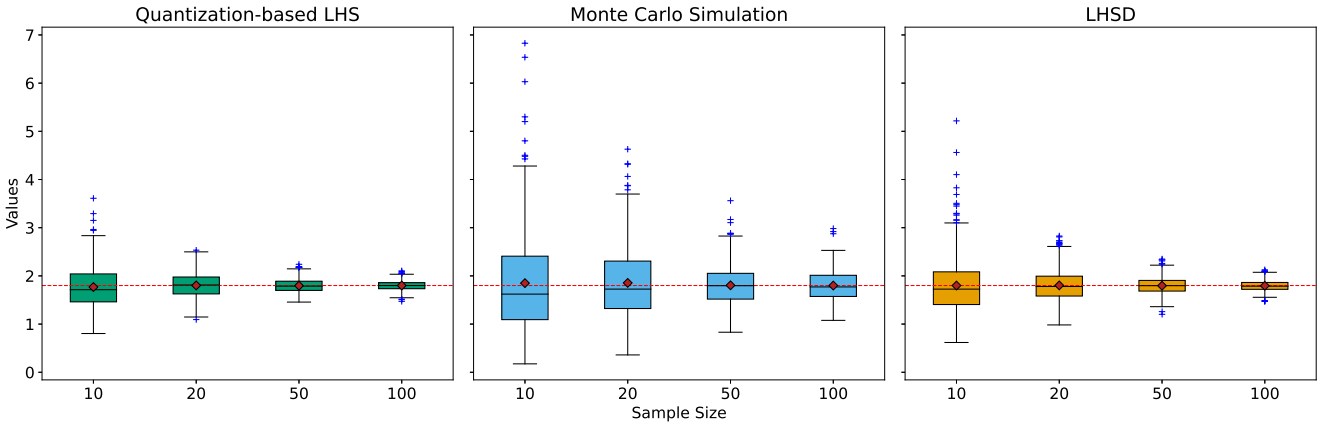

Figure 5: Estimation of $\mathbb{E}((X_1 + X_2)^2 X_{\text{indep}})$ where $X_{\text{dep}} = (X_1, X_2)$ is a centered bivariate Gaussian vector with covariance $cov(X_1, X_2) = 0.8$ and $X_{\text{indep}} \sim \mathcal{U}([0,1])$ with sample size $N \in \{10, 20, 50, 100\}$ and 500 repetitions per N. The red dashed line is the theoretical value.

3.3 Double Quantization-based Latin hypercube sampling

Considering two independent random vectors $X_{G_1} \in \mathbb{R}^s$ and $X_{G_2} \in \mathbb{R}^{d-s}$ where G_1 (resp. G_2) stands for "Group 1" (resp. 2), each composed of several dependent variables. We want to compute $\mathbb{E}[f(X_{G_1}, X_{G_2})]$ where $f : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function. Let consider two samples $U = (U_1, \ldots, U_N)$ and $V = (V_1, \ldots, V_N)$ of X_{G_1} and X_{G_2} obtained via RQ from Algorithm 3. These two samples preserve the correlation inside each group of dependent variables. In the same manner as in the previous section, the idea is to associate to each stratum of X_{G_1} a stratum of X_{G_2} using a random permutation π. The sampling scheme is described in Algorithm 5.

Algorithm 5 Double Quantization-based LHS

Let $X_{G_1} \in \mathbb{R}^s$ be a random vector composed of s dependent components.
Apply Algorithm 3 to provide a RQ sample $U = (U_1, \ldots, U_N)$ of X_{G_1}.
Let $X_{G_2} \in \mathbb{R}^{d-s}$ be a random vector composed of $d - s$ dependent components.
Apply Algorithm 3 to provide a RQ sample $V = (V_1, \ldots, V_N)$ of X_{G_2}.
Let π be a random permutation of $\{1, \ldots, N\}$ in $\{1, \ldots, N\}$.
return $((U_1, V_{\pi(1)}), \ldots, (U_N, V_{\pi(N)}))$

Definition 3.3. Let $((U_i, V_{\pi(i)}))_{i=1\ldots N}$ a sample provided by Algorithm 5 where

- $U_i \sim \mathcal{L}(X_{G_1} | X_{G_1} \in C_i^{X_{G_1}})$ and $(C_i^{X_{G_1}})_{i=1,\ldots,N}$ is the Voronoi tessellation associated to the quantization of X_{G_1}
- $V_j \sim \mathcal{L}(X_{G_2} | X_{G_2} \in C_j^{X_{G_2}})$ and $(C_j^{X_{G_2}})_{j=1,\ldots,N}$ is the Voronoi tessellation associated to the quantization of X_{G_2}

We define the following Q2LHS estimator :

$$\mu_{Q2LHS} := \frac{1}{\sum_{i=1}^N p_i q_{\pi(i)}} \sum_{i=1}^N p_i q_{\pi(i)} f\left(U_i, V_{\pi(i)}\right) \tag{3}$$

where $\forall 1 \leq i \leq N$, $p_i = \mathbb{P}(X_{G_1} \in C_i^{X_{G_1}})$ and $\forall 1 \leq j \leq N$, $q_j = \mathbb{P}(X_{G_2} \in C_j^{X_{G_2}})$.

Figure 6 shows 1000 iterations of the estimate of $\mathbb{E}(X_{G_1} X_{G_2}^2 + X_{G_2}^2)$ where $X_{G_1} \sim \mathcal{LN}(0, 1)$ (log-normal distribution) and $X_{G_2} \sim \mathcal{N}(0, 1)$ with its confidence interval. It is observed that the Q2LHS estimator is asymptotically unbiased. Figure 7 compares the Q2LHS estimator with Monte Carlo and LHSD with $N \in \{10, 20, 50, 100\}$. The estimator is asymptotically unbiased and performs better than conventional Monte Carlo and LHS estimators, with smaller variance and fewer outliers.

4 Application to kernel-based sensitivity analysis

When building a surrogate model (or metamodel) for a numerical computation code, it is useful to carry out a preliminary selection of the most influential input variables [27]. This step is called screening and is crucial when the number of input parameters is large. It allows tuning the meta-model in a limited dimension space and thus requires a reduced number of costly evaluations of computer experiments. The goal of this section is to show how Quantization-based LHS allows estimating Sensitivity Analysis HSIC measures taking dependence into account.

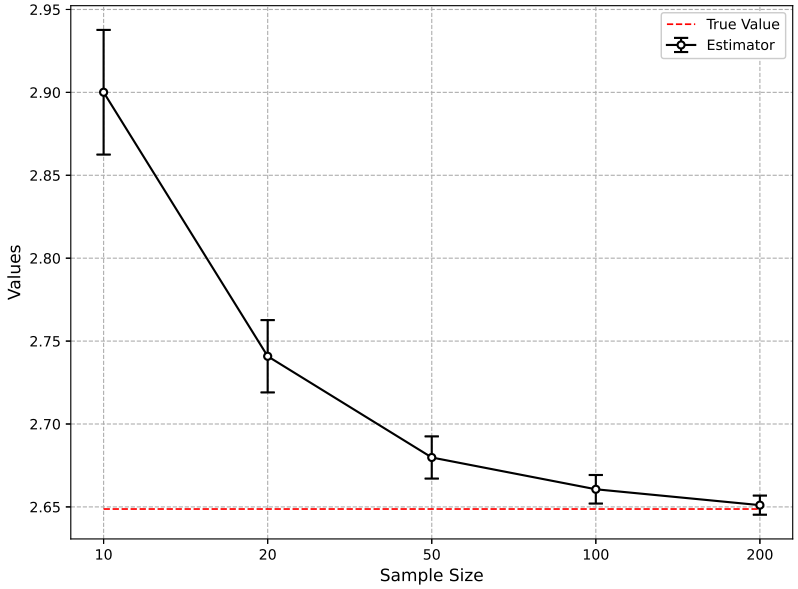

Figure 6: Estimation of $\mathbb{E}(\mu_{Q2LHS})$ with the sample size when estimating $\mathbb{E}(X_{G_1}X_{G_2}^2 + X_{G_2}^2)$ where $X_{G_1} \sim \mathcal{LN}(0,1)$ and $X_{G_2} \sim \mathcal{N}(0,1)$. 1000 repetitions of the estimation are performed with $N \in \{10, 20, 50, 100\}$. The 95% confidence intervals are shown.

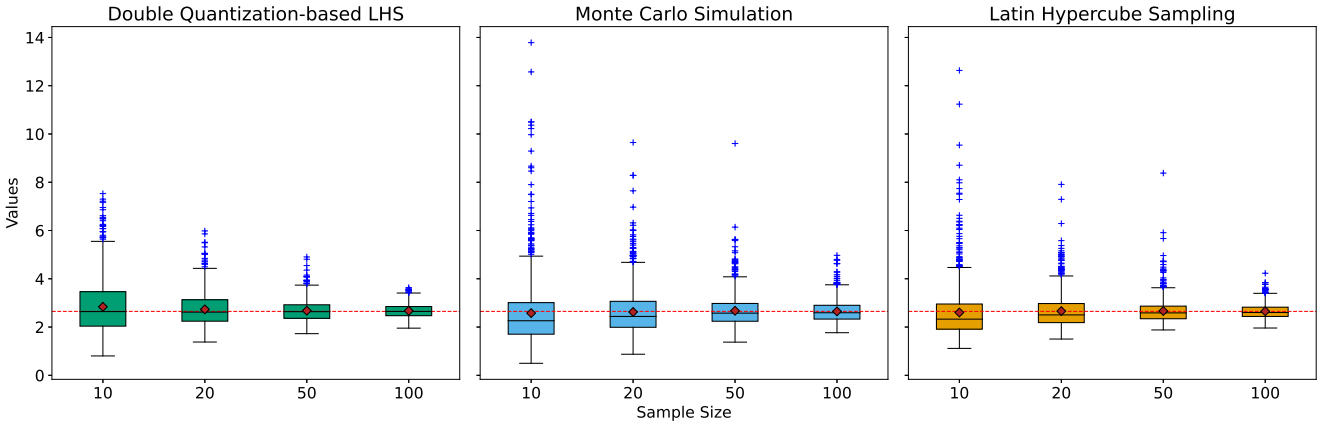

Figure 7: Estimation of $\mathbb{E}(X_{G_1}X_{G_2}^2 + X_{G_2}^2)$ where $X_{G_1} \sim \mathcal{LN}(0,1)$ and $X_{G_2} \sim \mathcal{N}(0,1)$ with sample size $N = 10, 20, 50, 100$ and 1000 repetitions per N. The red dashed line is the theoretical value i.e $e^{0.5} + 1 \approx 2.64$.

4.1 Kernel-based sensitivity analysis

Consider a numerical model $\mathcal{M}$ such that for $X \in \mathbb{R}^d$, $\mathcal{M}(X) \in \mathbb{R}$. Let $K \neq \emptyset$ and $\mathcal{H}$ be a Hilbert space of real functions in K.

Definition 4.1 (Reproducing kernel)**.** A kernel $k : K \times K \to \mathbb{R}$ of $\mathcal{H}$ is reproducing if we have : $\forall x \in K, \quad k(\cdot, x) \in \mathcal{H}$ and if it verifies the reproducing property :

$$\forall f \in \mathcal{H}, \quad x \in K, \quad f(x) = \langle f, k(\cdot, x) \rangle$$

The space $\mathcal{H}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if, for all $x \in K$, the Dirac function $\delta_x : \mathcal{H} \to K$ defined as

$$\forall f \in \mathcal{H}, \quad \delta_x(f) = f(x)$$

is continuous.

For more details on RKHS, the reader is referred to [28].

Definition 4.2 (Kernel embedding)**.** Let $\mathcal{M}_1^+$ the space of probability measures on K. Consider $\mathcal{H}$ the RKHS induced by a kernel $k : K \times K \to \mathbb{R}$. We define the kernel mean embedding as

$$\mu : \begin{cases} \mathcal{M}_1^+ \to \mathcal{H} \\ \mathbb{P} \mapsto \int k(\cdot, x) \mathrm{d}\mathbb{P}(x) \end{cases}$$

Consider $\mathbf{X} = (X_1, \ldots, X_d)$ a random vector defined on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and Y a scalar output (which can be extended to vector or functional outputs) where $Y := \mathcal{M}(X)$ and $M : \mathcal{X} \to \mathbb{R}$ is a black-box numerical model. For a given set of indices $A \subset \{1, \ldots, d\}$, we define the random vector X_A as $(X_i)_{i \in A}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution $\mathbb{P}_{X_A}$.

Definition 4.3 (HSIC measure)**.** Let $A \subset \{1, \ldots, d\}$. Let $\mathcal{H}$ be the RKHS of functions of $\mathcal{X}_A$ in $\mathbb{R}$ with kernel $k := \bigotimes_{i \in A} k_i$, and let $\mathcal{F}$ be the RKHS of functions of $\mathcal{Y}$ in $\mathbb{R}$ with kernel k_Y. The Hilbert-Schmidt independence criterion (HSIC) measures the distance between the embeddings of two distributions : the joint probability distribution $\mathbb{P}_{(X_A, Y)}$ of (X_A, Y) and the product of the marginal probability distributions $\mathbb{P}_{X_A}$ and $\mathbb{P}_Y$ and is given by:

$$\mathrm{HSIC}(X_A, Y) = \mathrm{MMD}\left(\mathbb{P}_{(X_A,Y)}, \mathbb{P}_{X_A} \otimes \mathbb{P}_Y\right)^2 = \left\| \mu\left(\mathbb{P}_{(X_A,Y)}\right) - \mu\left(\mathbb{P}_{X_A}\right) \otimes \mu\left(\mathbb{P}_Y\right) \right\|^2_{\mathcal{H} \times \mathcal{F}}$$

where $\mu\left(\mathbb{P}_{(X_A,Y)}\right) = \mathbb{E}\left[k(X_A, \cdot) k_Y(Y, \cdot)\right]$ is the kernel mean embedding of the joint distribution, and $\mu\left(\mathbb{P}_{X_A}\right) \otimes \mu\left(\mathbb{P}_Y\right) = \mathbb{E}\left[k(X_A, \cdot)\right] \mathbb{E}\left[k_Y(Y, \cdot)\right]$ is the kernel mean embedding of the product of marginal distributions.

We can note that, if X_A and Y are independent, $\mathrm{HSIC}(X_A, Y) = 0$. Thus, HSIC enables the identification of input variables (or sets of variables) that influence the output by measuring dependence. Moreover, HSIC can be calculated for groups of random variables indexed by A, which is useful for analyzing groups of dependent variables. An illustrated example of the HSIC measure with RKHS is given in Figure 8. The reproducibility property of RKHS allows us to derive an expression based exclusively on the expectations of the kernels, as summarized in the following proposition:

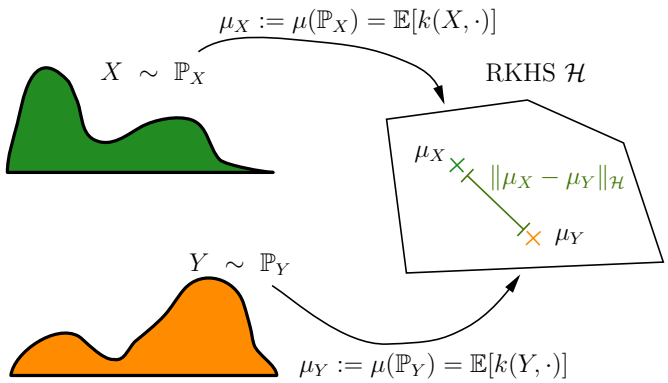

Figure 8: Illustration of the embedding of two probability distributions in the RKHS in order to compare them.

Proposition 4.1. Given an i.i.d copy (X'_A, Y') of (X_A, Y) such that $\mathbb{E}\left[k(X_A, X'_A)\right] < +\infty$ and $\mathbb{E}\left[k_Y(Y, Y')\right] < +\infty$, we have :

$$\begin{aligned}\mathrm{HSIC}(X_A, Y) = \mathbb{E}\left[k(X_A, X'_A)k_Y(Y, Y')\right] + \mathbb{E}\left[k(X_A, X'_A)\right]\mathbb{E}\left[k_Y(Y, Y')\right] \\ - 2\mathbb{E}\left[\mathbb{E}\left[k(X_A, X'_A)\,|\,X_A\right]\mathbb{E}\left[k_Y(Y, Y')\,|\,Y\right]\right]\end{aligned}$$

Using this simplified expression for the HSICs, we can easily estimate them using standard methods such as Monte Carlo. Unbiased (U-statistic) and biased (but asymptotically unbiased, V-statistic) estimators are introduced in [5, 29]. V-statistics are commonly used. To simplify notation, we will assume that $X := X_A$.

4.2 Estimation based on crude Monte Carlo

Given two i.i.d samples $(X_i, Y_i)_{1\leq 1\leq N}$ and $(X'_i, Y'_i)_{1\leq 1\leq N}$ of (X, Y), a first estimator of $\mathrm{HSIC}(X, Y)$ is given by the following V-statistic :

$$\begin{aligned}\widehat{\mathrm{HSIC}}_V(X, Y) = \frac{1}{N^2}\sum_{i,j=1}^{N} k_X(X_i, X'_j)k_Y(Y_i, Y'_j) + \frac{1}{N^4}\sum_{i,j=1}^{N} k_X(X_i, X'_j)\sum_{i,j=1}^{N} K_Y(Y_i, Y'_j) \\ -\frac{2}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} k_X(X_i, X'_j)\frac{1}{N}\sum_{j=1}^{N} k_Y(Y_i, Y'_j)\right)\end{aligned}$$

The authors in [5] introduced another estimator which is used in the following to assess crude Monte Carlo method :

Proposition 4.2.

$$\widehat{\mathrm{HSIC}}(X, Y) = \frac{1}{N^2}\mathrm{tr}\left(L_X H L H\right) \tag{4}$$

with

- L_X and L are the Gram matrices defined as

$$L_X = \left(k\left(X_i, X_j\right)\right)_{1\leq i,j\leq N} \text{ and } L = \left(k_Y\left(Y_i, Y_j\right)\right)_{1\leq i,j\leq N}$$

- $H = \left(\delta_{ij} - \frac{1}{N}\right)_{1 \leq i,j \leq N}$ where δ_{ij} is the Kronecker delta.

The main advantage of this formulation is that it requires only one i.i.d. sample of (X, Y).

4.3 Estimation based on random quantization

As shown in section 3 quantization-based LHS can be used to evaluate expectation while preserving dependency among a group of inputs. The idea is here to apply these results to the computation of HSIC. Let's define the function f such that for all $(x, x') \in \mathbb{R}^2$, $f(x, x') = k_X(x, x')k_Y\left(\mathcal{M}(x), \mathcal{M}(x')\right)$. Let $(U_i)_{i=1\ldots N}$ a sample provided by Algorithm 3 where $U_i \sim \mathcal{L}(X|X \in C_i)$, $(C_i)_{i=1,\ldots,N}$ is the Voronoi tessellation associated to the quantization of X and $p_i = \mathbb{P}\left[X \in C_i\right]$.

$$\widehat{\mathrm{HSIC}}_{RQ}(X,Y) = \sum_{i,j=1}^{N} p_i p_j f(U_i, U_j) + \sum_{i,j=1}^{N} p_i p_j k_X\left(U_i, U_j\right) \sum_{i,j=1}^{N} p_i p_j k_Y\left(\mathcal{M}\left(U_i\right), \mathcal{M}\left(U_j\right)\right) - 2\sum_{i=1}^{N} p_i \left[\left(\sum_{j=1}^{N} p_j k_X\left(U_i, U_j\right)\right)\left(\sum_{j=1}^{N} p_j k_Y\left(\mathcal{M}(U_i), \mathcal{M}\left(U_j\right)\right)\right)\right] \quad (5)$$

4.4 HSIC-based independence test

The main interest of HSIC is to identify input parameters that do not affect the output. In order to obtain a distance in the RKHS, the kernels must be characteristic, i.e. injective. Therefore, the following equivalence holds for $A \subset \{1, \ldots, d\}$:

$$X_A \perp\!\!\!\perp Y \iff \mathrm{HSIC}(X_A, Y) = 0$$

HSIC can be used to construct a statistical test of independence based on this result, introduced by [30]. The null hypothesis $\mathcal{H}_0$: "X_A and Y are independent" is equivalent to $\mathrm{HSIC}(X_A, Y) = 0$. The statistic corresponding to this test is :

$$\widehat{\mathcal{S}} = N \times \widehat{\mathrm{HSIC}}(X_A, Y)$$

The p-value represents the probability that, under the null hypothesis $\mathcal{H}_0$, the observed value $\widehat{S_{obs}} = N \times \widehat{\mathrm{HSIC}}(X_A, Y)_{obs}$ is greater than $\widehat{S}$:

$$\mathrm{p_{val}} = \mathbb{P}\left[\widehat{\mathcal{S}} \geq \widehat{\mathcal{S}}_{\mathrm{obs}} | \mathcal{H}_0\right]$$

Hence, $\mathcal{H}_0$ is rejected if $\mathrm{p_{val}} < \alpha$, where α is the first order risk of the test, i.e., the risk of falsely rejecting $\mathcal{H}_0$. In practice, $\widehat{S} \mid \mathcal{H}_0$ distribution is not known. It can be approximated asymptotically to a gamma distribution (see [30]), which requires a sample size of several hundred. Alternatively, a test based on permutations and Bootstrap can be used (see [31]).

5 Numerical experiments

In this section, we compare the Quantization-based LHS approach to Monte Carlo and LHSD on operational environmental models. The first case examines flood risk, where there is perfect knowledge of the dependency structure and the characteristics of the marginal laws. The second case studies the sizing of a grass strip in an agricultural context, where dependencies are unknown.

5.1 Case study I: sampling for a 1D hydro-dynamical model of flood risk

In this first real application, we simulate the risk of a site being flooded by a river. The study concerns an industrial site near a river, protected by a dyke. The objective is to analyze the water level relative to the dyke height to prevent flooding. The model, which is a crude simplification of the 1-D Saint-Venant equations assuming uniform and constant flow rate, is defined as follows:

$$S = Z_v + H - H_d - C_b$$
$$H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6}$$

Here, S represents the maximal overflow [m], and H is the water height [m]. The model depends on eight random variables $(Q, K_s, Z_v, Z_m, H_d, C_b, L, B)$, which are summarized in Table 1. H_d is a design parameter, while the randomness of other inputs arises from their spatio-temporal variability or estimation inaccuracies. More details on this model can be found in [32, 33].

Input	Description	Unit	Probability Distribution
Q	Maximum annual flow rate	$m^3 s^{-1}$	Truncated Gumbel $G(1013, 558)$ over $[500, 3000]$
K_s	Strickler coefficient	$m^3 s^{-1}$	Truncated Normal $\mathcal{N}(30, 8)$ over $[15, \infty)$
Z_v	Downstream river level	m	Triangular $T(49, 50, 51)$
Z_m	Upstream river level	m	Triangular $T(54, 55, 56)$
H_d	Height of the dike	m	Uniform $\mathcal{U}([7, 9])$
C_b	Bank level	m	Triangular $T(55, 55.5, 56)$
L	Length of the river section	m	Triangular $T(4990, 5000, 5010)$
B	Width of the river	m	Triangular $T(295, 300, 305)$

Table 1: Description of the model inputs.

The eight inputs of the flood problem are dependent, and a Gaussian copula is proposed for the joint distribution:

$$C(u_1, \ldots, u_d) = \Phi_{\text{joint}} \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d) \right)$$

where, Φ_{joint} is the cumulative distribution function of the multivariate Gaussian distribution with covariance matrix Σ, and Φ, the cumulative distribution function of the standard normal distribution. In the case of the flood, dependency only exists pairwise, with the following correlation coefficients: $\rho(Q, K_s) = 0.5$, $\rho(Z_v, Z_m) = \rho(L, B) = 0.3$ [33].

This study compares the results obtained with LHSD and QLHS to investigate stratified random sampling for dependent inputs with a well-known dependence structure. The results show that both LHSD and QLHS offer better performance than Monte Carlo, although μ_{QLHS} has a higher variance than LHSD (Figure 9). This was expected given that LHSD possesses analytical knowledge of the copula and the c.d.f and inverse c.d.f of the marginals. Moreover, LHSD is based on LHS and therefore inherits its properties, i.e. the more additive the function f is in the d components of X, the greater the variance reduction [13]. Since the S function is nearly additive (the sum of Sobol's first-order indices is ≈ 0.993, [34]), it can be expected that LHSD will perform better than QLHS, which lacks such properties. In environmental modeling, however, the inputs must be measured

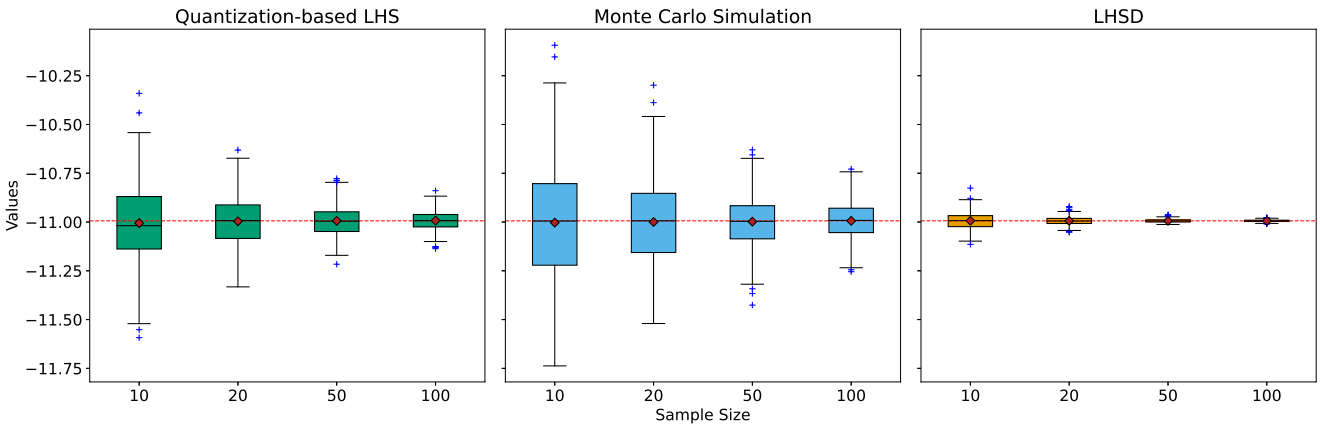

Figure 9: Estimation of $\mathbb{E}[S]$ with 500 repetitions per sample size $N \in \{10, 20, 50, 100\}$. The true value (red dotted line) is obtained using a Monte Carlo estimate with 10^8 points.

on the field or come from empirical relationships, for example. In most cases, information on the dependency structure between inputs may be completely unknown, or limited, for example coming from a random generator. In that case, the quantization-based LHS design proposed in section 3.1 is particularly adapted since it only requires the application of k-means over a large sample size. In the next section, these sampling strategies are implemented and compared on a digital twin of an agricultural catchment, without any analytical knowledge of dependency.

5.2 Case study II : sampling of Soil water retention for pesticide transfer modeling

5.2.1 Model and data description

In order to reduce the river's pollution in agricultural catchments, some best management practices consist in applying vegetative filter strips (VFSs) that reduce significantly surface runoff and erosion from the cultivated fields [35, 36]. These nature-based solutions must be designed optimally to be efficient and socially accepted, considering the local conditions of soil, climate, topography, and cultural practices. To that aim, [37, 38] developed the decision-making tool BUVARD_MES for french farmers or stakeholders in the water quality domain, based on the benchmark numerical model VFSMOD [39, 40, 41] (see figure 10). In this study case, BUVARD_MES is extended on the digital twin of the Morcille catchment (Figure 11), a vineyard agricultural place in the Beaujolais region (France), where water, sediment and pesticide are intensively measured for more than 30 years [18]. This digital twin, deeply tested and described in [42], allows simulating transfers in fields and VFSs in all possible places of the catchment, thus running on a large sample of inputs.

5.2.2 Soil water retention estimation

In the model, infiltration in presence of a water table is represented by the SWINGO algorithm [43], which depends on Van Genuchten soil hydraulic functions (VG, [8] , eq. 6).

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{(1 + (\alpha|h|)^n)^{1-1/n}} \tag{6}$$

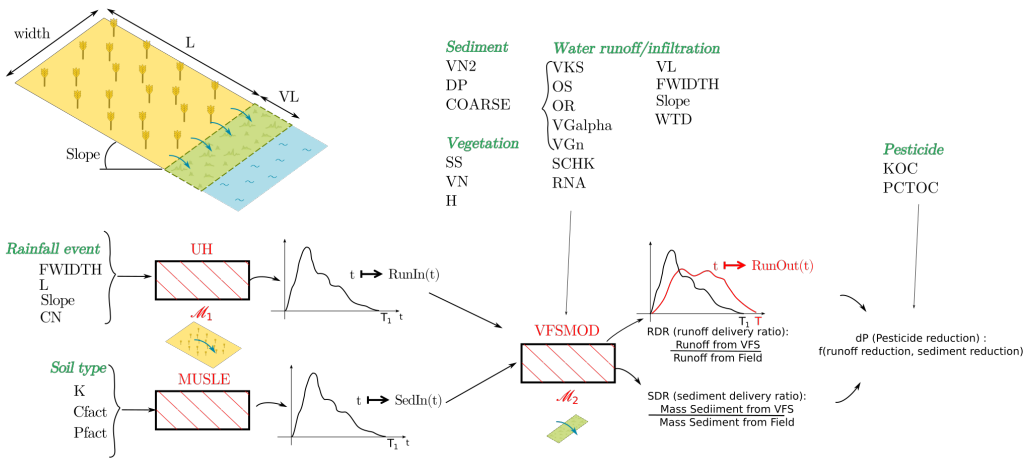

Figure 10: BUVARD_MES model and its sub-models, with inputs for climate, soil, vegetation properties of the fields and VFSs. The group of (Van Genuchten) dependent parameters is indicated by a brace.

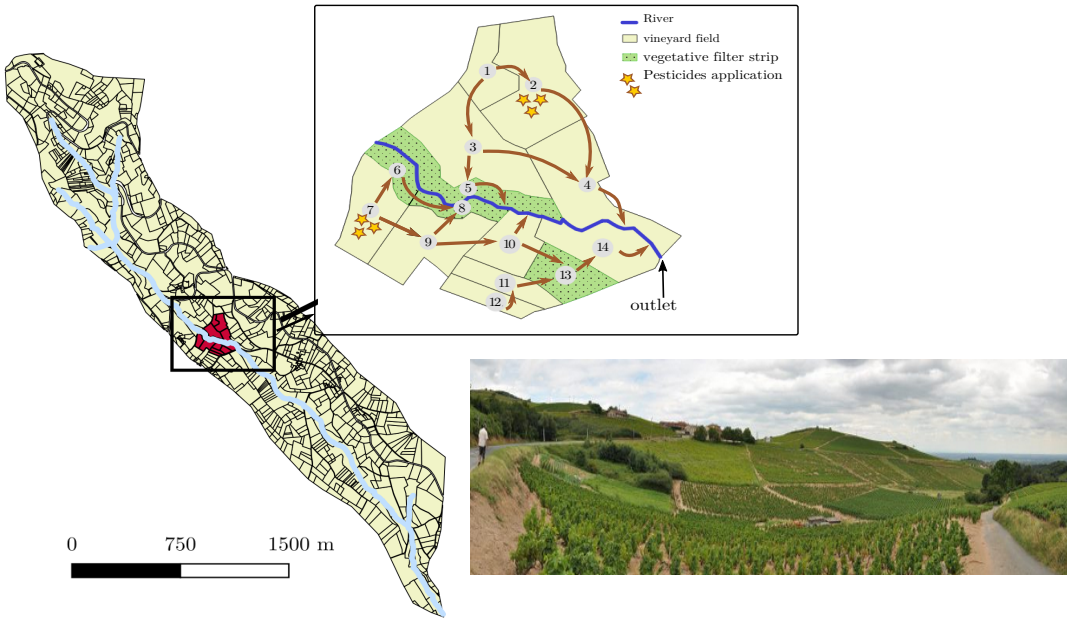

Figure 11: The Morcille catchment (left and bottom) and its digital twin, in the Beaujolais vineyard region (France). Example of surface properties extracted from the virtual catchment (top).

where θ_s is the saturated water content, θ_r is the residual water content, α is linked to the inverse of the air entry suction and n is related to the pore-size distribution.
This conductivity is described by:

$$K_v(h) = K_{\text{sat}}\sqrt{S(h)}\left(1 - \left(1 - S(h)^{\frac{1}{1-1/n}}\right)^{1-1/n}\right)^2 \tag{7}$$

where K_{sat} is the hydraulic conductivity at saturation and :

$$S(h) = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r}$$

Dependencies between the soil properties in the VG equations are known to exist, but their structure is not explicitly known, despite many studies. For example, [44] constraints the sampling by simultaneously estimating soil water characteristics and capillary length with pedotransfer functions, and [45] estimates a stochastic relation between some of the VG parameters on some specific soils. In order to account for this unknown information, two properties are considered to describe the inputs in BUVARD_MES for the sampling: the first set of inputs consider them as independent and thus random, and the second set is made of dependent variables (the Van Genuchten set, 5 parameters). For this set, a random generator was used on the data to generate the joint distribution.

Figure 12 illustrates $\theta(h)$ curves with RQ (Algorithm 3) and LHS sampling, clearly showing the values of θ_r (minimum water content) and θ_s (maximum water content), which correspond to physical values and exhibit a trend consistent with reality. The LHS curves, however, are not in agreement with observed physical behavior.

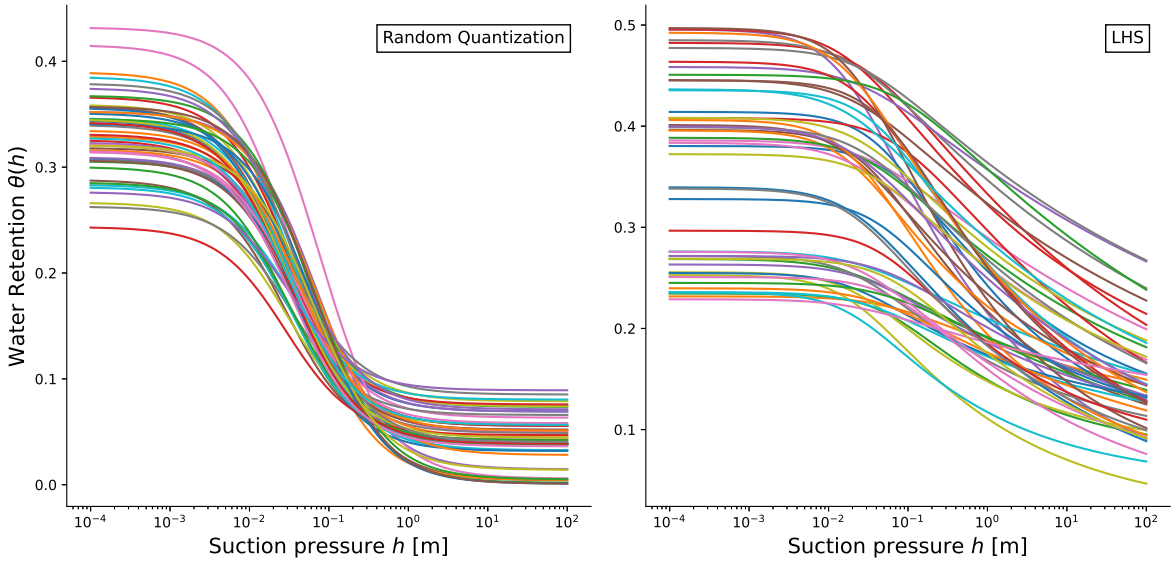

Figure 12: Water retention curves $\theta(h)$ for different Van Genuchten parameters samples based on Random Quantization (left) and LHS (right).

For the LHSD case, a Gaussian copula was fitted using maximum likelihood. As we do not have access to the quantile and cumulative distribution functions, we used their empirical versions.

Results given in Figure 13 show that the RQ estimator is unbiased with lower variance than Monte Carlo and LHSD. This confirms the relevance of this method based on quantization, considering the simplicity of implementing this approach compared to LHSD. Indeed, it only requires a simple k-means, while LHSD requires estimating a good copula and distribution function estimates, which can be a time-consuming process. Figure 14 shows another illustration on the conductivity curve. Despite the difficulty of the problem, which lies in the small value to be estimated, all three methods provide satisfactory results. μ_{RQ} outperforms both LHSD and Monte Carlo by providing a lower variance unbiased estimate.

5.2.3 Sensitivity analysis with HSIC

To reduce the complexity of the BUVARD-MES model, which is time-consuming to compute, it is necessary to reduce the dimensions in order to keep only those that influence the output. To address this issue, HSIC independence tests are performed as described in section 4. The group of dependent variables is considered as a single input, referred to as Van Genuchten. For LHSD, the same parameterization as in the previous example is maintained.

In practice, the analysis included 10 uncorrelated inputs that described the geometric properties of the contributing surface (CA) and the VFS (Area, Length, Slope, Curve Number of the field ; Slope, Width, Organic Matter, Clay content and Water table depth of the VFS ; and the pesticide property Koc, see Table 2), as well as properties related to pesticides and organic matter. Additionally, the group of 5 correlated Van Genuchten inputs were used to describe soil conductivity and water retention capacity.

The independent inputs are associated with a univariate Radial Basis Function (RBF) kernel, while the dependent group is associated with an adapted RBF kernel. To ensure consistency in the parameterization of this kernel, data were standardized, as the use of a standard deviation for the whole group is restrictive due to the 5 variables having a very variable order of magnitude (ranging from 10^{-5} to 10^{1}). As the output is scalar, a univariate RBF is used. If the p-value is less than 5%, $\mathcal{H}_0$ is rejected. Otherwise, $\mathcal{H}_0$ is accepted, i.e. the considered input is independent of the output.

$$\forall (x, x') \in \mathbb{R}^2, \quad k_{\text{indep}}(x, x') = \exp\left(-\frac{(x - x')^2}{2\theta^2}\right), \quad \theta \in \mathbb{R}$$

$$\forall (x, x') \in \left(\mathbb{R}^d\right)^2, \quad k_{\text{dep}}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\theta^2}\right), \quad \theta \in \mathbb{R}$$

The results of this independence test are summarized in Table 2. The reference values have been obtained by an asymptotic test using a gamma distribution and a Monte Carlo draw of 10,000 points. For other methods (QLHS, MC, LHSD), p-values are obtained by bootstrap [31]. The results between the reference and the proposed QLHS method are in agreement despite the small sample size. Furthermore, by using either the LHSD or Monte Carlo approach (with 400 points), we can accept the hypothesis that Van Genuchten is independent of runoff efficiency. This contradicts the reference and 'expert' knowledge.

6 Conclusion

This article proposes a new design of computer experiment, called QLHS, which naturally incorporates dependency. It is based on vector quantization, specifically k-means, ensuring ease of

Input	p-value				Decision			
	Ref	MC	QLHS	LHSD	Ref	MC	QLHS	LHSD
Area CA	0	0	0	0	✓	✓	✓	✓
Length CA	0.38	0.40	0.86	0.75	✗	✗	✗	✗
Width CA	0	0	0	0	✓	✓	✓	✓
Slope CA	0.54	0.032	0.33	0.87	✗	✓	✗	✗
Slope VFS	0	0.012	0.0018	0.0004	✓	✓	✓	✓
Width VFS	0	0	0	0	✓	✓	✓	✓
OM VFS	0.27	0.81	0.87	0.23	✗	✗	✗	✗
WTD VFS	0	0	0	0	✓	✓	✓	✓
C VFS	0.87	0.09	0.29	0.85	✗	✗	✗	✗
Koc	0.38	0.09	0.67	0.89	✗	✗	✗	✗
CN CA	0	0	0	0	✓	✓	✓	✓
Van Genuchten	0	0.18	0.031	0.096	✓	✗	✓	✗

Table 2: HSIC independence test on BUVARD-MES with 400 points per sample method. 'CA' stands for the contributive area of the VFS (the field), 'VFS' stands for vegetative filter strip. WTD is Water Table depth, C is clay content, Koc is the pesticide soil adsorption coefficient, CN is the Curve Number. The HSIC for MC and LHSD were computed with Equation 4. For QLHS, with Equation 5. ✓ : The output is dependent of the input. ✗ : The output is independent of the input.

implementation while considering groups of dependent inputs. The sampling strategy is built in an LHS way, ensuring comprehensive coverage of each marginal including groups of dependent inputs, and requires few evaluations. It allows for unbiased estimation of expectations in various configurations. The methodology has been applied to several case studies, including HSIC kernel sensitivity analysis. We show that the use of QLHS allows for high-performance sensitivity analysis with a smaller sample size compared to existing sampling approaches.

Consequently, on the basis of this methodology, and while ensuring that the dependency structure of the inputs is taken into account, a screening step can be carried out that allows the input dimension to be reduced in order to limit the calls to the computational code and to build an accurate metamodel.

Acknowledgments

This research was carried out with the support of the CIROQUO Applied Mathematics consortium, which brings together partners from both industry and academia to develop advanced methods for computational experimentation. The research is part of Water4All's AQUIGROW project, which aims to enhance the resilience of groundwater services under increased drought risk. The project has received funding from the European Union's Horizon Europe Program under Grant Agreement 101060874.

References

[1] Á. Bárkányi, T. Chován, S. Németh, and J. Abonyi. "Modelling for Digital Twins—Potential Role of Surrogate Models". In: *Processes* 9.3 (2021). ISSN: 2227-9717. DOI: 10.3390/pr9030476.

[2] B. Wang, G. Zhang, H. Wang, J. Xuan, and K. Jiao. "Multi-physics-resolved digital twin of proton exchange membrane fuel cells with a data-driven surrogate model". In: *Energy and AI* 1 (2020), p. 100004. ISSN: 2666-5468. DOI: https://doi.org/10.1016/j.egyai.2020.100004.

[3] S. Chakraborty, S. Adhikari, and R. Ganguli. "The role of surrogate models in the development of digital twins of dynamic systems". In: *Applied Mathematical Modelling* 90 (2021), pp. 662–681. ISSN: 0307-904X. DOI: https://doi.org/10.1016/j.apm.2020.09.037.

[4] A. Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis: The Primer*. en. John Wiley & Sons, 2008. ISBN: 978-0-470-72517-7.

[5] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: vol. 3734. 2005. ISBN: 978-3-540-29242-5. DOI: 10.1007/11564089_7.

[6] M. El Amri and A. Marrel. "More powerful HSIC-based independence tests, extension to space-filling designs and functional data". In: *International Journal for Uncertainty Quantification* 14.2 (2024). Publisher: Begel House Inc.

[7] N. Fellmann, C. Blanchet-Scalliet, C. Helbert, A. Spagnol, and D. Sinoquet. "Kernel-based sensitivity analysis for (excursion) sets". In: (2024). arXiv: 2305.09268 [math.ST].

[8] M. Van Genuchten. "A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1". In: *Soil Science Society of America Journal* 44 (Sept. 1980). DOI: 10.2136/sssaj1980.03615995004400050002x.

[9] M. D. McKay, R. J. Beckman, and W. J. Conover. "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code". In: *Technometrics* 21.2 (1979). Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], pp. 239–245. ISSN: 0040-1706. DOI: 10.2307/1268522.

[10] I.M Sobol'. "On the distribution of points in a cube and the approximate evaluation of integrals". In: *USSR Computational Mathematics and Mathematical Physics* 7.4 (1967), pp. 86–112. ISSN: 0041-5553. DOI: 10.1016/0041-5553(67)90144-9.

[11] E. Rouzies, C. Lauvernet, B. Sudret, and A. Vidard. "How to perform global sensitivity analysis of a catchment-scale, distributed pesticide transfer model? Application to the PESHMELBA model". In: *Geoscientific Model Development Discussions* 2021 (2021), pp. 1–44. DOI: 10.5194/gmd-2021-425.

[12] R.L. Iman and W.J. Conover. "Small sample sensitivity analysis techniques for computer models.with an application to risk assessment". In: *Communications in Statistics - Theory and Methods* 9.17 (1980). Publisher: Taylor & Francis, pp. 1749–1842. ISSN: 0361-0926. DOI: 10.1080/03610928008827996.

[13] M. Stein. "Large Sample Properties of Simulations Using Latin Hypercube Sampling". In: *Technometrics* 29.2 (1987). Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], pp. 143–151. ISSN: 00401706. DOI: 10.2307/1269769.

[14] A. Mondal and A. Mandal. "Stratified random sampling for dependent inputs in Monte Carlo simulations from computer experiments". In: *Journal of Statistical Planning and Inference* 205 (Mar. 2020), pp. 269–282. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2019.08.001.

[15] Y. Chen, M. Welling, and A. Smola. *Super-Samples from Kernel Herding*. 2012. arXiv: 1203.3472 [cs.LG].

[16] Y. Saka, M. Gunzburger, and J. Burkardt. "Latinized, improved LHS, and CVT point sets in hypercubes". In: *IEEE Transactions on Information Theory - TIT* 4 (2007).

[17] S. Corlay and G. Pagès. In: *Monte Carlo Methods and Applications* 21.1 (2015), pp. 1–32. DOI: doi:10.1515/mcma-2014-0010.

[18] V. Gouy et al. "Ardières-Morcille in the Beaujolais, France: A research catchment dedicated to study of the transport and impacts of diffuse agricultural pollution in rivers". In: *Hydrological Processes* 35.10 (2021), e14384. DOI: 10.1002/hyp.14384.

[19] A. B. Owen. "Monte Carlo variance of scrambled net quadrature". In: *SIAM Journal on Numerical Analysis* 34.5 (1997), pp. 1884–1910.

[20] A. B. Owen. "A central limit theorem for Latin hypercube sampling". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 54.2 (1992), pp. 541–551.

[21] M. Sklar. "Fonctions de répartition à N dimensions et leurs marges". In: *Annales de l'ISUP* VIII.3 (1959), pp. 229–231.

[22] L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993. ISBN: 9780130151575.

[23] N.M. Nasrabadi and R.A. King. "Image coding using vector quantization: a review". In: *IEEE Transactions on Communications* 36.8 (1988), pp. 957–971. ISSN: 1558-0857. DOI: 10.1109/26.3776.

[24] G. Pagès. *Numerical Probability: An Introduction with Applications to Finance*. Universitext. Springer International Publishing, 2018. ISBN: 9783319902760. DOI: 10.1007/978-3-319-90276-0.

[25] S. Graf and H. Luschgy. "Foundations of Quantization for Probability Distributions". In: *Lecture Notes in Mathematics -Springer-verlag-* 1730 (Jan. 2000), pp. 1–+. DOI: 10.1007/BFb0103947.

[26] S. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[27] Sébastien Da Veiga, Fabrice Gamboa, Bertrand Iooss, and Clémentine Prieur. *Basics and Trends in Sensitivity Analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2021. DOI: 10.1137/1.9781611976694.

[28] I. Steinwart and A. Christmann. "Kernels and Reproducing Kernel Hilbert Spaces". In: *Support Vector Machines*. Springer New York, 2008, pp. 110–163. ISBN: 978-0-387-77242-4. DOI: 10.1007/978-0-387-77242-4_4.

[29] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. "Supervised feature selection via dependence estimation". In: ICML '07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 823–830. ISBN: 9781595937933. DOI: 10.1145/1273496.1273600.

[30] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. "A Kernel Statistical Test of Independence". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc., 2007.

[31] M. De Lozzo and A. Marrel. "New improvements in the use of dependence measures for sensitivity analysis and screening". In: *Journal of Statistical Computation and Simulation* 86.15 (2016). Publisher: Taylor & Francis, pp. 3038–3058. ISSN: 0094-9655. DOI: 10.1080/00949655.2016.1149854.

[32] B. Iooss. "Revue sur l'analyse de sensibilité globale de modèles numériques". In: *Journal de la Societe Française de Statistique* 152.1 (2011). in french, pp. 1–23.

[33] G. Chastaing, F. Gamboa, and C. Prieur. "Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis". In: *Electronic Journal of Statistics* 6 (2012), pp. 2420–2448. DOI: 10.1214/12-EJS749.

[34] M. Lamboni, B. Iooss, A.-L. Popelin, and F. Gamboa. "Derivative-based global sensitivity measures: General links with Sobol' indices and numerical tests". In: *Mathematics and Computers in Simulation* 87 (2013), pp. 45–54. ISSN: 0378-4754. DOI: https://doi.org/10.1016/j.matcom.2013.02.002.

[35] J.-G. Lacas, M. Voltz, V. Gouy, N. Carluer, and J.-J. Gril. "Using grassed strips to limit pesticide transfer to surface water: a review". In: *Agronomy for Sustainable Development* 25.2 (2005), pp. 253–266. ISSN: 1774-0746. DOI: 10.1051/agro:2005001.

[36] S. Reichenberger, M. Bach, A. Skitschak, and H.-G. Frede. "Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness; A review". In: *Science of The Total Environment* 384.1–3 (2007), pp. 1–35. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2007.04.046.

[37] N. Carluer, C. Lauvernet, D. Noll, and R. Muñoz-Carpena. "Defining context-specific scenarios to design vegetated buffer zones that limit pesticide transfer via surface runoff". In: *Science of the Total Environment* 575 (2017), pp. 701–712. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2016.09.105.

[38] F. Veillon, N. Carluer, C. Lauvernet, and M. Rabotin. "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* (NAMUR, May 18–20, 2022). in french. 2022.

[39] R. Muñoz-Carpena, J. E. Parsons, and J. W. Gilliam. "Modeling hydrology and sediment transport in vegetative filter strips". In: *Journal of Hydrology* 214.1-4 (1999), pp. 111–129. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(98)00272-8.

[40] R. Muñoz-Carpena and J. E. Parsons. "A design procedure for vegetative filter strips using VFSMOD-W". en. In: *Transactions of the ASAE* 47.6 (2004), pp. 1933–1941. ISSN: 2151-0059. DOI: 10.13031/2013.17806.

[41] C. Lauvernet and R. Muñoz-Carpena. "Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – Part 2: model coupling, application, factor importance, and uncertainty". In: *Hydrology and Earth System Sciences* 22.1 (2018), pp. 71–87. DOI: 10.5194/hess-22-71-2018.

[42] M. Fressard, Carluer. N., and J. Pic. *PULSE : Paysage, Particules, Pesticides, Rapport final, Action n° 74 du Programme 2020 au titre de l'accord cadre Agence de l'Eau ZABR.* Tech. rep.

[43] R. Muñoz-Carpena, C. Lauvernet, and N. Carluer. "Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – Part 1: nonuniform infiltration and soil water redistribution". In: *Hydrology and Earth System Sciences* 22.1 (2018), pp. 53–70. DOI: 10.5194/hess-22-53-2018.

[44] P. Lehmann, S. Bickel, Z. Wei, and D. Or. "Physical Constraints for Improved Soil Hydraulic Parameter Estimation by Pedotransfer Functions". In: *Water Resources Research* 56.4 (2020). DOI: doi.org/10.1029/2019WR025963.

[45] C. M. Regalado and R. Muñoz-Carpena. "Estimating the saturated hydraulic conductivity in a spatially variable soil with different permeameters: a stochastic Kozeny–Carman relation". In: *Soil and Tillage Research* 77.2 (2004), pp. 189–202. ISSN: 0167-1987. DOI: 10.1016/j.still.2003.12.008.

A Proofs

A.1 Proof of Proposition 3.1

Proposition 3.1. μ_{RQ} is an unbiased estimator of m, and its variance is given by :

$$\mathrm{Var}\left(\mu_{RQ}\right) = \sum_{i=1}^{N} \mathbb{P}[X \in C_i]^2 \mathrm{Var}\left(f(U_i)\right)$$

Proof. For the bias :

$$\begin{aligned}
\mathbb{E}[f(X)] &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{1}_{X \in C_i} f(X)\right] \\
&= \sum_{i=1}^{N} \mathbb{P}[X \in C_i] \mathbb{E}\left[f(X) \,|\, X \in C_i\right] \\
&= \sum_{i=1}^{N} \mathbb{P}[X \in C_i] \mathbb{E}\left[f(U_i)\right] \\
&= \mathbb{E}\left[\mu_{RQ}\right]
\end{aligned}$$

For the variance :

$$\begin{aligned}
\mathrm{Var}\left(\mu_{RQ}\right) &= \sum_{i=1}^{N} \mathbb{P}[X \in C_i]^2 \mathrm{Var}\left(f(U_i)\right) + 2 \sum_{1 \le i < j \le N} \mathbb{P}[X \in C_i] \mathbb{P}[X \in C_j] \mathrm{Cov}\left(f(U_i), f(U_j)\right) \\
&= \sum_{i=1}^{N} \mathbb{P}[X \in C_i]^2 \mathrm{Var}\left(f(U_i)\right)
\end{aligned}$$

□

A.2 Proof of Proposition 3.2

Proposition 3.2. μ_{QLHS} is an unbiased estimator of m.

Proof. In the following, for $i = 1, \ldots, N$, we define $p_i := \mathbb{P}[X \in C_i]$. For $N \in \mathbb{N}^*$, we denote by S_N

the permutation group of order $N!$.

$$
\begin{aligned}
\mathbb{E}[\mu_{QLHS}] &= \sum_{i=1}^{N} p_i \mathbb{E}\left[f(U_i, V_{\pi(i)})\right] \\
&= \sum_{i=1}^{N} p_i \mathbb{E}\left[\mathbb{E}\left[f(U_i, V_{\pi(i)})\right] \mid \pi\right] \\
&= \sum_{i=1}^{N} p_i \sum_{a \in S_N} \mathbb{P}[\pi = a] \mathbb{E}\left[f(U_i, V_{\pi(i)}) \mid \pi = a\right] \\
&= \sum_{i=1}^{N} p_i \sum_{a \in S_N} \frac{1}{N!} \mathbb{E}\left[f(U_i, V_{a(i)})\right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{\substack{a \in S_N \\ a(i)=j}} \frac{1}{N!} p_i \mathbb{E}\left[f(U_i, V_j)\right] \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{(N-1)!}{N!} p_i \mathbb{E}\left[f(U_i, V_j)\right] \\
&= \frac{N(N-1)!}{N!} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{N} p_i \mathbb{E}[f(U_i, V_j)] \\
&= \mathbb{E}\left[f(X, Y)\right]
\end{aligned}
$$

Because $\{a \in S_N \,|\, a(i) = j\} \cong S_{N-1}$. Hence, $\text{Card}\left(\{a \in S_N \,|\, a(i) = j\}\right) = (N-1)!$. □

B Soil water retention estimation

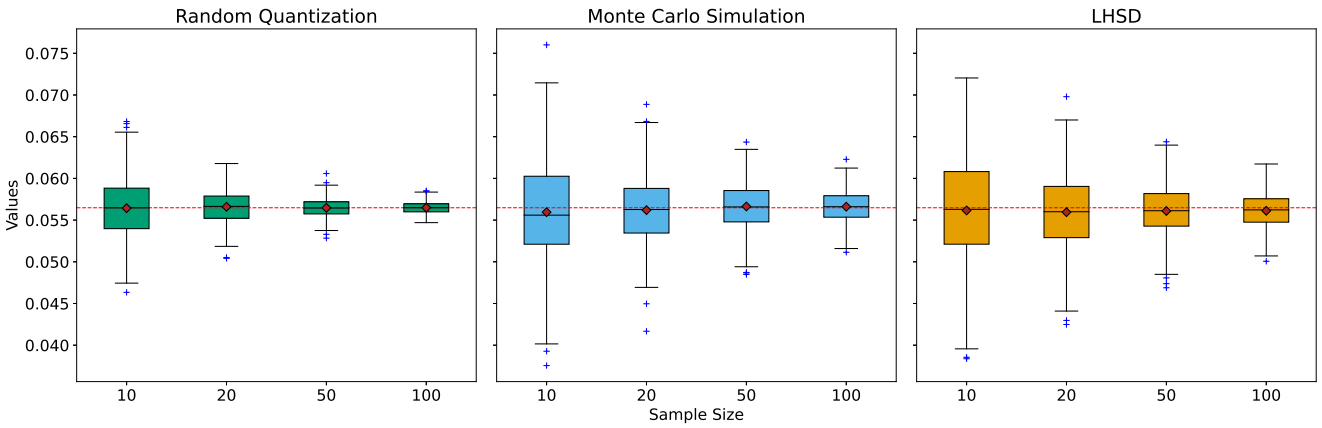

Figure 13: Comparison of RQ, Monte Carlo, and LHSD on the Water content for $h = 1$m for sample sizes $N \in \{10, 20, 50, 100\}$, with 500 replicates per sample size. The LHSD was modelled using a Gaussian copula and estimated through maximum likelihood and empirical quantile function.

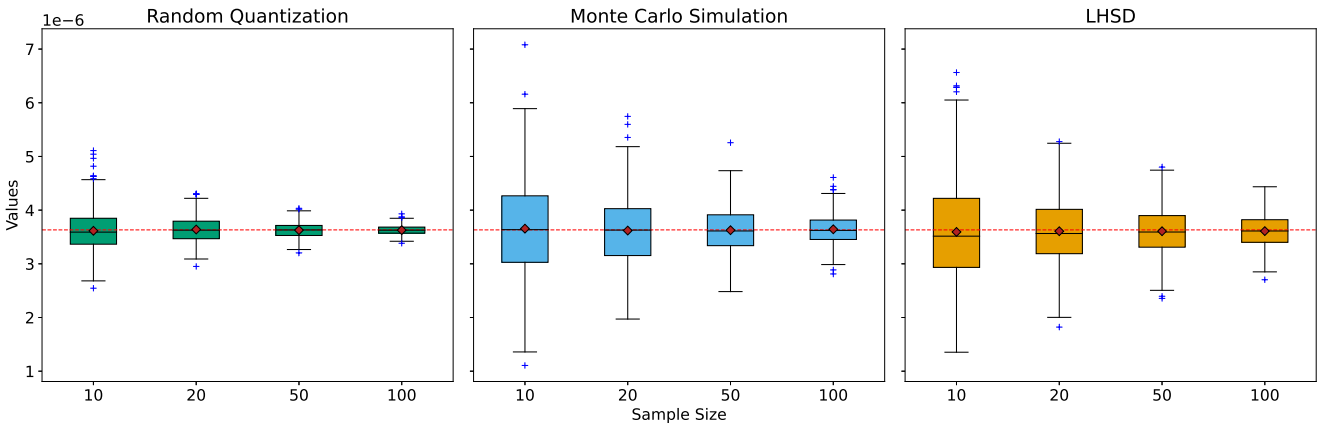

Figure 14: Comparison of RQ, Monte Carlo, and LHSD on the conductivity curve estimation for $h = 10^{-3}$ was conducted using sample sizes $N \in \{10, 20, 50, 100\}$, with 500 replicates per sample size. The LHSD was modelled using a Gaussian copula and estimated through maximum likelihood and empirical quantile function.