Computers & Geosciences 184 (2024) 105539

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

An autocorrelated conditioned Latin hypercube method for temporal or spatial sampling and predictions

Van Huong Le, Rodrigo Vargas *

Department of Plant and Soil Sciences, University of Delaware, Newark, DE, 19716, USA

ARTICLE INFO

Keywords:
Sampling design
Joint probability distribution
Temporal or spatial variogram
Geostatistical simulation
Bernstein copula
Global optimization

ABSTRACT

A data-driven method is presented for improving sampling designs from times series (1D approach) or spatial arrays (2D approach) of digital information. We present the autocorrelated conditioned Latin Hypercube Sampling (acLHS). This method combines a conditioned Latin Hypercube (cLHS) to obtain a representative sample of the joint probability distribution function and an autocorrelation model to reproduce the spatial or temporal dependency function (i.e., temporal or spatial variability). The acLHS method was tested with two case studies using data of soil CO_2 efflux (i.e., the CO_2 flux from soils to the atmosphere known as soil respiration) that are useful for carbon cycle science. First, we used data representing a time series (1D approach), and then spatial data (2D approach) across the conterminous United States (CONUS). Results show that acLHS was more efficient than other sampling methods (i.e., fixed sampling, cLHS) as it better reproduced the joint probability distribution and the temporal or spatial variability of the variable of interest. Finally, we use a Bernstein copula-based stochastic co-simulation method (BCSCS) and demonstrated that the acLHS reduces modeling prediction errors compared with other methods. The acLHS is a flexible method that can be applied to any variable of interest as a time series (1D approach) or as a spatial format (2D approach).

1. Introduction

A fundamental question for any sampling design is identifying where and when to measure. The aim of data-driven temporal or spatial (or geostatistical) sampling is to identify temporal locations in a time series, or spatial positions in an area that contributes to a representative sample of the geostatistical behaviors of the variables of interest. This implies that the obtained sample must have comparable statistical properties and similar spatial or temporal variability as the original data set or phenomena of interest. Furthermore, representative samples should also be useful to derive models and predict the variable of interest across time or space. A known challenge is that accurate temporal or spatial sampling is limited by several factors, including conceptual, logistical, technological, and physical constraints, collectively known as interoperability barriers (Vargas et al., 2017). Consequently, defining efficient methods for improving sampling designs is a crucial task. Improving sampling designs has many practical applications in environmental sciences, including modeling the spatial distribution of soil properties (Carter and Gregorich, 2006; Brus and Heuvelink, 2007; Oliver and Webster, 2015; Molla et al., 2022), optimization of environmental observatory networks (Villarreal et al., 2018; Barnett et al., 2019; Villarreal et al., 2019; Xiaojing et al., 2022), or monitoring greenhouse gas fluxes (Vickers et al., 2009; Barton et al., 2015; He et al., 2016; Vargas and Le, 2022) among many other scientific applications.

There are two main approaches for improving sampling designs: probability-based, which follows the probability distribution function, and configuration-based, following the temporal or spatial patterns. These approaches have also been referred to as design-based and model-based, respectively (Brus and De Gruijter, 1997). Probability-based sampling is essentially based on univariate and multivariate probability spaces and is focused on maximizing the reproducibility of the statistical properties (e.g., the mean, median, quantiles) of the resulting samples (McKay et al., 2000). Arguably, one of the most popular methods is the conditioned Latin Hypercube Sampling (cLHS), which presents a stratified random procedure to sample variables of interest from their multivariate distributions (Minasny and McBratney, 2006). This method can be modified to add practical constraints (e.g., travel time, terrain traversal, point clustering; (Roudier et al., 2012)), sample more at the edge of the distribution (Minasny and McBratney, 2010), or consider a high density of similar information in the sampling design (Brungard and Johnanson, 2015). The cLHS focuses on reproducing the univariate probability distribution functions and the dependency relationships between the variables from the original data to the samples. That said, cLHS is not designed to maximize

* Corresponding author.
E-mail addresses: vanle@udel.edu (V.H. Le), rvargas@udel.edu (R. Vargas).

https://doi.org/10.1016/j.cageo.2024.105539
Received 2 May 2023; Received in revised form 20 October 2023; Accepted 14 January 2024
Available online 15 January 2024

the representation of the temporal or spatial variability of the variable of interest. This limitation can influence the uncertainty of further predictions using samples derived from this approach.

Configuration-based sampling is focused on representing temporal or spatial coverage (De Gruijter et al., 2006; Walvoort et al., 2010), variability (Bogaert and Russo, 1999; Lark, 2002) and prediction (Zhu and Stein, 2006; Ma et al., 2020). For logistical reasons and simplicity, the most common approach is a fixed sampling (FS) or regular (systematically aligned) sampling (Pebesma and Bivand, 2005). This approach can be as simple as systematically aligned samples for the convenience of the experimenter, or it can be varied to sample points evenly in the areas of spherical caps using a Fibonacci lattice (González, 2010) or establish even sampling intervals in spatial strata that are constructed by k-means clustering (Walvoort et al., 2010). Configuration-based sampling focuses on reproducing the temporal or spatial variability and distribution, but it has limitations in reproducing the statistical properties and dependency relationships of the variables of interest.

Previous attempts have been made to combine probability-based with configuration-based approaches. For example, Gao et al. (2016), Wan et al. (2021) added spatial stratification (i.e., including X and Y coordinates as covariates) to combine cLHS with spatial coverage. This approach is an important advancement because the statistical properties, the dependency relationships between the variables, and the spatial coverage of the variable of interest are guaranteed. This approach does not ensure that the spatial dependence function will be reproduced. This function is essential since, from it, the variable of interest can be better interpolated and predicted across space. According to Le et al. (2020), a random variable is characterized by its univariate probability distribution function, the dependence function with other variables, and the spatial or temporal dependence function. Once we know these three functions, the variable of interest can be accurately modeled. Therefore, there is a need to propose a sampling method to combine probability-based approaches with configuration-based approaches to reproduce those three functions from the original data.

In this study, we combine probability-based (design-based) with configuration-based (model-based) approaches and propose the autocorrelated conditioned Latin Hypercube Sampling (acLHS) as a new method. The acLHS incorporates information on the spatial or temporal autocorrelation function (i.e., semivariogram) as an objective function in the optimization scheme of the commonly used cLHS method. The acLHS focus on maximizing the representativeness of variables' univariate probability distribution functions, the dependency relationships between them, and the autocorrelation function in time or space of the variable of interest. Thus, acLHS is a novel and flexible approach to improving sampling designs for time series or spatial information. In addition to Pearson's linear correlation coefficient information, we include rank correlation coefficients (i.e., Spearman and Kendall) to provide information on non-linear dependency relationships and interpretability in the resulting optimized samples. We propose that acLHS is an efficient approach to represent better the temporal or spatial distributions of the variable of interest and improve prediction estimates derived from these samples.

The proposed acLHS is applied in two case studies using temporal and spatial information of soil CO_2 efflux, which is the efflux of CO_2 from soils to the atmosphere (i.e., soil respiration) and is relevant for the global carbon cycle (Vargas et al., 2011; Phillips et al., 2017). The global soil CO_2 efflux has been estimated to be around 88 Pg/yr (Warner et al., 2019); therefore, accurate measurements to represent the temporal variability (Vargas et al., 2010) and spatial representation (Stell et al., 2021) are needed to improve local-to-global estimates. Temperature has been used as an essential variable to predict soil CO_2 efflux (Rayment and Jarvis, 2000; Pumpanen et al., 2003; Jassal et al., 2004; Curiel Yuste et al., 2010; Capooci and Vargas, 2022) and here we explore the temporal and spatial relationships between these two variables using the acLHS. The first case study represents a time series of soil CO_2 efflux and soil temperature (1D approach), and the second one represents the spatial distribution of soil CO_2 efflux and soil temperature across the conterminous United States (CONUS; 2D approach). We compared the acLHS with a fixed sampling approach (as a standard and simple configuration-based method) and using a cLHS approach (as a common probability-based method). Finally, data simulations are performed using the Bernstein copula-based temporal or spatial stochastic co-simulation method (Díaz-Viera et al., 2018; Le et al., 2020; Le, 2021; Le and Vargas, 2021, 2024) from the samples resulting from each sampling method. Our results show that acLHS is a flexible method that could be used to improve sampling designs and subsequent prediction efforts.

2. Methodology

We propose a methodology to optimize a data sampling design using an autocorrelated conditioned Latin Hypercube Sampling (acLHS). Here, we present a workflow that consists of four steps: (1) Input data, (2) Sampling design, (3) Sample-based modeling, and (4) Prediction (Fig. 1).

2.1. Input data

The input data represents two case studies: time series of soil CO_2 efflux and temperature (1D approach); and spatial information of soil CO_2 and temperature across CONUS (2D approach). Let $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be observations of the variables X and Y either as a time series or in a spatial distribution array, where X is an independent environmental variable (e.g., temperature), Y is a dependent variable such as soil CO_2 efflux (i.e., soil respiration). Variables X and Y must have the same quantity of data distributed over the same time frame (e.g., 1D array) or across the same spatial locations (i.e., 2D array). In this work, we use a 365-day full-year time series with a fixed 1-day temporal separation and a 903-point spatial distribution array across the conterminous United States (CONUS) with an equidistance of 100 km in longitudinal and latitudinal directions. We clarify that there are no specific requirements for initial observations to use this algorithm. Initial observations are completely defined by the user and the algorithm resolves the optimization based on the user's selection. This is a data-driven approach, so the initial observations influence the algorithm's output as they will influence the statistical properties and the temporal or spatial dependency that the algorithm optimizes for. Therefore, the algorithm is flexible to optimize spatial or temporal sampling, and the user decides which is the input information to inform this analytical optimization. The next step is to explore and compute their geostatistical properties, such as the univariate empirical probability distribution of X and Y; the scatterplot between X and Y, their correlation coefficients (i.e., Pearson, Spearman, and Kendall); and the empirical semivariogram of the variable of interest Y.

2.2. Sampling design

The sampling design is performed by applying different methods (i.e., fixed sampling, cLHS, and acLHS). Note that many previous studies have shown that the cLHC is more efficient for sampling than a random sampling approach (McKay et al., 2000; Minasny and McBratney, 2006; Worsham et al., 2012). Therefore, we have decided not to include the random sampling method in this work. Validation is then performed to conclude which sampling method yields a result that best represents the original data. This validation is performed by comparing the sampling results with the original data on aspects of geostatistical properties such as univariate probability distribution of X and Y, dependency relationship coefficients between X and Y (i.e., Pearson, Spearman, and Kendall), and the temporal or spatial dependence function (i.e., semivariogram) of the variable of interest Y. We used the Kolmogorov–Smirnov test to compare the univariate probability distribution and the L1 norm (i.e., the sum of absolute difference) for the correlation coefficients and empirical semivariogram.

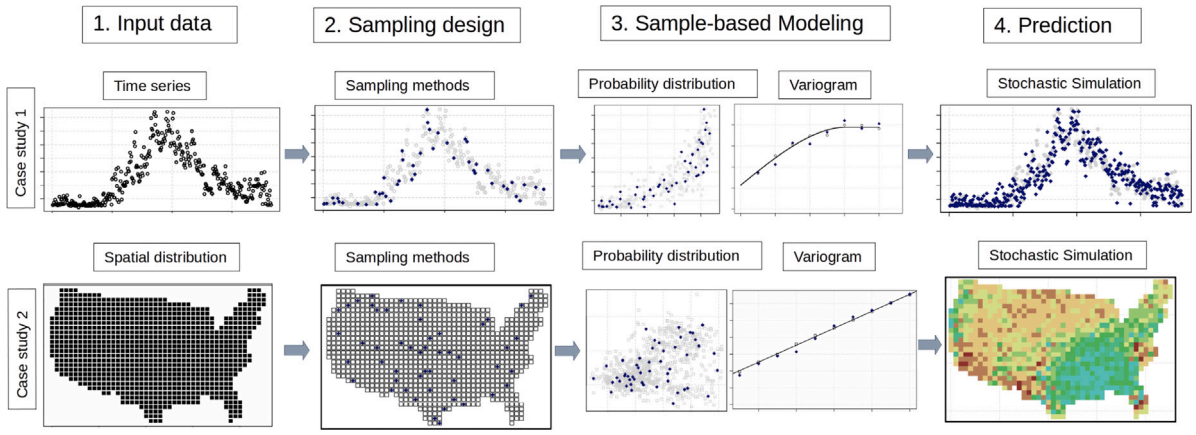

Fig. 1. Workflow of the methodology.

2.2.1. Autocorrelated conditioned Latin Hypercube Sampling method

First, we describe the autocorrelated conditioned Latin Hypercube Sampling (acLHS). In geostatistical sampling, the objective is to obtain a sample subset s of size k that is representative of the observations S; this means that s and S must have the equivalent geostatistical properties, such as the univariate statistical properties of X and Y, the dependency relationships between X and Y, and last but not least, the temporal or spatial autocorrelation of the variable of interest Y. The acLHS is an extension of the widely used cLHS (McKay et al., 2000; Minasny and McBratney, 2006), which consists of adding the rank correlation coefficients (i.e., Kendall and Spearman) and the autocorrelation function (i.e., semivariogram model) to the cLHS. The cLHS is a data-driven sampling method, and it focuses on obtaining a sample subset s that has similar statistical properties and linear dependency relationship between variables of S; however, cLHS does not emphasize nonlinear dependence relationships (i.e., Kendall and Spearman) or the temporal or spatial autocorrelation function (i.e., semivariogram) of the variable of interest Y. In contrast, the acLHS is designed to ensure that all geostatistical properties are reproducible, such as the univariate probability distribution function of X and Y, the linear and nonlinear dependency relationships between X and Y (i.e., Pearson, Kendall and Spearman), and the temporal or spatial autocorrelation function of the variable of interest Y. Here we describe step by step the acLHS method.

1. Divide the quantile distribution of X and Y into k strata, each stratum i with the same probability of $\frac{1}{k}$ and its boundary $[\tilde{Q}_X^i, \tilde{Q}_X^{i+1}]; [\tilde{Q}_Y^i, \tilde{Q}_Y^{i+1}]$, for $i = \{1, \ldots, k\}$. For each stratum or interval $[\tilde{Q}_Y^i, \tilde{Q}_Y^{i+1}]$, a sample of Y must be taken. Through this stratification strategy, an initial representative subset of the univariate probability distribution of the variable Y will be obtained. Note that this initial representative subset has its corresponding temporal or spatial coordinates. Our next step is to perturb to obtain an optimal sample from each stratum that minimizes the following functions.
2. To ensure that the univariate probability distribution of the variable X of the sample subset s and of the observations S are statistically equivalent, we must insert the objective function 1:

$$OF_1 = \sum_{i=1}^{k} |count(\tilde{Q}_X^i \le x_i < \tilde{Q}_X^{i+1}) - 1| \quad (1)$$

   To ensure a good spatial or temporal distribution of the sample points, the spatial or temporal coordinates can be entered together with the variable X.
3. An objective function 2 is added to guarantee that the dependency relationships between the variables X and Y of the sample subset s and of the observations S are similar:

$$OF_2 = |r_{XY}^s - r_{XY}^S| + |\rho_{XY}^s - \rho_{XY}^S| + |\tau_{XY}^s - \tau_{XY}^S| \quad (2)$$

   where $r_{XY}^s, \rho_{XY}^s, \tau_{XY}^s, r_{XY}^S, \rho_{XY}^S, \tau_{XY}^S$ are Pearson's linear correlation coefficient, and Spearman's and Kendall's rank correlation coefficient between X and Y of s and of S, respectively.
4. Objective function 3 is to aim that the spatial or temporal autocorrelation of Y, the sample subset s, and the observations S are equal:

$$OF_3 = \sum_{h=\Delta u}^{m\Delta u} |\gamma_Y^s(h) - \gamma_Y^S(h)| \quad (3)$$

   where $m \ge 1$ and $\gamma_Y^s(h)$, $\gamma_Y^S(h)$ are empirical semivariograms of s and of S, respectively. These are calculated as follows:

$$\gamma(h = \Delta u) = \frac{1}{2N(\Delta u)} \sum_{i=1}^{N(\Delta u)} [Y(u_i + \Delta u) - Y(u_i)]^2 \quad (4)$$

   where $N(\Delta u)$ is the number of pairs $Y(u_i + \Delta u)$ and $Y(u_i)$ that are separated by a time or distance lag Δu.
5. The sum value of the three objective functions is defined as:

$$OF = w_1 OF_1 + w_2 OF_2 + w_3 OF_3 \quad (5)$$

   where w_i are weights of each objective function. The weight of the objective functions can vary depending on the case study. We explored the sensitivity of the weights of each function to optimize the algorithm while minimizing the error. The selected weights for Case Study 1 were: $w_1 = 0.3$, $w_2 = 100$, $w_3 = 1.0$. The selected weights for Case Study 2 were: $w_1 = 10$, $w_2 = 1000$, $w_3 = 0.001$. There are multiple methods for global optimization processes including, but not limited to, ant colony optimization (Aragón-Royón et al., 2020), swarm-based optimization (Ciupke, 2016), simulated annealing (Xiang et al., 2013), and differential evolution (Storn and Price, 1997), among others. In this study, we chose differential evolution optimization because of several advantages such as simplicity, efficiency, and ease to use (Rout et al., 2013; Storn and Price, 1997). We clarify that this algorithm can be applied using other global optimization methods.
6. Iteration of steps 2–6.

Repeat steps 2 through 6 until the sum value of the objective functions (OF) reaches the error tolerance or until a maximum specified number of iterations (e.g., 1000 iterations) have been completed. We clarify that the algorithm either stops when it reaches a declared error or a maximum specified number of iterations, whatever is reached first. In this study, we declare an error tolerance of 10^{-6} and 10,000 iterations for Case Study 1 and an error tolerance of 10^{-6} and 20,000 iterations for Case Study 2. In both case studies, the optimization process does not reach error tolerance, and only reaches the maximum number of iterations. Note that this global optimization process, like

all optimization methods, always has costs and benefits (Qin et al., 2009). The cost here is the high computational time and the benefit is obtaining the representative optimized samples. Furthermore, this method is considered a data-driven sampling because the optimized samples are based on information from target and independent variables (i.e., univariate and multivariate distributions, and temporal or spatial correlation function).

At the end of this workflow, the acLHS will represent the probability space with the univariate distributions, the dependency relationships maintained, and the temporal or spatial autocorrelation of the variable of interest Y will be reproducible. This algorithm is implemented using the RGEOSTAD tools (Díaz-Viera et al., 2021) and is coded using the R software (R Core Team, 2022). Note that the proposed method "acLHS" is an extension of cLHS to obtain representative samples in the following aspects: univariate and multivariate statistics; and temporal or spatial dependence. Based on those representative samples, we can build a model to generate additional samples, as described by Iman and Conover (1980). Furthermore, in the next sections, we will also adopt Latin Hypercube sampling to generate representative realizations or simulations (conditional) from a non-Gaussian random field model using the BCSCS method, as described in Pebesma and Heuvelink (1999) and Kyriakidis and Gaganis (2013).

2.3. Sample-based modeling

Sample-based modeling is then performed using Bernstein copula-based stochastic co-simulation method (BCSCS) (Díaz-Viera et al., 2018; Le et al., 2020; Le, 2021; Vázquez-Ramírez et al., 2023) from the sample resulting from different methods mentioned above. We used the BCSCS method because of its capability to simulate the geostatistical properties of the target variables (Le et al., 2020). Therefore, we build a geostatistical model based on the samples obtained using the aforementioned sampling methods. This modeling aims to obtain a representative geostatistical model of the natural phenomenon we are investigating (e.g., soil respiration). Then, using this model, the variable of interest Y can be simulated or predicted using the variable X as an auxiliary variable. For this, the BCSCS method is applied from the samples obtained to construct the characteristic functions such as the univariate probability distribution functions of X and Y, the dependency relationships between X and Y, and the temporal or spatial dependence of the variable of interest Y. The univariate probability distribution functions of X and Y are modeled using the Bernstein polynomial approximation. The dependency relationship between X and Y is modeled using the Bernstein copula, and the temporal or spatial autocorrelation function of Y is modeled by a semivariogram function (more detail can be seen in Le et al. (2020), Le (2021)).

2.4. Prediction

Finally, we predict the temporal or spatial distribution of the target variable (i.e., CO_2 efflux) conditioned by temperature, as an essential environmental control, using inferred models based on samples from a time series (1D approach) or spatial distribution (2D approach) from the previous step. The representative geostatistical models obtained from the samples from the different sampling methods are used to predict the variable of interest Y conditioned by all the information of the variable X as an auxiliary variable. These predictions can be considered as the simulations of the variable Y, aiming to represent the geostatistical properties of the samples obtained from the different sampling methods. Finally, the prediction results of varying sampling methods are compared with the original data on their temporal or spatial distribution and the geostatistical properties. We propose that if the samples are better representative of the original population or data, their predictions should be close to the original data. We highlight that this framework is flexible and can be applied to any target variable of interest.

3. Application

3.1. Case study 1: Temporal sampling

The input data are time series of CO_2 efflux [μmol m^{-2} s^{-1}; dependent variable] and Temperature [°C; independent variable] from a temperate forest described in the previous studies (Petrakis et al., 2018; Barba et al., 2021). We used daily resolution of these variables taken during one year of measurements. The target sampling was 48 days out of 365 possible data points, equivalent to about one sample per week during a year (i.e., about 4 times per month and 12 months times 4 is 48 samples). We applied three sampling methods: (1) Fixed sampling (FS), is a sampling method with a fixed equidistant in time or x, y coordinates. In this case, data points are selected approximately every 8 days within the 365 data points. This method represents a traditional configuration-based approach. It is common practice that researchers visit a study site once a week to simplify logistical challenges (e.g., transportation, limited human resources), and aims to represent the general temporal variability within a year. (2) Traditional cLHS using the clhs package in R (Roudier, 2011) with two variables: CO_2 efflux and Temperature. This method represents a common probability-based approach focused on representing the statistical properties of CO_2 efflux and Temperature. (3) The proposed acLHS, described in the previous section, where X is Temperature and Y is CO_2 efflux. The acLHS is a method to combine probability-based and configuration-based approaches. For this case study, 10000 iterations were chosen to find an optimal sample with a computational time of about 8 min using a PC with an Intel® Core™ i7-6700HQ CPU @ 2.60 GHz × 8 processors and 16 GB of RAM. Figure S1 illustrates the optimization process of the objective function for acLHS, where the value of the objective function decreases as the number of iterations applied by the differential evolution method increases. Note that the first sampling method only takes into account the temporal configuration (i.e., time), while the last two sampling methods use the information of the target variable (i.e., soil CO_2 efflux) and the covariate (i.e., temperature).

All three sampling methods (FS, cLHS, and acLHS) selected 48 samples distributed across the year (Fig. 2). The FS method purposely targets this temporal coverage across a year (i.e., fixed systematic sampling). The other methods (i.e., cLHS, acLHS) did not predispose this temporal coverage. Still, the natural variability (i.e., univariate probability distribution functions, the dependency relationship, the temporal dependency function) of CO_2 efflux and Temperature resulted in samples distributed across the year. The scatterplots between the CO_2 efflux and Temperature derived from all three sampling methods show a apparently similar dispersion when compared to the original data (Figure S2). Still, the methods have substantial differences, as described here.

First, the samples from the acLHS method reproduce better the correlation coefficients (i.e., Pearson, Kendall, Spearman) of the original data than the samples from other sampling methods (Table S1). Second, the points of the cumulative univariate probability distributions of the samples from cLHS and acLHS are comparable to the original data. In contrast, samples from the FS sometimes deviate (Figure S3 and Figure S4). The Kolmogorov–Smirnov test shows p-values equal to 1 between the original data and samples from cLHS and acLHS, while samples from FS have p-values < 1. These results show that the cumulative probability distributions of the samples derived using cLHS or acLHS are closer to the original data distribution than those derived from FS. Third, the samples from acLHS reproduce the experimental temporal semivariogram of the original CO_2 efflux data, while the samples from cLHS do not (Fig. 3). We highlight that the experimental temporal semivariogram of the samples from FS cannot be calculated for lags < 8 days because the fixed sample resolution is 8 days. Therefore, samples from FS cannot reproduce the temporal dependency of the original data.

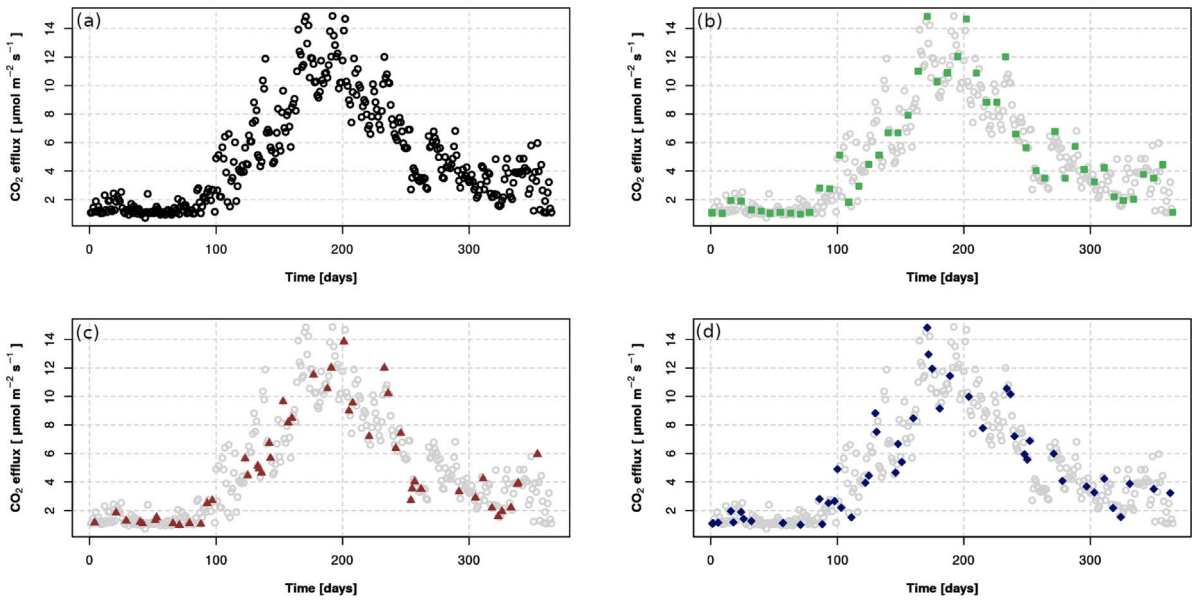

Fig. 2. Temporal distribution of: (a) CO_2 efflux data in black circles, and sampling days resulted from three sampling methods: (b) FS in green squares, (c) cLHS in red triangles, and (d) acLHS in blue diamonds.

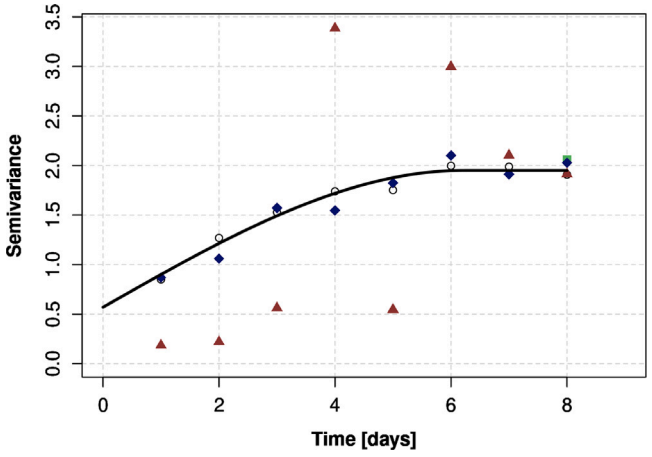

Fig. 3. The experimental temporal semivariogram for the CO_2 efflux data in black circles, and for samples resulted from three sampling methods: FS in green squares, cLHS in red triangles and acLHS in blue diamonds, and the fitted semivariogram model.

We used BCSCS for predictions of CO_2 efflux using data selected by each of the three sampling methods. Predictions of CO_2 efflux were conditioned with information from Temperature (independent variable) for calculations on 365 days (Fig. 4). We argue that a geostatistical simulation from the samples is preferred to generating realizations with similar or equivalent univariate probability distribution functions and dependency relationships and the temporal or spatial autocorrelation function of those samples (Le et al., 2020; Le, 2021).

The sum of the absolute differences between the CO_2 efflux original data and the simulations from FS, cLHS, and acLHS are 521.654, 572.232, 405.600 [µmol m^{-2} s^{-1}], respectively. This indicates that the prediction derived from the acLHS has a 29% and 22% less error than the predictions from cLHS and FS, respectively. The scatterplots between predicted CO_2 efflux and Temperature show relatively similar dispersion compared to the original data (Figure S5). Quantitatively, the predictions from the acLHS method reproduce better the correlation coefficients (i.e., Pearson, Kendall, Spearman) of the data than the simulations from the samples of cLHS and FS (Table S2). In addition, the points of the cumulative univariate probability distributions of the predictions from cLHS and acLHS are comparable to the original data points. In contrast, the predictions from the FS sometimes deviate (Figure S6). In addition, the p-values of the Kolmogorov–Smirnov test between the original CO_2 efflux data and predictions from cLHS and acLHS are equal to 0.975, 0.951 respectively, while the FS is 0.644. Finally, the predictions derived from acLHS reproduce the experimental temporal semivariogram of the original CO_2 efflux data, while the predictions from cLHS and FS do not (Fig. 5). Quantitatively, the sum of the semivariogram absolute differences between the original CO_2 efflux data and predictions from FS, cLHS, and acLHS methods are 14.653, 17.690, and 0.847, respectively.

Overall, our results show that the acLHS method is an improved approach to obtaining representative samples to better reproduces the univariate probability distribution functions of Temperature and CO_2 efflux, the dependency relationships between them, and the temporal autocorrelation function of CO_2 efflux. Consequently, predictions of CO_2 efflux conditioned by Temperature are improved when using samples derived from the acLHS method.

3.2. Case study 2: Spatial sampling

The input data are the spatial distribution of soil CO_2 efflux [g C m^{-2} $year^{-1}$; dependent variable] and Temperature [°C; independent variable] across CONUS. These data points were extracted from spatial explicit information of global soil respiration (Stell et al., 2021). We up-scaled the native resolution of the spatial information from 1 km to 100 km to obtain a dataset with 903 data points. This was done to simplify the case study as a proof-of-concept and facilitate BCSCS simulations. From 903 data points across CONUS, we tested the three methods mentioned above (i.e., FS, cLHS, acLHS) to extract 50 representative data points. We chose 50 samples because 50 samples are comparable with our selection of 48 samples for the 1D approach. This was an arbitrary number that represented a distance of about 450 km among sampling points. The data points using the FS method were selected by systematically aligned sampling (Pebesma and Bivand, 2005). Then, 50 data points across CONUS were selected using cLHS as a probability-based approach, and another 50 data points using acLHS to combine probability-based and configuration-based methods. We performed 20000 iterations to find an optimal sample with a computational time of about 18 min using a PC with an Intel® Core™ i7-8700 CPU @ 3.20 GHz × 12 processors and 32 GB of RAM. Figure S7

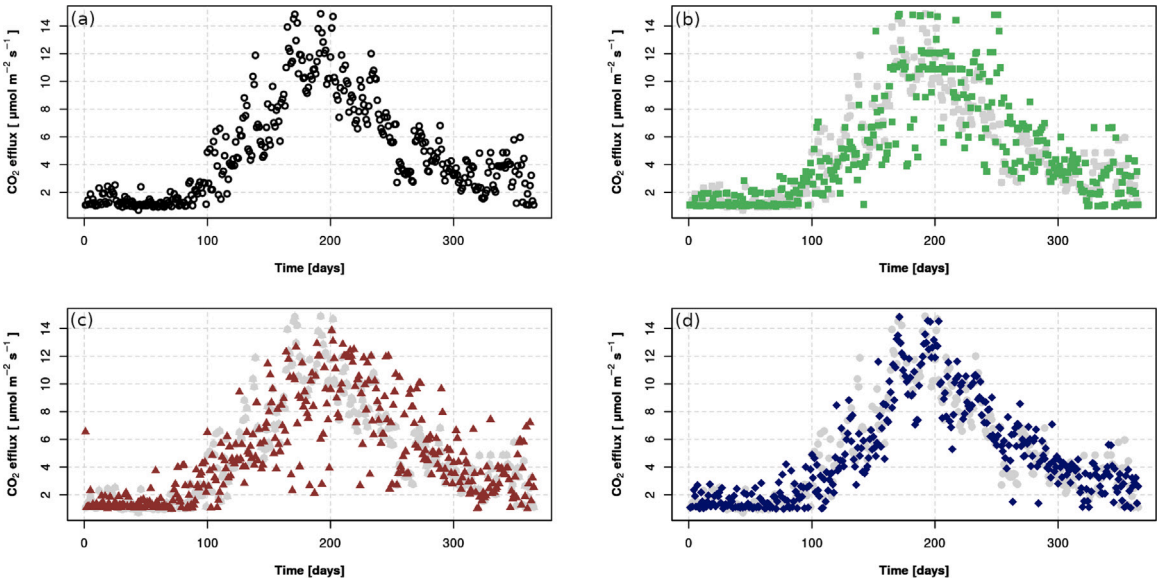

Fig. 4. Temporal distribution of: (a) CO_2 efflux data in black circles, and simulations resulted from three sampling methods: (b) FS in green squares, (c) cLHS in red triangles, and (d) acLHS in blue diamonds.

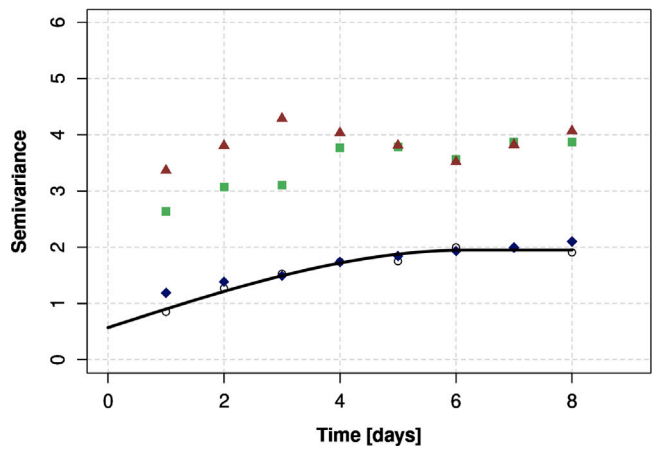

Fig. 5. The experimental temporal semivariogram (black circles) and the fitted model (black line) of the CO_2 efflux data; and the experimental temporal semivariogram of the simulations from samples the 3 sampling methods: FS in green squares, cLHS in red triangles and acLHS in blue diamonds.

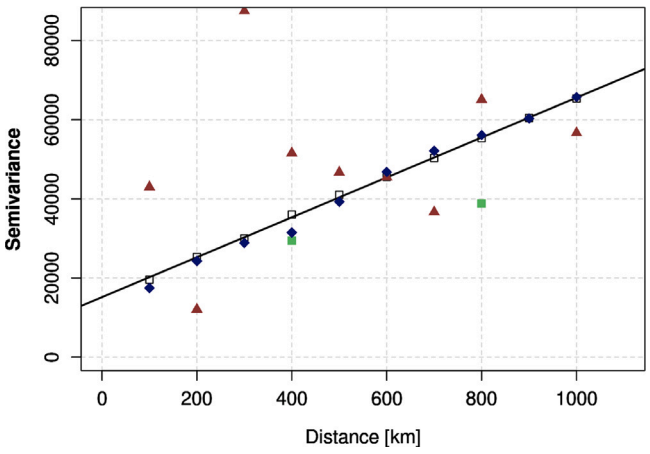

Fig. 7. The experimental spatial semivariogram (black squares) and the fitted model (black line) of the soil CO_2 efflux data; and the experimental spatial semivariogram of the resulting samples from the 3 sampling methods: FS in green squares, cLHS in red triangles, and acLHS in blue diamonds.

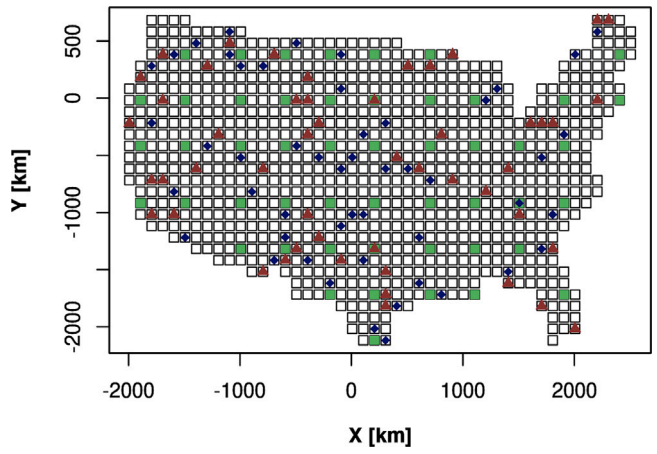

Fig. 6. The sampling results by the 3 methods: FS in green squares, cLHS in red triangles, and acLHS in blue diamonds; and data in black squares.

illustrates the optimization process of the objective function for acLHS, where the value of the objective function decreases as the number of iterations applied by the differential evolution method increases. Note that the first sampling method only takes into account the spatial configuration (i.e., x and y coordinates), while the last two sampling methods use the information of the target variable (i.e. soil CO_2 efflux) and the covariate (i.e., temperature).

The spatial distribution of the samples across CONUS reflects the predisposed conditions of each of the three methods (i.e., FS, cLHS, and acLHS; Fig. 6). The scatterplots between soil CO_2 efflux and Temperature from all three sampling methods show an apparent similar dispersion compared to the original data (Figure S8). Still, important differences are similar to those described in the first case study. First, the samples of the acLHS method reproduce better the correlation coefficients (i.e., Pearson, Kendall, Spearman) of the data than the samples from other sampling methods (Table S3). Second, the points of the cumulative univariate probability distributions of the samples from cLHS and acLHS are comparable to the original data. In contrast, samples from the FS sometimes deviate (Figure S9 and Figure S10). In

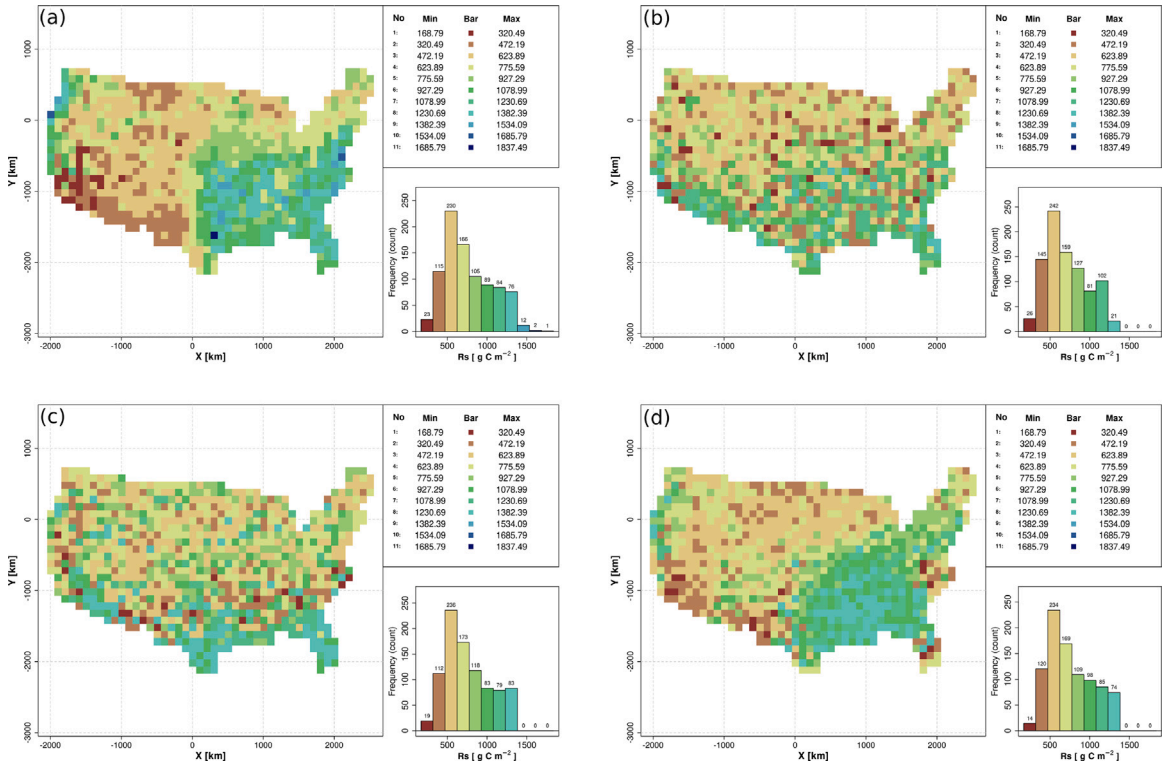

Fig. 8. The spatial distribution of soil CO_2 efflux (Rs) simulations applying the BCSCS method for the samples of the 3 sampling methods: (a) data, (b) FS, (c) cLHS, and (d) acLHS.

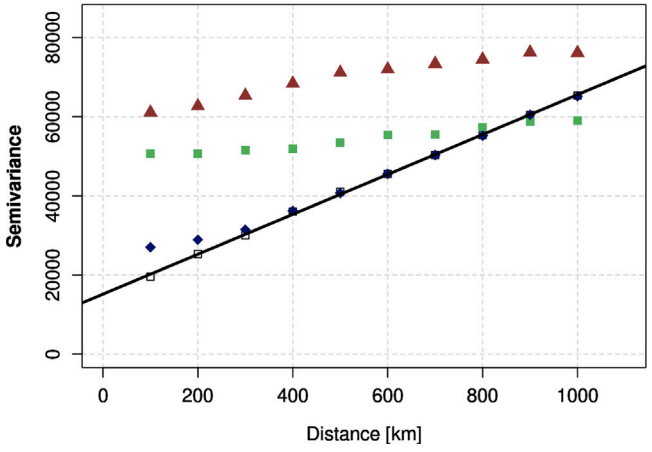

Fig. 9. The experimental spatial semivariogram and the fitted model of the soil CO_2 efflux data; and the experimental spatial semivariogram of the simulations from samples the 3 sampling methods: FS in green, cLHS in red, and acLHS in blue.

addition, the Kolmogorov–Smirnov test also shows p-values equal to 1 between the original data and samples from cLHS and acLHS, while samples from FS have p-values < 1. These results also show that the cumulative probability distributions of the samples derived using cLHS and acLHS are closer to the original data distribution than those derived from FS. Third, the cLHS samples do not reproduce the experimental spatial semivariogram of the original data, and neither can the FS samples since the sampling distance is fixed at 400 km (Fig. 7). In contrast, the acLHS sample accurately reproduces the original data's experimental spatial semivariogram (Fig. 7).

We also used BCSCS to predict spatial variability of soil CO_2 efflux conditioned with information on Temperature across CONUS using samples derived from each tested method. Our results show that the spatial patterns of soil CO_2 efflux predictions derived using the acLHS method were more similar to the original data than the predictions from FS and cLHS (Fig. 8). We calculated the sum of the soil CO_2 efflux across CONUS and found that the original data estimates a total of 6.91 [PgC year^{-1}]. Our results show that the FS method estimates with an absolute difference from the original data 2.33 [PgC year^{-1}], the cLHS 2.56 [PgC year^{-1}], and the acLHS 1.71 [PgC year^{-1}]. This indicates that the prediction derived using the acLHS reduces a 33.6% error concerning the cLHS and a 26.7% error concerning the FS. The scatterplots between predicted soil CO_2 efflux and Temperature show relatively similar dispersion compared to the original data (Figure S11). Quantitatively, the spatial predictions from the acLHS method reproduce better the correlation coefficients (i.e., Pearson, Kendall, Spearman) of the original data than the simulations from cLHS and FS (Table S4). In addition, the points of the cumulative univariate probability distributions of the simulations from cLHS and acLHS are comparable to points of the original data. In contrast, the points of the FS sometimes deviate (Figure S12). In addition, the p-values of the Kolmogorov–Smirnov test between the original soil CO_2 efflux and the predictions from the cLHS and acLHS are 0.907 and 0.968, respectively. In contrast, the *p*-value of the FS prediction is 0.002, demonstrating statistical differences from the original data. Finally, the simulations derived from the acLHS samples accurately reproduce the experimental spatial semivariogram of the original soil CO_2 data, while the simulations from FS and cLHS do not (Fig. 9). Quantitatively, the semivariogram differences between the original soil CO_2 efflux data and simulations from FS, cLHS, and acLHS are 131307.4, 271898, and 13345.71, respectively.

Overall, our results show that the acLHS method is also an improved approach to obtaining representative spatial samples to reproduce the univariate probability distribution functions of the target variables (i.e., spatial variability of soil CO_2 efflux and Temperature), the spatial relationships between them, and the spatial autocorrelation function of soil CO_2 efflux.

4. Conclusions

Where and when to measure or collect a sample is critical for environmental research. Biased sampling designs could result in flawed estimates, conclusions, and predictions, so improvements and optimization approaches are needed. Here, we introduce a novel approach to optimize sample selection that could be applied in time series (1D) and spatial arrays (2D). We propose the autocorrelated conditioned Latin hypercube sampling (acLHS) as an approach to consider the univariate probability distribution functions of the variables studied, the dependency relationships between them, and the temporal or spatial autocorrelation function of the variable of interest. This data-driven approach informs the user about how to optimally subsample information so that not all the original information is needed to recreate the statistical properties (i.e., probability distributions) and the temporal or spatial dependency. Our results show that the acLHS improves traditional approaches for representing the variable of interest and prediction estimates. We present two case studies using the information on soil CO_2 efflux, an essential variable for the global carbon budget, in a time series (1D) and across the conterminous United States (2D) to demonstrate the applicability and performance of the acLHS method. Our results show the strengths of the acLHS method and how it can be applied for sampling variables of interest relevant to environmental sciences.

In our example, we used full realizations in a time series or in space to demonstrate (i.e., proof of concept) how the algorithm works. The algorithm shows when or where to sample to reduce the number of measurements needed to represent the statistical properties and the temporal or spatial dependency of the variable of interest. The algorithm's output can be used to inform a future sampling design (either in space or time) if the statistical properties and temporal or spatial dependence are preserved. Consequently, future sampling needs fewer samples, but these will represent the statistical properties and temporal or spatial dependence of a full realization in space or time, saving time and resources for the researcher. Another application is for modeling or simulation approaches, where parameters can be calculated by using a subsample of the full realizations (in space or time) derived from our algorithm so computing costs are reduced. In signal processing, these approaches belong to the overall domain of data compression, so less information is needed to preserve the full original information. This work extends the cLHS by optimizing subsample selection, including information on statistical probabilities and temporal or spatial dependencies.

CRediT authorship contribution statement

Van Huong Le: Conceptualization, Methodology, Software, Writing – original draft, Review, Editing. **Rodrigo Vargas:** Conceptualization, Supervision, Funding acquisition, Data curation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rodrigo Vargas reports financial support was provided by National Science Foundation. Rodrigo Vargas reports financial support was provided by NASA.

Data availability

Data and code are included in the GitHub repository.

Acknowledgments

This work was supported by NASA Carbon Monitoring System grant number (80NSSC21K0964) and by the National Science Foundation, United States grant number (2103845).

Code availability section

Name of the code/library: acLHS
Contact: rvargas@udel.edu
Hardware requirements: NA.
Program language: R version 4.2.1
Software required: RStudio 2022.07.1
Program size: 4.1 MB
The source codes are available for downloading at the link: https://github.com/vargaslab/acLHS

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cageo.2024.105539.

References

Aragón-Royón, F., Jiménez-Vílchez, A., Arauzo-Azofra, A., Benitez, J.M., 2020. FSinR: an exhaustive package for feature selection. arXiv:2002.10330.

Barba, J., Poyatos, R., Capooci, M., Vargas, R., 2021. Spatiotemporal variability and origin of CO_2 and CH_4 tree stem fluxes in an upland forest. Glob. Change Biol. 27 (19), 4879–4893. http://dx.doi.org/10.1111/gcb.15783.

Barnett, D.T., Adler, P.B., Chemel, B.R., Duffy, P.A., Enquist, B.J., Grace, J.B., Harrison, S., Peet, R.K., Schimel, D.S., Stohlgren, T.J., et al., 2019. The plant diversity sampling design for the national ecological observatory network. Ecosphere 10 (2), e02603. http://dx.doi.org/10.1002/ecs2.2603.

Barton, L., Wolf, B., Rowlings, D., Scheer, C., Kiese, R., Grace, P., Stefanova, K., Butterbach-Bahl, K., 2015. Sampling frequency affects estimates of annual nitrous oxide fluxes. Sci. Rep. 5 (1), 1–9. http://dx.doi.org/10.1038/srep15912.

Bogaert, P., Russo, D., 1999. Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. Water Resour. Res. 35 (4), 1275–1289. http://dx.doi.org/10.1029/1998WR900078.

Brungard, C., Johnanson, J., 2015. The gate's locked! I can't get to the exact sampling spot... can I sample nearby? Pedometron 37, 8–10, URL http://www.pedometrics.org/Pedometron/Pedometron37.pdf.

Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80 (1–2), 1–44. http://dx.doi.org/10.1016/S0016-7061(97)00072-4.

Brus, D.J., Heuvelink, G.B., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1–2), 86–95. http://dx.doi.org/10.1016/j.geoderma.2006.10.016.

Capooci, M., Vargas, R., 2022. Diel and seasonal patterns of soil CO2 efflux in a temperate tidal marsh. Sci. Total Environ. 802, 149715. http://dx.doi.org/10.1016/j.scitotenv.2021.149715.

Carter, M.R., Gregorich, E.G., 2006. Soil Sampling and Methods of Analysis, second ed. CRC Press, http://dx.doi.org/10.1201/9781420005271.

Ciupke, K., 2016. Particle swarm optimization. R J. URL https://journal.r-project.org.

Curiel Yuste, J., Ma, S., Baldocchi, D., 2010. Plant-soil interactions and acclimation to temperature of microbial-mediated soil respiration may affect predictions of soil CO2 efflux. Biogeochemistry 98 (1), 127–138. http://dx.doi.org/10.1007/s10533-009-9381-1.

De Gruijter, J., Brus, D.J., Bierkens, M.F., Knotters, M., et al., 2006. Sampling for Natural Resource Monitoring, Vol. 665. Springer, http://dx.doi.org/10.1007/3-540-33161-1.

Díaz-Viera, M.A., Hernández-Maldonado, V., Méndez-Venegas, J., Mendoza-Torres, F., Le, V.H., Vázquez-Ramírez, D., 2021. RGEOESTAD: Un programa de código abierto para aplicaciones geoestadísticas basado en R-project. URL https://github.com/esmg-mx/RGEOSTAD.

Díaz-Viera, M.A., Le, V.H., Vázquez-Ramírez, D., 2018. A prediction of the spatial distribution of petrophysical properties with Bernstein copula using seismic attributes as secondary variables. In: InterPore2018 New Orleans. URL https://events.interpore.org/event/2/contributions/828/.

Gao, B., Pan, Y., Chen, Z., Wu, F., Ren, X., Hu, M., 2016. A spatial conditioned Latin hypercube sampling method for mapping using ancillary data. Trans. GIS 20 (5), 735–754. http://dx.doi.org/10.1111/tgis.12176.

González, Á., 2010. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. Math. Geosci. 42 (1), 49–64. http://dx.doi.org/10.1007/s11004-009-9257-x.

He, Y., Gibbons, J., Rayment, M., 2016. A two-stage sampling strategy improves chamber-based estimates of greenhouse gas fluxes. Agricult. Forest Meteorol. 228, 52–59. http://dx.doi.org/10.1016/j.agrformet.2016.06.015.

Iman, R.L., Conover, W.J., 1980. Small sample sensitivity analysis techniques for computer models with an application to risk assessment. Comm. Statist. Theory Methods 9 (17), 1749–1842.

Jassal, R., Black, T., Drewitt, G., Novak, M., Gaumont-Guay, D., Nesic, Z., 2004. A model of the production and transport of CO2 in soil: Predicting soil CO2 concentrations and CO2 efflux from a forest floor. Agricult. Forest Meteorol. 124 (3–4), 219–236. http://dx.doi.org/10.1016/j.agrformet.2004.01.013.

Kyriakidis, P., Gaganis, P., 2013. Efficient simulation of (Log)Normal random fields for hydrogeological applications. Math. Geosci. 45 (5), 531–556.

Lark, R., 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105 (1–2), 49–80. http://dx.doi.org/10.1016/S0016-7061(01)00092-1.

Le, V.H., 2021. Copula-Based Modeling for Petrophysical Property Prediction Using Seismic Attributes as Secondary Variables (Ph.D. thesis). Universidad Nacional Autónoma de México, URL http://132.248.9.195/ptd2021/marzo/0810397/Index.html.

Le, V.-H., Díaz-Viera, M.A., Vázquez-Ramírez, D., del Valle-García, R., Erdely, A., Grana, D., 2020. Bernstein copula-based spatial cosimulation for petrophysical property prediction conditioned to elastic attributes. J. Pet. Sci. Eng. 193, 107382. http://dx.doi.org/10.1016/j.petrol.2020.107382.

Le, V.-H., Vargas, R., 2021. Copula-based dependency model for CO_2 efflux prediction and its uncertainty quantification. In: Data Science Institute Symposiums. URL https://dsi.udel.edu/events/dsi-symposium-2021/posters/#session|1.

Le, V.H., Vargas, R., 2024. Beyond a deterministic representation of the temperature dependence of soil respiration. Sci. Total Environ. 912, 169391. http://dx.doi.org/10.1016/j.scitotenv.2023.169391.

Ma, T., Brus, D.J., Zhu, A.-X., Zhang, L., Scholten, T., 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. Geoderma 370, 114366. http://dx.doi.org/10.1016/j.geoderma.2020.114366.

McKay, M.D., Beckman, R.J., Conover, W.J., 2000. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42 (1), 55–61. http://dx.doi.org/10.1080/00401706.2000.10485979.

Minasny, B., McBratney, A.B., 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388. http://dx.doi.org/10.1016/j.cageo.2005.12.009.

Minasny, B., McBratney, A., 2010. Conditioned latin hypercube sampling for calibrating soil sensor data to soil properties. In: Proximal Soil Sensing. Springer, pp. 111–119. http://dx.doi.org/10.1007/978-90-481-8859-8_9.

Molla, A., Zuo, S., Zhang, W., Qiu, Y., Ren, Y., Han, J., 2022. Optimal spatial sampling design for monitoring potentially toxic elements pollution on urban green space soil: A spatial simulated annealing and k-means integrated approach. Sci. Total Environ. 802, 149728. http://dx.doi.org/10.1016/j.scitotenv.2021.149728.

Oliver, M.A., Webster, R., 2015. Basic Steps in Geostatistics: The Variogram and Kriging. Springer, http://dx.doi.org/10.1007/978-3-319-15865-5.

Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5 (2), 9–13, URL https://CRAN.R-project.org/doc/Rnews/.

Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics 41 (4), 303–312.

Petrakis, S., Barba, J., Bond-Lamberty, B., Vargas, R., 2018. Using greenhouse gas fluxes to define soil functional types. Plant Soil 423, 285–294. http://dx.doi.org/10.1007/s11104-017-3506-4.

Phillips, C.L., Bond-Lamberty, B., Desai, A.R., Lavoie, M., Risk, D., Tang, J., Todd-Brown, K., Vargas, R., 2017. The value of soil respiration measurements for interpreting and modeling terrestrial carbon cycling. Plant Soil 413 (1), 1–25. http://dx.doi.org/10.1007/s11104-016-3084-x.

Pumpanen, J., Ilvesniemi, H., Hari, P., 2003. A process-based model for predicting soil carbon dioxide efflux and concentration. Soil Sci. Am. J. 67 (2), 402–413. http://dx.doi.org/10.2136/sssaj2003.4020.

Qin, A.K., Huang, V.L., Suganthan, P.N., 2009. Differential evolution algorithm with strategy adaptation for global numerical optimization. IEEE Trans. Evol. Comput. 13 (2), 398–417. http://dx.doi.org/10.1109/TEVC.2008.927706.

R Core Team, 2022. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing. Vienna, Austria, URL https://www.R-project.org/.

Rayment, M., Jarvis, P., 2000. Temporal and spatial variation of soil CO2 efflux in a Canadian boreal forest. Soil Biol. Biochem. 32 (1), 35–45. http://dx.doi.org/10.1016/S0038-0717(99)00110-8.

Roudier, P., 2011. Clhs: A R package for conditioned Latin hypercube sampling. URL https://cran.r-project.org/web/packages/clhs/index.html.

Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned latin hypercube sampling algorithm incorporating operational constraints. In: Digital Soil Assessments and Beyond. CRC Press, London, pp. 227–231. http://dx.doi.org/10.1201/b12728.

Rout, U.K., Sahu, R.K., Panda, S., 2013. Design and analysis of differential evolution algorithm based automatic generation control for interconnected power system. Ain Shams Eng. J. 4 (3), 409–421. http://dx.doi.org/10.1016/j.asej.2012.10.010, URL https://www.sciencedirect.com/science/article/pii/S2090447912000986.

Stell, E., Warner, D., Jian, J., Bond-Lamberty, B., Vargas, R., 2021. Spatial biases of information influence global estimates of soil respiration: How can we improve global predictions? Global Change Biol. 27 (16), 3923–3938. http://dx.doi.org/10.1111/gcb.15666.

Storn, R., Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11 (4), 341–359. http://dx.doi.org/10.1023/a:1008202821328.

Vargas, R., Alcaraz-Segura, D., Birdsey, R., Brunsell, N.A., Cruz-Gaistardo, C.O., de Jong, B., Etchevers, J., Guevara, M., Hayes, D.J., Johnson, K., et al., 2017. Enhancing interoperability to facilitate implementation of REDD+: Case study of Mexico. Carbon Manag. 8 (1), 57–65. http://dx.doi.org/10.1080/17583004.2017.1285177.

Vargas, R., Carbone, M.S., Reichstein, M., Baldocchi, D.D., 2011. Frontiers and challenges in soil respiration research: From measurements to model-data integration. Biogeochemistry 102 (1), 1–13. http://dx.doi.org/10.1007/s10533-010-9462-1.

Vargas, R., Detto, M., Baldocchi, D.D., Allen, M.F., 2010. Multiscale analysis of temporal variability of soil CO2 production as influenced by weather and vegetation. Global Change Biol. 16 (5), 1589–1605. http://dx.doi.org/10.1111/j.1365-2486.2009.02111.x.

Vargas, R., Le, V.H., 2022. The paradox of assessing greenhouse gases from soils for nature-based solutions. Biogeosci. Discuss. 2022, 1–44. http://dx.doi.org/10.5194/bg-2022-153.

Vázquez-Ramírez, D., Le, V.H., Díaz-Viera, M.A., del Valle-García, R., Erdely, A., 2023. Joint stochastic simulation of petrophysical properties with elastic attributes based on parametric copula models. Geofísica Int. 62 (2), URL http://revistagi.geofisica.unam.mx/index.php/RGI/article/view/1593.

Vickers, D., Thomas, C., Law, B.E., 2009. Random and systematic CO2 flux sampling errors for tower measurements over forests in the convective boundary layer. Agric. Forest Meteorol. 149 (1), 73–83. http://dx.doi.org/10.1016/j.agrformet.2008.07.005.

Villarreal, S., Guevara, M., Alcaraz-Segura, D., Brunsell, N.A., Hayes, D., Loescher, H.W., Vargas, R., 2018. Ecosystem functional diversity and the representativeness of environmental networks across the conterminous United States. Agricult. Forest Meteorol. 262, 423–433. http://dx.doi.org/10.1016/j.agrformet.2018.07.016.

Villarreal, S., Guevara, M., Alcaraz-Segura, D., Vargas, R., 2019. Optimizing an environmental observatory network design using publicly available data. J. Geophys. Res.: Biogeosci. 124 (7), 1812–1826. http://dx.doi.org/10.1029/2018JG004714.

Walvoort, D.J., Brus, D., De Gruijter, J., 2010. An r package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36 (10), 1261–1267. http://dx.doi.org/10.1016/j.cageo.2010.04.005.

Wan, C., Kuzyakov, Y., Cheng, C., Ye, S., Gao, B., Gao, P., Ren, S., Yun, W., 2021. A soil sampling design for arable land quality observation by using SPCOSA–CLHS hybrid approach. Land Degradat. Dev. 32 (17), 4889–4906. http://dx.doi.org/10.1002/ldr.4077.

Warner, D., Bond-Lamberty, B., Jian, J., Stell, E., Vargas, R., 2019. Spatial predictions and associated uncertainty of annual soil respiration at the global scale. Glob. Biogeochem. Cycles 33 (12), 1733–1745. http://dx.doi.org/10.1029/2019GB006264.

Worsham, L., Markewitz, D., Nibbelink, N.P., West, L.T., 2012. A comparison of three field sampling methods to estimate soil carbon content. For. Sci. 58 (5), 513–522.

Xiang, Y., Gubian, S., Suomela, B., Hoeng, J., 2013. Generalized simulated annealing for efficient GlobalOptimization: The gensa package for R. R J. 5/1, URL https://journal.r-project.org.

Xiaojing, W., Honglin, H., Li, Z., Lili, F., Xiaoli, R., Weihua, L., Changxin, Z., Naifeng, L., 2022. Spatial sampling design optimization of monitoring network for terrestrial ecosystem in China. Sci. Total Environ. 157397.. http://dx.doi.org/10.1016/j.scitotenv.2022.157397.

Zhu, Z., Stein, M.L., 2006. Spatial sampling design for prediction with estimated parameters. J. Agric. Biol. Environ. Stat. 11 (1), 24–44. http://dx.doi.org/10.1198/108571106X99751.