Received: 20 November 2019 | Revised: 24 March 2020 | Accepted: 25 March 2020

DOI: 10.1111/ejss.12962

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

How to compare sampling designs for mapping?

Alexandre M.J.-C. Wadoux[1] | **Dick J. Brus**[2]

[1]Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia

[2]Biometris, Wageningen University and Research, Wageningen, The Netherlands

Correspondence
Dick J. Brus, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands.
Email: dick.brus@wur.nl

Funding information
No funding has been received to carry out this study.

Abstract

If a map is constructed through prediction with a statistical or non-statistical model, the sampling design used for selecting the sample on which the model is fitted plays a key role in the final map accuracy. Several sampling designs are available for selecting these calibration samples. Commonly, sampling designs for mapping are compared in real-world case studies by selecting just one sample for each of the sampling designs under study. In this study, we show that sampling designs for mapping are better compared on the basis of the distribution of the map quality indices over repeated selection of the calibration sample. In practice this is only feasible by subsampling a large dataset representing the population of interest, or by selecting calibration samples from a map depicting the study variable. This is illustrated with two real-world case studies. In the first case study a quantitative variable, soil organic carbon, is mapped by kriging with an external drift in France, whereas in the second case a categorical variable, land cover, is mapped by random forest in a region in France. The performance of two sampling designs for mapping are compared: simple random sampling and conditioned Latin hypercube sampling, at various sample sizes. We show that in both case studies the sampling distributions of map quality indices obtained with the two sampling design types, for a given sample size, show large variation and largely overlap. This shows that when comparing sampling designs for mapping on the basis of a single sample selected per design, there is a serious risk of an incidental result.

Highlights

- We provide a method to compare sampling designs for mapping.
- Random designs for selecting calibration samples should be compared on the basis of the sampling distribution of the map quality indices.

KEYWORDS

Kriging, machine learning, pedometrics, random forest, soil sampling, validation

1 | INTRODUCTION

In recent years, there has been an increase in digital soil mapping (DSM) activities (Arrouays, Lagacherie, & Hartemink, 2017). Digital maps of soil properties are predicted from a geostatistical model or machine learning algorithm fitted on a sample of units selected from the area to be mapped. Because the sample is the basis for mapping, its size and spatial pattern play a key role in the resulting soil map accuracy. Hereafter a sample used for fitting a statistical

model or training a machine learning algorithm is referred to as a calibration sample. Sampling designs used for selecting calibration samples that are subsequently used for mapping are referred to as sampling designs for mapping.

In the statistical and DSM literature, several solutions have been proposed to select the sampling locations for fitting or training a model and mapping. For an overview of common sampling designs for soil mapping, we redirect readers to De Gruijter, Brus, Bierkens, and Knotters (2006) and Brus (2019). The large number of available sampling designs has logically led to studies comparing the effect sampling designs have on the resulting mapping accuracy. Schmidt et al. (2014) compared three different sampling designs and their effect on the mapping accuracy of five soil properties at field scale. In this study, with each design a single sample is collected in the field and used for model fitting and prediction. Besides, to validate a map obtained with the calibration sample of a given design, the calibration samples of the other designs were used, so that no reliable conclusions can be drawn from this study. Similarly, Werbylo and Niemann (2014) evaluated stratified random and conditioned Latin hypercube sampling designs for soil moisture downscaling at three local-scale catchments. The selection of samples with the designs under study is repeated 100 times and compared using the averaged values of the Nash-Sutcliffe coefficient of efficiency between the observed and predicted downscaled patterns. The authors found mixed results, stratified random sampling being more efficient than conditioned Latin hypercube sampling for small sample sizes (fewer than 30 units) while it was the opposite for larger sample sizes.

Repeated selection of samples with a probability sampling design leads to different samples and different estimates of the population mean or total. This is also the case for commonly used sampling designs for mapping, such as spatial coverage sampling supplemented by short distance points (Lark & Marchant, 2018), feature space coverage sampling (Brus, 2019), conditioned Latin hypercube sampling (Minasny & McBratney, 2006) and model-based designs for mapping, such as the designs proposed by Van Groenigen (2000), Brus and Heuvelink (2007), Marchant and Lark (2007) or Wadoux, Marchant, and Lark (2019), among others. In all these sampling designs a random number generator is used at some stage in the selection process. In the design proposed by Lark and Marchant (2018) the points of the supplemental sample are selected randomly at a fixed but random distance from a (random) subset of the spatial coverage sample. In feature space coverage sampling, *k*-means is used to minimize a criterion. The initial clustering is chosen randomly. In conditioned Latin hypercube sampling and model-based sampling designs for mapping a criterion is minimized by simulated annealing in which proposal samples are generated by random selection of one point of the current sample and shifting it to a random selected location. The randomness in the selection of the locations of a calibration sample may have an impact on the resulting map accuracy. This has as yet been disregarded in previous studies evaluating and comparing sampling designs for mapping.

The aim of our paper is to show the importance of repeated selection of calibration samples from real-world or simulated populations when comparing sampling designs for mapping in which randomness is involved. Similar to comparing probability sampling designs for estimating the population mean on the basis of the sampling distribution of the estimated population mean, not just on the basis of the error obtained with a single probability sample, sampling designs should be compared on the basis of the distribution of map quality indices over repeated selection of samples. We illustrate this with two real-world case studies, one for a quantitative variable and one for a categorical variable. We compare simple random sampling (SRS) and conditioned Latin hypercube sampling (cLHS) at various sample sizes.

2 | THEORY AND METHODS

2.1 | Map quality indices

A wide variety of map quality indices is available (Congalton, 1991; Janssen & Heuberger, 1995; Stehman, 1997). Commonly used quality indices of continuous maps are the population means of the prediction error (ME) and of the squared prediction error (MSE), defined as:

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon(\mathbf{s}_i), \tag{1}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon(\mathbf{s}_i)^2, \tag{2}$$

where $\varepsilon(\mathbf{s}_i) = \hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i)$ is the error, in which z and $\hat{z}$ denote the measured and predicted soil variable at location $\mathbf{s}_i$, $i = 1, ..., N$, respectively, and N is the total number of population units.

The quality of categorical maps is commonly quantified by the overall accuracy (OA), defined as:

$$\mathrm{OA} = \frac{1}{N}\sum_{i=1}^{N}a(\mathbf{s}_i), \tag{3}$$

where $a(\mathbf{s}_i)$ is an indicator defined as follows:

$$a(\mathbf{s}_i) = \begin{cases} 1 & \hat{c}(\mathbf{s}_i) = c(\mathbf{s}_i) \\ 0 & \hat{c}(\mathbf{s}_i) \neq c(\mathbf{s}_i) \end{cases}, \tag{4}$$

with $\hat{c}(\mathbf{s}_i)$ the predicted class for population unit i and $c(\mathbf{s}_i)$ the true class for that unit. For infinite populations, the summation in Equations (1)–(3) is approximated by an integral.

2.2 | Evaluation of sampling designs for mapping

We propose to evaluate random designs for mapping on the basis of the sampling distribution of map quality indices: f_p(ME), f_p(MSE) and f_p(OA). The subscript p refers to a sampling design for selecting the calibration samples. Of special interest are the expectation and variance of these distributions. The expectation and variance of the MSE are defined as:

$$E_p(MSE) = \sum_{\mathcal{S}} MSE(\mathcal{S})\ p(\mathcal{S}), \tag{5}$$

$$V_p(MSE) = \sum_{\mathcal{S}} \left(MSE(\mathcal{S}) - E_p(MSE)\right)^2 p(\mathcal{S}), \tag{6}$$

with $MSE(\mathcal{S})$ being the population MSE for calibration sample $\mathcal{S}$ and $p(\mathcal{S})$ the selection probability of sample $\mathcal{S}$. By replacing MSE in Equations (5) and (6) by ME and OA we obtain the definition of the expectation and variance of the mean error and overall accuracy, respectively. Note that $E_p(MSE) = V_p(ME) + \{E_p(ME)\}^2$. Although an infinite number of samples $\mathcal{S}$ can be selected, in practice we proceed by selecting a large but finite number of samples from the population.

These are sampling distributions of the overall map quality. For mapping we propose to evaluate the sampling designs also on the basis of the sampling distributions for individual units (points), $f_p\{\varepsilon(\mathbf{s}_i)\}$, $f_p\{\varepsilon^2(\mathbf{s}_i)\}$ and $f_p\{a(\mathbf{s}_i)\}$, $i = 1, \ldots, N$. Maps of the expectation and variance of these point-wise distributions may reveal performance characteristics that remain undiscovered when looking at the distribution of overall map quality indices only. The expectation and variance of the squared error are defined as:

$$E_p\left(\varepsilon_i^2\right) = \sum_{\mathcal{S}} \varepsilon_i^2(\mathcal{S})\ p(\mathcal{S}), \tag{7}$$

$$V_p\left(\varepsilon_i^2\right) = \sum_{\mathcal{S}} \left(\varepsilon_i^2(\mathcal{S}) - E_p\left(\varepsilon_i^2\right)\right)^2 p(\mathcal{S}), \tag{8}$$

with $\varepsilon_i = \varepsilon(\mathbf{s}_i)$. The expectation and variance of the errors and accuracy indicators for individual units are obtained by replacing ε_i^2 in Equations 7 and 8 by ε_i and a_i, respectively. Note that $E_p\left(\varepsilon_i^2\right) = V_p(\varepsilon_i) + \left\{E_p(\varepsilon_i)\right\}^2$.

The sampling distributions are approximated by independent selection of a large number, say R, of calibration samples. The expectation and variance of the population mean of the quality indices and of the point-wise map quality indices are estimated by the average and variance across these R calibration samples. In the case study with the continuous variable $R = 1{,}000$, whereas $R = 400$ in the case study for the categorical variable. For the squared errors the estimators are:

$$\hat{E}_p(MSE) = \frac{1}{R}\sum_{\mathcal{S}=1}^{R} MSE(\mathcal{S}), \tag{9}$$

$$\hat{V}_p(MSE) = \frac{1}{R-1}\sum_{\mathcal{S}=1}^{R}\left(MSE(\mathcal{S}) - \frac{1}{R}\sum_{\mathcal{S}=1}^{R} MSE(\mathcal{S})\right)^2, \tag{10}$$

$$\hat{E}_p\left(\varepsilon_i^2\right) = \frac{1}{R}\sum_{\mathcal{S}=1}^{R} \varepsilon_i^2(\mathcal{S}), \tag{11}$$

$$\hat{V}_p\left(\varepsilon_i^2\right) = \frac{1}{R-1}\sum_{\mathcal{S}=1}^{R}\left(\varepsilon_i^2(\mathcal{S}) - \frac{1}{R}\sum_{\mathcal{S}=1}^{R} \varepsilon_i^2(\mathcal{S})\right)^2. \tag{12}$$

To avoid confusion we would like to stress that in this paper the indices to quantify the quality of a map are not defined across realizations of a model used for prediction (mapping), but across realizations of a sampling design used to select a calibration sample for mapping, as indicated by the subscript p in f_p, E_p and V_p. This is also the case when a statistical model is used for prediction (mapping), such as in the second case study hereafter, in which kriging with an external drift is used. This implies that, given a calibration sample, the prediction errors at points, as well as the population mean of the (squared) errors are fixed quantities, not random variables. By considering all calibration samples that can be selected by the sampling design, both the point-wise prediction errors, as well as the population mean of the errors become random quantities. The distributions of these random quantities are not model distributions, but sampling distributions, as indicated by the subscript p in f_p. In a model-based approach the statistical inference is conditioned on the calibration sample. No other samples than the one actually selected are considered. Randomness is introduced via the statistical model that is used in the inference, so that the prediction errors at points and the population mean of the (squared) errors become random variables, despite the conditioning on the calibration sample. The distributions of these random variables are model distributions, i.e., distributions defined over all possible realizations of the statistical model, which are

fundamentally different from sampling distributions, which we considered in this paper.

2.3 | Sampling designs for mapping

Two sampling designs for mapping are compared, conditioned Latin hypercube (cLHS) and simple random sampling (SRS).

Conditioned Latin hypercube sampling (cLHS, Minasny and McBratney (2006)) is an adaptation of the experimental design Latin hypercube sampling (LHS) for observational research. The adaptation is needed because not all combinations of factor levels may be represented in the population of interest. In cLHS the factor levels are marginal strata of equal size, i.e., with equal number of pixels. In total, there are n^c marginal strata, with n the total sample size and c the number of covariates. With continuous covariates only, a cLHS is selected by minimizing a weighted sum of two components. The first component is the sum over all marginal strata of the absolute difference of the marginal stratum sample size and the targeted sample size of one unit. The second component is the sum of the absolute difference of the entries of the sample correlation matrix and population correlation matrix. So in cLHS the marginal distributions of the covariates are uniformly covered, while accounting for the correlation between the covariates.

Simple random sampling is the simplest form of sampling design. It does not require any prior knowledge on the spatial variation and does not exploit environmental covariates. In SRS, each unit in the population has equal probability of being selected and the units are selected independently from each other.

3 | CASE STUDIES

3.1 | Mapping topsoil organic carbon content

We used the measurements collected over France within the framework of the European Land Use/Cover Area frame Statistical Survey (LUCAS). The database is composed of $N = 2{,}947$ georeferenced values of the topsoil (0–30 cm) organic carbon (SOC, in g kg^{-1}) as measured by an automated vario MAX CN analyzer (Elementar Analysensysteme GmbH, Germany) (Tóth, Jones, & Montanarella, 2013). The SOC values were log-transformed to correct for the highly skewed (skewness = 6.12) distribution. In this study, the $N = 2{,}947$ log-transformed SOC values are considered as our population of interest. In other words, we ignore that the LUCAS data are a sample of another area of interest (France). In addition, we collected five environmental covariates for modelling the mean (spatial trend). The covariates were either resampled using bilinear interpolation or aggregated to conform with a resolution of 1 km × 1 km, and their value was extracted to the location of the $N = 2{,}947$ LUCAS sampling locations. The covariates were the Landsat Band 3 (red) for the year 2014, the long-term averaged mean annual surface temperature (daytime) MODIS (in degrees Kelvin), the total annual precipitation (in mm/year), the elevation (in metre) and the multi-resolution index of valley bottom flatness (MRVBF) in metre × 100.

In this study, predictions of the topsoil log-transformed organic carbon were obtained by kriging with an external drift (Webster & Oliver, 2007). We fitted an exponential variogram model to the residuals of the soil property using ordinary least square in an automatic fitting and prediction procedure (Hiemstra, 2015). The linear trend was composed of the covariates described above.

For each calibration sample the MSE was computed by using the calibrated model to predict log-transformed organic carbon for all LUCAS points, including the points used for calibrating the model.

3.2 | Land cover classification

In the second case study, we used the CORINE Land Cover (CLC) inventory map updated in 2018 (Feranec, Soukup, Hazeu, & Jaffrain, 2016) as our variable of interest. The CLC map is a categorical map of 44 classes covering the whole of Europe with grid cells of 100 m × 100 m resolution. We used a subset of 39,151 km^2 of this map, covering the French region Centre-Val de Loire. In this regional area, 26 out of the 44 land cover classes are present. To speed up computation, we further selected a large sub-grid of the CLC map with a spacing of 400 m, resulting in $N = 247{,}061$ grid points. This large subsample is used as a basis to collect the calibration samples. A set of 12 environmental covariates were used as predictor in the model. The covariates were the water table depth in metre, the average soil and sedimentary-deposit thickness in metre, the Landsat Band 4 (NIR) for the year 2014, the Landsat Band 3 (red) for the year 2014, the Landsat Band 5 (SWIR) for year 2014, the Landsat Band 7 (SWIR) for year 2014, the long-term averaged mean annual surface temperature (daytime) MODIS (in Kelvin), the total annual precipitation (in mm/year), the elevation (in metre), the terrain slope in radian × 100, SAGA Wetness Index in metre × 10 and the

multiresolution index of valley bottom flatness (MRVBF) in metre × 100.

For categorical variables, predictions were made by a random forest (Breiman, 2001) model using the covariates listed above. We used the implementation provided by Wright and Ziegler (2017), and set the tuning parameters nodesize and mtry to their default values and ntree to 1,000.

4 | RESULTS

4.1 | Quality of organic carbon map

The estimated *p*-expectation of the population ME ($\hat{E}_p(ME)$) is about zero for both sampling designs SRS and cLHS and all sample sizes (Figure 1). The sampling distribution of the population ME becomes narrower around zero for larger sample sizes. There is no clear visual difference between the distributions of the ME for the two sampling designs SRS and cLHS, except for the somewhat narrower sampling distributions for cLHS with sample sizes of 50 and 200. This is confirmed by the minimum and maximum values presented in Table 1. Overall, all statistics characterizing the sampling distribution of the population ME are about equal for SRS and cLHS, for all sample sizes.

Figure 2 shows maps of the estimated *p*-expectation of the error at individual points ($\hat{E}_p(\varepsilon_i)$). The maps show that for a few points the expected values clearly differ from zero. There is no spatial correlation among the expectations of the error at points. This was tested by visual inspection of the sample variograms computed on the expectations of the error at points. Visually there is no clear difference between SRS and cLHS. This is confirmed by the summary statistics of the *p*-expectation of the errors at the $N = 2{,}947$ points provided in Table 2: all summary statistics were about equal for SRS and cLHS for all calibration sample sizes.

The median of the approximated sampling distribution of the population MSE decreases with the calibration sample size (Figure 3). The sampling distribution of the population MSE becomes narrower with increasing sample size. Visually there is no clear difference between the sampling distributions of the population MSE for SRS and cLHS. Overall, Table 3 shows that all statistics characterizing the sampling distribution of the population MSE are about equal for SRS and cLHS, for all sample sizes.

Figure 4 shows maps of the square root of the *p*-expectation of the squared error at individual points ($\sqrt{\hat{E}_p(\varepsilon_i^2)}$). Hereafter, we will shortly refer to these values as the RMSE values at points. Note that we use the square root for visualization purposes, but that the values

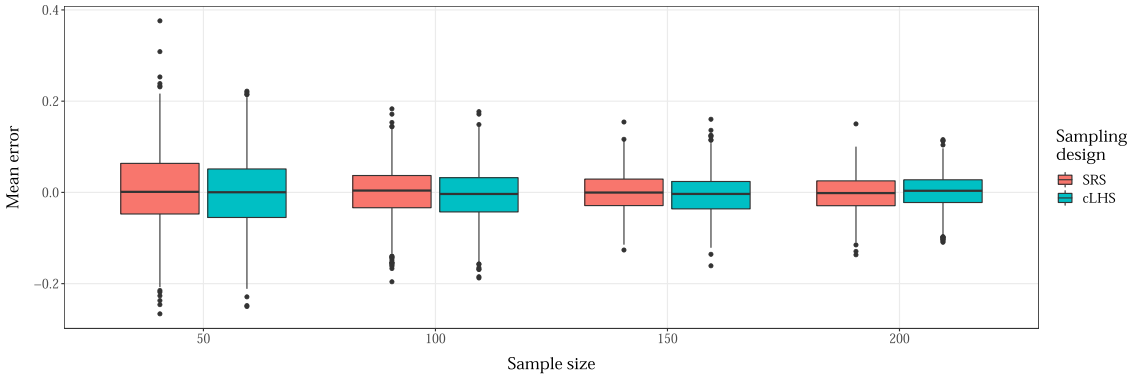

FIGURE 1 Approximated sampling distributions of the population ME, for the first case study, for SRS and cLHS and various sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Summary statistics of the approximated sampling distributions of the population ME, for the first case study, SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	−0.27	−0.25	0	0	0	0	0.08	0.08	0.38	0.22
100	−0.20	−0.25	0	0	0	−0.01	0.06	0.06	0.18	0.18
150	−0.13	−0.25	0	0	0	0	0.04	0.04	0.15	0.16
200	−0.14	−0.25	0	0	0	0	0.04	0.04	0.15	0.12

Abbreviation: cLHS, conditioned Latin hypercube sampling; ME, mean error; *Sd*, standard deviation; SRS, simple random sampling.

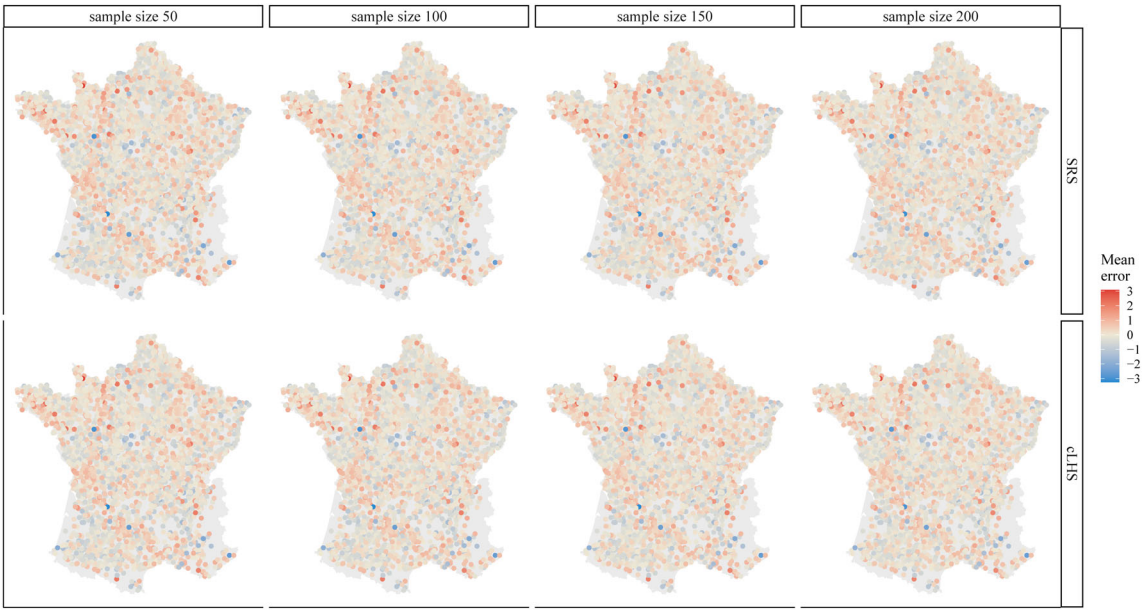

FIGURE 2 Maps of the estimated expectation of the error at individual points ($\hat{E}_p(\varepsilon_i)$) for the first case study, for SRS and cLHS and various calibration sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Summary statistics of the *p*-expectation of the prediction error at the $N = 2{,}947$ points, for the first case study, for SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	−3.24	−3.23	−0.02	−0.03	0	0	0.53	0.53	3.08	3.05
100	−3.20	−3.23	−0.03	−0.03	0	−0.01	0.51	0.51	2.98	2.99
150	−3.17	−3.23	−0.03	−0.04	0	0	0.50	0.50	2.92	2.93
200	−3.07	−3.23	−0.03	−0.03	0	0	0.49	0.49	2.85	2.93

Abbreviation: cLHS, conditioned Latin hypercube sampling; *Sd*, standard deviation; SRS, simple random sampling.

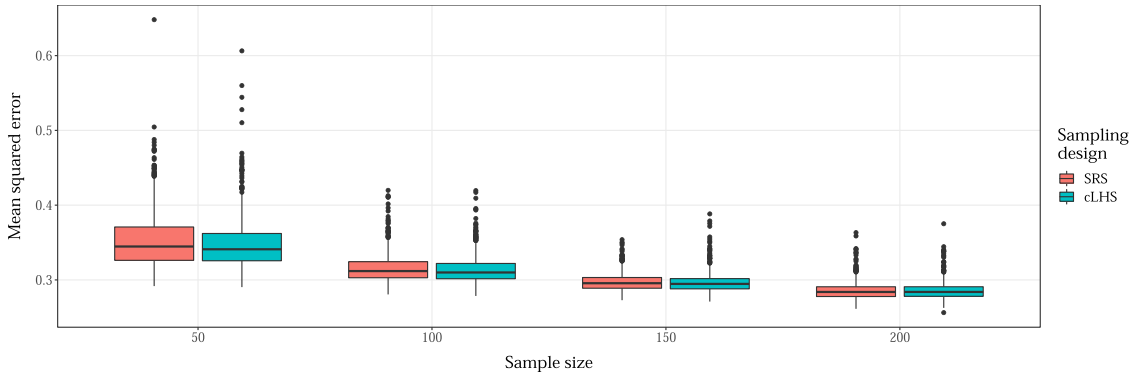

FIGURE 3 Approximated sampling distributions of the population MSE, for the first case study, for SRS and cLHS and various sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

reported in Tables 3 and 4 are not transformed. The maps show that for a few points the RMSE values are large (larger than 2 $\sqrt{\log(\mathrm{g\,kg^{-1}})}$), which means that on average over repeated calibration sampling the predictions at these points are inaccurate. There seems to be no clear spatial correlation among the RMSE values at points. Visually, there is also no difference between the RMSE values at points for SRS and cLHS. This is confirmed by

TABLE 3 Summary statistics of the approximated sampling distributions of the population MSE, for the first case study, for SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		*Sd*/Mean		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	0.29	0.29	0.34	0.34	0.35	0.35	0.04	0.03	0.10	0.09	0.65	0.61
100	0.28	0.29	0.31	0.31	0.32	0.31	0.02	0.02	0.06	0.06	0.42	0.42
150	0.27	0.29	0.30	0.29	0.30	0.30	0.01	0.01	0.04	0.04	0.35	0.39
200	0.26	0.29	0.28	0.28	0.29	0.29	0.01	0.01	0.04	0.04	0.36	0.38

Abbreviation cLHS, conditioned Latin hypercube sampling; MSE, mean squared error; *Sd*, standard deviation; SRS, simple random sampling.

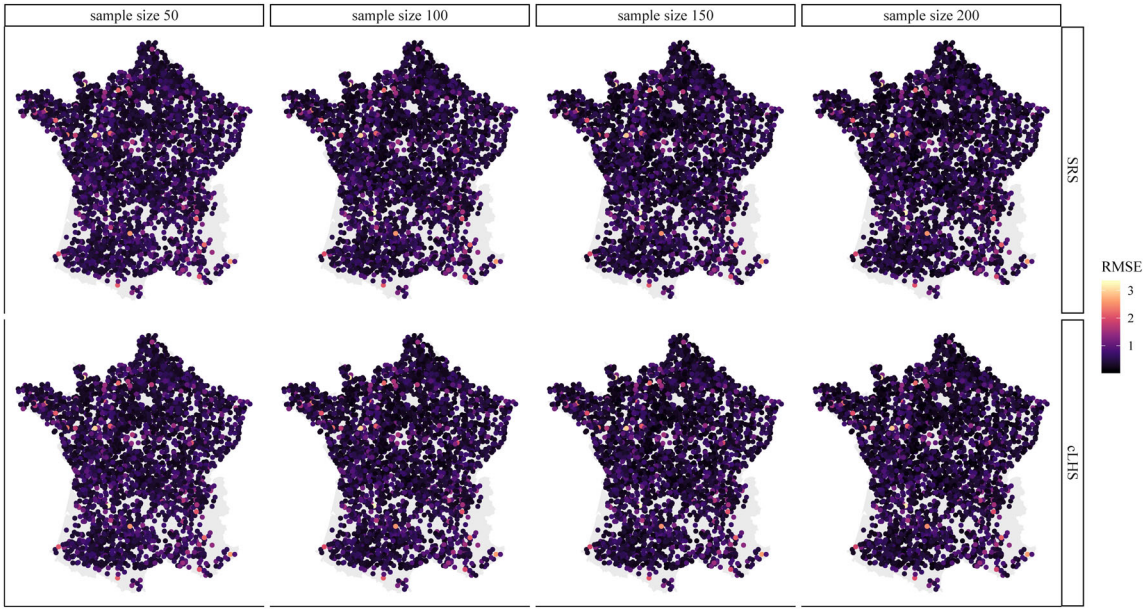

FIGURE 4 Maps of the square root of the estimated *p*-expectation of the squared error at individual points ($\hat{E}_p(\varepsilon_i^2)$) for the first case study, for SRS and cLHS and various calibration sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

Table 4: the summary statistics of the RMSE values at $N = 2{,}947$ points were about equal for SRS and cLHS, for all calibration sample sizes.

4.2 | Quality of land cover map

The median of the approximated sampling distribution of the population OA increases with the calibration sample size (Figure 5). Values of the median of the population OA for the cLHS design are slightly higher compared to SRS, for all sample sizes. Figure 5 also shows that the sampling distributions of the population OA are all comprised in a small range of values between about 0.5 and 0.75. Overall, Table 5 shows that all statistics characterizing the sampling distribution of the population OA are about equal for SRS and cLHS, for all sample sizes.

Figure 6 shows maps of the expectation of the classification indicator at individual points ($\hat{E}_p(a_i)$). Hereafter we will refer to $\hat{E}_p(a_i)$ as the mean accuracy at points. The maps show no clear visual difference in the mean accuracy at points. Minor differences between calibration sample sizes are visible in the South of the study area, where the mean accuracy at points increases with the sample size. For some of the population units, the mean accuracies are equal to zero, which means that the landcover class is never correctly classified (over repeated selection of the calibration sample and prediction). There is a clear spatial pattern in the maps. The poorly classified areas correspond to the cities (e.g. the city of Orléans) and to the main roads. All summary statistics of the mean accuracy at the $N = 247{,}061$ points are about equal for cLHS and SRS (Table 6).

TABLE 4 Summary statistics of the *p*-expectation of the squared errors at the $N = 2{,}947$ points, for the first case study, for SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		*Sd*/Mean		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	0.03	0.03	0.18	0.17	0.35	0.35	0.60	0.59	1.69	1.71	10.71	10.68
100	0.02	0.03	0.14	0.14	0.32	0.31	0.58	0.58	1.83	1.83	10.57	10.69
150	0.01	0.03	0.13	0.13	0.30	0.30	0.56	0.56	1.90	1.91	10.59	10.70
200	0.01	0.03	0.12	0.12	0.29	0.29	0.55	0.55	1.93	1.93	10.23	9.77

Abbreviation: cLHS, conditioned Latin hypercube sampling; *Sd*, standard deviation; SRS, simple random sampling.

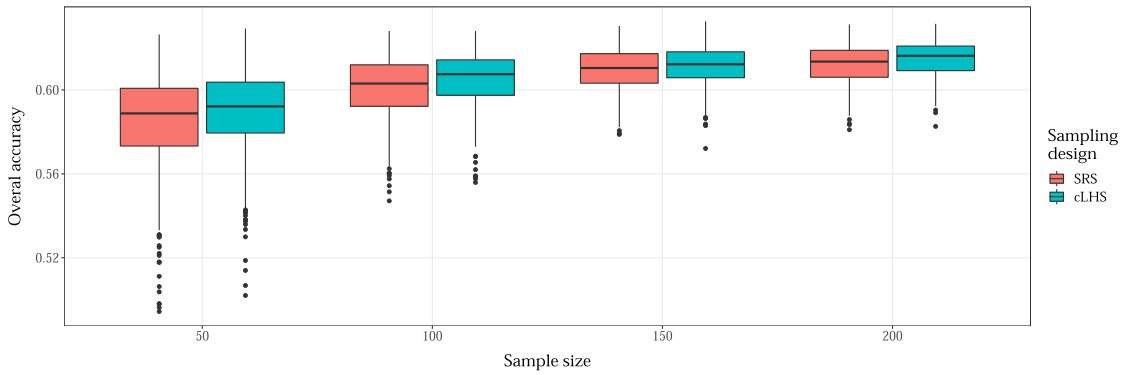

FIGURE 5 Approximated sampling distributions of the population OA, for the second case study, for SRS and cLHS and various sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Summary statistics of the approximated sampling distribution of the population OA, for the second case study, for SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		*Sd*/Mean		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	0.49	0.50	0.59	0.59	0.58	0.59	0.02	0.02	0.04	0.04	0.63	0.63
100	0.55	0.50	0.60	0.61	0.60	0.60	0.01	0.01	0.02	0.02	0.63	0.63
150	0.58	0.50	0.61	0.61	0.61	0.61	0.01	0.01	0.02	0.02	0.63	0.63
200	0.58	0.50	0.61	0.62	0.61	0.61	0.01	0.01	0.02	0.01	0.63	0.63

Abbreviation: cLHS, conditioned Latin hypercube sampling; OA, overall accuracy; *Sd*, standard deviation; SRS, simple random sampling.

5 | DISCUSSION

Our results show the importance of repeating the selection of the calibration samples when comparing sampling designs with randomness in the sample selection procedure. By repeating the selection of the calibration samples, we obtained distributions of the map quality indices. Figures 1, 3 and 5 show that this sampling distribution can be wide and that the distributions of different types of sampling designs can largely overlap for a given calibration sample size. In our case study, differences between the designs were hardly visible: both sampling designs performed about equal on average. If we select a single SRS and a single cLHS, there was about 50% probability that using an SRS for calibration resulted in a map with higher accuracy than using the cLHS for calibration, and reversely. This implies that there is a serious risk of an incidental result based on single calibration samples, showing the superiority of one of the sampling designs, whereas based on the mean accuracy over repeated calibration sampling, this is not correct.

In our case no differences can be seen between cLHS and SRS, neither in the sampling distributions of population means of (squared) prediction errors (ME, MSE and OA) nor in the sampling distributions at points. However, in other cases it may happen that the sampling

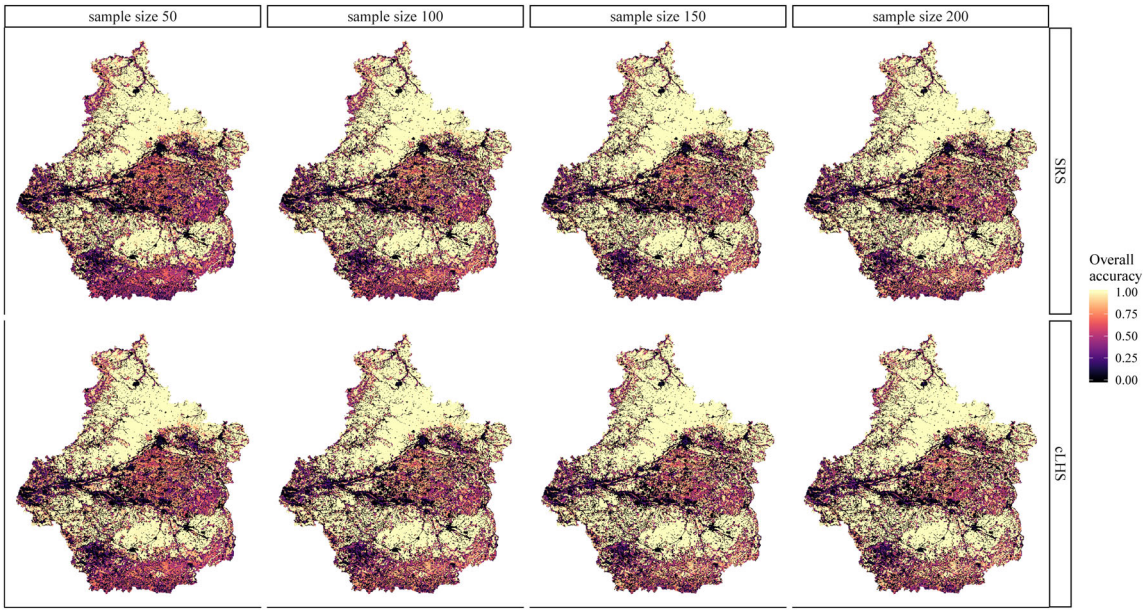

FIGURE 6 Maps of the estimated mean accuracy at points of the second case study, for both SRS and cLHS and different calibration sample sizes [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Summary statistics of the estimated mean accuracy ($\hat{E}_p(a_i)$) at the $N = 247{,}061$ points, for the second case study, for SRS and cLHS and various sample sizes

	Minimum		Median		Mean		*Sd*		*Sd*/Mean		Maximum	
	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS	SRS	cLHS
50	0	0	0.73	0.75	0.58	0.59	0.41	0.41	0.69	0.69	1	1
100	0	0	0.80	0.82	0.60	0.60	0.41	0.41	0.69	0.69	1	1
150	0	0	0.84	0.85	0.61	0.61	0.42	0.42	0.69	0.69	1	1
200	0	0	0.85	0.86	0.61	0.61	0.42	0.42	0.69	0.68	1	1

Abbreviation: cLHS, conditioned Latin hypercube sampling; *Sd*, standard deviation; SRS, simple random sampling.

distributions of overall map quality indices such as OA and MSE are about equal, but not so the distributions at points. For instance, when the *p*-expectation of the population ME is close to zero, the *p*-expectation of the error at points still can largely differ from zero as long as positive and negative errors are in balance. A sampling design with a *p*-expectation of ME close to zero and besides small values for the *p*-expectation of the point-wise errors, is to be preferred over a sampling design with larger values for the *p*-expectation of the point-wise errors.

Similarly, two designs with about equal values for the *p*-expectation of the population MSE can be quite different when looking at the spatial variation of the *p*-expectation of the squared errors at points.

In our case studies, SRS and cLHS were equivalent in terms of map accuracy. Several studies (e.g., Castro-Franco, Costa, Peralta, & Aparicio, 2015; Chu, Lin, Jang, & Chang, 2010; Contreras, Ballari, De Bruin, & Samaniego, 2019; Domenech, Castro-Franco, Costa, & Amiotti, 2017; Schmidt et al., 2014) concluded that cLHS in combination with kriging or random forest for mapping gave the most accurate prediction. These studies promote the use of cLHS as an effective sampling design for mapping. We emphasize that the results of the studies previously cited are possible outcomes (as shown by Figures 3 and 5), but, their conclusion on a sampling design performing better than another is potentially incidental because the selection of the calibration sample was not repeated. Our case studies, conversely, confirm some earlier conclusions made by Worsham, Markewitz, Nibbelink, and West (2012) and later Wadoux, Brus, and Heuvelink (2019) and Ma, Brus, Zhu, Zhang, and Scholten (2020). Worsham et al. (2012) compared SRS, stratified random sampling and cLHS for selecting calibration samples on the basis of the root mean squared

error of the mapped soil C content over a 12ha field. By repeating the selection of the calibration sample 10 times from the population, they showed that while there was a clear advantage in terms of resulting map accuracy (RMSE) for stratified random sampling and cLHS over SRS, the authors did not find an apparent improvement when using cLHS over stratified random sampling. This was further confirmed by Wadoux et al. (2019) and Ma et al. (2020) when comparing sampling designs for mapping with random forest. There is need for further research in this direction.

In practice we do not have exhaustive knowledge of the population values, so that the map accuracy obtained with a given calibration sample must be estimated from a probability sample. The validation sampling error contributes to the total variance of the map quality index. We estimated this contribution by estimating the expectation over repeated calibration sampling of the validation sampling variance of the estimated map quality index. The results (reported in the Appendix in the Supporting Information) show that the contribution of the validation sampling error to the total variance of the map quality index was large in both case studies, even with a validation sample size of 200. In most DSM studies, the validation sample size is limited, and often much smaller than the calibration sample size. One can then expect that the uncertainty about the map accuracy is large. In these cases, it is best to compute confidence intervals of the map quality indices (ME, MSE and OA) and to test whether differences in the estimated map quality indices are significant using a paired *t*-test or Wilcoxon signed-rank test. This is only feasible when the validation locations are selected by probability sampling (Brus, Kempen, & Heuvelink, 2011). Since in practice the objective is to obtain a map (not to estimate the map quality index) and it is likely that an additional sampling effort is integrated into the calibration sample rather than used for validation, we did not pursue any further in this direction.

5.1 | Various types of study to compare designs for mapping

We emphasize the need for various types of study to compare sampling designs for mapping: (a) real-world case studies, (b) studies where a very large dataset is treated as the population of interest, (c) studies in which a map of the study variable is treated as error free so that we have exhaustive knowledge of the study variable, and (d) geostatistical simulation studies.

Real-world case studies are by far the most common approaches to date. In a real-world case study, the calibration samples of the sampling designs under study are collected in a study area, used to calibrate a model, and compared based on some map quality indices. An advantage of these studies is that the data are real-world data that generally do not perfectly behave according to our probability models. An important disadvantage is that in general we cannot afford to repeat the selection of calibration samples, so our conclusion about the relative performance of sampling designs for mapping is necessarily conditioned on the two samples selected. Another disadvantage is that the map accuracy is unknown and must be estimated from a probability sample.

Similar to the real-world case studies, an advantage of studies where a large dataset is treated as the population of interest is that the data are real-world data. Another advantage is that the selection of calibration samples can be repeated, so that the sampling distribution of the map quality index can be assessed, and more general conclusions about the relative performance of sampling designs for mapping can be drawn. The main drawback is that the population is a sample of the true population of interest. The dataset must be sufficiently large to cover the characteristics of the true population. Also, the sampling fraction of the calibration sample must be very small, so that approximately errorless estimates of the map quality index can be obtained.

When using a map as reality, we either have a very large but finite population of raster cells or an infinite population (polygon map). As a consequence, selection of calibration samples can be repeated, and the map quality can be assessed from a very large validation sample, so that the computed map quality index can be treated as errorless. However, we treat the predictions as depicted on the map as errorless predictions. But actually we are comparing two predictions, one of which is treated as the ground truth. The map quality indices are only realistic estimates of the map accuracy in real-world surveys of the area depicted on the map when the quality of the map used as reality is very high. It is hard to say whether the computed map quality index over- or under-estimates the map quality with real-world surveys. Under the assumption that the systematic error in the map quality index is equal for the sampling designs under study, this type of study still may give valuable information about the relative performance of sampling designs for mapping based on the sampling distribution.

In geostatistical simulation studies, a large number of spatial populations can be generated using various models of spatial variation. With this type of study, the sampling distributions of the *model expectation* of the map quality index can be computed for the sampling designs under study. This gives insight into the relative performance of sampling designs for mapping under various models of spatial variation.

We recommend that, before a novel sampling design for selecting calibration samples is published, the performance of this novel design is compared with existing sampling designs, not only on the basis of the map accuracies obtained with a single sample per design, but preferably on the basis of the sampling distributions of the map accuracy over repeated calibration sampling. This is to avoid that a novel design is embraced by many scientists, and after many applications it appears that the novel design performed worse than existing designs.

The alternative to these empirical studies is to reason from theory that the proposed design performs better than existing designs. For instance, in an experimental design with numerous factors and many levels for each factor, Latin hypercube sampling is more efficient than a fully random design of the same size (Pebesma & Heuvelink, 1999). Minasny and McBratney (2006) therefore proposed the cLHS design for observational research, implicitly assuming that the efficiency of the experimental design is maintained when applied in observational studies, despite that the sampling design is necessarily constrained to factor level combinations that are present in the study area. They fully relied on this assumption, and in their paper they did not compare the sampling designs on the basis of the quality indices of maps obtained with these samples. Now there is growing evidence (e.g. by Worsham et al. (2012); Wadoux, Brus, and Heuvelink (2019); Ma et al. (2020) that the performance of this design is quite poor compared to other sampling designs for mapping. Studies are needed to understand how this poor performance can be explained.

6 | CONCLUSIONS

Based on the results and the discussion of these results we draw the following conclusions:

- Sampling designs for selecting calibration samples in which randomness is involved should be compared on the basis of the sampling distribution of map quality indices at the level of the population as well as the level of individual points.
- In the two case studies, with simple random sampling and conditioned Latin hypercube there was considerable variation in the map quality index over repeated sampling for all calibration sample sizes.
- When sampling designs for mapping are compared on the basis of one sample per design, the difference in the map quality index between the two sampling designs for mapping may largely deviate from the difference in the expectation of the map quality index over repeated sampling.
- In both case studies there was no benefit in using conditioned Latin hypercube sampling over simple random sampling for mapping.
- We recommend to compare sampling designs for mapping based on a combination of (a) real-world case studies, (b) studies in which calibration samples are repeatedly selected from a very large sample representing the population and/or from a map, and (c) simulation studies in which populations are generated with various models of spatial variation.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed.

ORCID

Alexandre M.J.-C. Wadoux https://orcid.org/0000-0001-7325-9716
Dick J. Brus https://orcid.org/0000-0003-2194-4783

REFERENCES

Arrouays, D., Lagacherie, P., & Hartemink, A. E. (2017). Digital soil mapping across the globe. *Geoderma Regional*, *9*, 1–4.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brus, D. J. (2019). Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma*, *338*, 464–480.

Brus, D. J., & Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, *138*, 86–95.

Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, *62*, 394–407.

Castro-Franco, M., Costa, J. L., Peralta, N., & Aparicio, V. (2015). Prediction of soil properties at farm scale using a model-based soil sampling scheme and random forest. *Soil Science*, *180*, 74–85.

Chu, H.-J., Lin, Y.-P., Jang, C.-S., & Chang, T.-K. (2010). Delineating the hazard zone of multiple soil pollutants by multivariate indicator kriging and conditioned Latin hypercube sampling. *Geoderma*, *158*, 242–251.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*, 35–46.

Contreras, J., Ballari, D., De Bruin, S., & Samaniego, E. (2019). Rainfall monitoring network design using conditioned Latin hypercube sampling and satellite precipitation estimates: An application in the ungauged Ecuadorian Amazon. *International Journal of Climatology*, *39*, 2209–2226.

De Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., & Knotters, M. (2006). *Sampling for natural resource monitoring*. Dordrecht, NL: Springer Science & Business Media.

Domenech, M. B., Castro-Franco, M., Costa, J. L., & Amiotti, N. M. (2017). Sampling scheme optimization to map soil depth to petrocalcic horizon at field scale. *Geoderma*, *290*, 75–82.

Feranec, J., Soukup, T., Hazeu, G., & Jaffrain, G. (2016). *European landscape dynamics: CORINE land cover data*. Boca Raton: CRC Press.

Hiemstra, P. (2015). *Package "automap"*. R package version 1.0–14. Retrieved from https://CRAN.R-project.org/package=automap

Janssen, P. H. M., & Heuberger, P. S. C. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*, 55–66.

Lark, R. M., & Marchant, B. P. (2018). How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters? *Geoderma*, *319*, 89–99.

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

Marchant, B. P., & Lark, R. M. (2007). Optimized sample schemes for geostatistical surveys. *Mathematical Geology*, *39*, 113–134.

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*, 1378–1388.

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

Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., & Scholten, T. (2014). A comparison of calibration sampling schemes at the field scale. *Geoderma*, *232*, 243–256.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*, 77–89.

Tóth, G., Jones, A. and Montanarella, L. (2013). *Lucas topsoil survey: Methodology, data and results* (JRC Technical Report). Luxembourg: Publications Office of the European Union.

Van Groenigen, J. W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, *97*, 223–236.

Wadoux, A. M. J.-C., Brus, D. J., & Heuvelink, G. B. M. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, *355*, 113913.

Wadoux, A. M. J.-C., Marchant, B. P., & Lark, R. M. (2019). Efficient sampling for geostatistical surveys. *European Journal of Soil Science*, *70*, 975–989.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists*. Chichester, UK: John Wiley & Sons.

Werbylo, K. L., & Niemann, J. D. (2014). Evaluation of sampling techniques to characterize topographically-dependent variability for soil moisture downscaling. *Journal of Hydrology*, *516*, 304–316.

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

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1–17.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wadoux AMJ-C, Brus DJ. How to compare sampling designs for mapping? *Eur J Soil Sci*. 2021;72:35–46. https://doi.org/10.1111/ejss.12962