Contents lists available at ScienceDirect

Geoderma

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

Sampling design optimization for soil mapping with random forest

Alexandre M.J-C. Wadoux[a,*], Dick J. Brus[b], Gerard B.M. Heuvelink[a]

[a] Soil Geography and Landscape Group, Wageningen University & Research, the Netherlands
[b] Biometris, Wageningen University & Research, the Netherlands

ARTICLE INFO

Keywords:
Pedometrics
Random forest
Optimal design
Conditioned Latin Hypercube
Spatial coverage
k-means
LUCAS
Spatial simulated annealing
Uncertainty assessment

ABSTRACT

Machine learning techniques are widely employed to generate digital soil maps. The map accuracy is partly determined by the number and spatial locations of the measurements used to calibrate the machine learning model. However, determining the optimal sampling design for mapping with machine learning techniques has not yet been considered in detail in digital soil mapping studies. In this paper, we investigate sampling design optimization for soil mapping with random forest. A design is optimized using spatial simulated annealing by minimizing the mean squared prediction error (MSE). We applied this approach to mapping soil organic carbon for a part of Europe using subsamples of the LUCAS dataset. The optimized subsamples are used as input for the random forest machine learning model, using a large set of readily available environmental data as covariates. We also predicted the same soil property using subsamples selected by simple random sampling, conditioned Latin Hypercube sampling (cLHS), spatial coverage sampling and feature space coverage sampling. Distributions of the estimated population MSEs are obtained through repeated random splitting of the LUCAS dataset, serving as the population of interest, into subsets used for validation, testing and selection of calibration samples, and repeated selection of calibration samples with the various sampling designs. The differences between the medians of the MSE distributions were tested for significance using the non-parametric Mann-Whitney test. The process was repeated for different sample sizes. We also analyzed the spread of the optimized designs in both geographic and feature space to reveal their characteristics. Results show that optimization of the sampling design by minimizing the MSE is worthwhile for small sample sizes. However, an important disadvantage of sampling design optimization using MSE is that it requires known values of the soil property at all locations and as a consequence is only feasible for subsampling an existing dataset. For larger sample sizes, the effect of using an MSE optimized design diminishes. In this case, we recommend to use a sample spread uniformly in the feature (i.e. covariate) space of the most important random forest covariates. The results also show that for our case study, cLHS sampling performs worse than the other sampling designs for mapping with random forest. We stress that comparison of sampling designs for calibration by splitting the data just once is very sensitive to the data split that one happens to use if the validation set is small.

1. Introduction

Conventional Digital Soil Mapping (DSM) employs geostatistical techniques to predict a continuous soil property at unobserved locations from measurements of this property at a finite number of sampling locations. Prediction is usually improved by exploiting the quantitative empirical relationship between the soil property and one or several environmental covariates. This leads to kriging with external drift, a basic technique in geostatistics, in which a soil property is modelled as a sum of a linear combination of covariates and a zero mean, spatially auto-correlated stochastic residual. Kriging models the soil property in a comprehensive, statistically sound way, but has several limitations (Webster and Oliver, 2007). First, it typically assumes that the residual is normally distributed, stationary and isotropic. Second, it considers that the model of spatial variation (i.e. the variogram) is estimated without error. Finally, the relation between the soil property and the covariates is usually assumed to be linear, and difficult to model when using a large number of cross-correlated covariates.

As an alternative, in recent decades (supervised) machine learning (ML) techniques have been applied for spatial prediction and DSM. ML refers to a large class of non-linear data-driven algorithms, originally developed for data mining and pattern recognition purposes. But ML is

[*] Corresponding author at: Soil Geography and Landscape Group, Wageningen University, Droevendaalsesteeg 3, Wageningen 6708BP, the Netherlands.
E-mail address: alexandre.wadoux@wur.nl (A.M.J.-C. Wadoux).

https://doi.org/10.1016/j.geoderma.2019.113913
Received 28 March 2019; Received in revised form 10 August 2019; Accepted 12 August 2019
Available online 20 August 2019

increasingly being used in other quantitative fields, such as in predictive soil mapping. ML techniques do not rely on rigid statistical assumptions about the distribution of the soil property and can handle numerous and correlated covariates as predictors, if at least a large calibration dataset is available. Examples on the use of ML techniques for DSM are Henderson et al. (2005) for mapping multiple soil properties at national-scale using decision trees, Behrens et al. (2005) for predicting soil units using artificial neural networks, and Grimm et al. (2008) to map soil organic carbon using random forest (RF). In this study we use the latter technique, whose use for soil mapping was recently formalized in Hengl et al. (2018).

Mapping requires calibrating a model using a sample from the target population. In consequence, the map accuracy is partly determined by the sample size and spatial locations of the sampling units with measurements of the target property that are used to calibrate the model. Various sampling designs are potentially suitable, depending on the intended mapping technique (Brus, 2019). In most cases, the soil is mapped using a known model of spatial variation (e.g. a variogram when using kriging). In this context, it is sensible to select a sample whose units are spread evenly throughout the area. This can be achieved by spatial coverage sampling (Royle and Nychka, 1998; Walvoort et al., 2010). If one assumes that the soil property is linked to environmental covariates, a robust strategy is to ensure that the measurements are also uniformly spread in the feature (i.e. covariates) space. This can be achieved using conditioned Latin Hypercube sampling (cLHS) (Minasny and McBratney, 2006) or feature space coverage sampling using the *k*-means (Hartigan and Wong, 1979) algorithm. The spatial coordinates can be added to the set of covariates so as to ensure a spread in both geographical and feature space. Brus (2019) noted that there is no single best sampling design, and that the best design depends on the technique used for mapping.

If the mapping technique is known beforehand, it is judicious to optimize a design for the intended use. In a model-based setting, we obtain an estimate of the prediction error variance, which can be minimized. As mentioned, for mapping with ordinary kriging this leads to a fairly uniform spread of the measurements in the geographic space, which can be obtained using a spatial coverage design (Brus et al., 2007). If one or several covariates are used as a trend in the kriging model, the optimized design shows a spread of the measurements in both geographic and feature space (Heuvelink et al., 2006). For mapping using ML techniques with covariates, Brus (2019) recommends to select the sample using feature space coverage sampling (FSCS) or cLHS. Both cLHS and FSCS aim for an even sampling density in the multivariate feature space, but in different ways. In cLHS it is done through minimization of a criterion which is a function of the marginal distributions and correlation matrix of the covariates using spatial simulated annealing, in FSCS it is achieved through minimization of a feature space distance criterion between sampling and prediction points using the *k*-means algorithm. This might be advantageous for ML techniques, which rely heavily on nonlinear relations, but this has not yet been confirmed by experimental results. In machine learning, we do not have a model-based estimate of the prediction error variance. Hence optimizing the sampling design is not straightforward, although it is possible to optimize the design using a universal prediction accuracy measure, such as the mean square error (MSE) of the prediction. To the best of our knowledge, little has been investigated on optimal sampling design for mapping using random forest.

A relevant contribution was made in Tuia et al. (2013) which optimized the allocation of new climatological stations in a case study in Austria. In this study support vector regression and active learning were used to derive the optimal locations of new stations so as to select the most important sampling units to be included in the sample, i.e. units that are used as support vectors. However, active learning is a sequential re-design technique which is appropriate to improve an already-calibrated ML model. Tuia et al. (2013) provides little insight into where to select the sampling locations when there is no prior ML model. In consequence, the conclusions of this study are of little use for practitioners who wish to map soil properties using machine learning.

The objective of this study is to investigate what makes a design optimal (sample size and sampling locations) for mapping using RF. To achieve this, we (i) estimate the population MSE with various sampling designs (viz. simple random sampling, cLHS, spatial coverage sampling (SCS), feature space coverage sampling (FSCS) and MSE optimized); (ii) statistically test the differences in the medians of the MSE sampling distributions of these designs through repeated selection of samples with a given design; and (iii) reveal the characteristics of the optimal design by analyzing the spread of the sampling locations in both geographic and feature space.

2. Materials and methods

2.1. Case study and data

We used the freely available soil database collected withing the framework of the European Land Use/Cover Area frame Statistical Survey (LUCAS) (Tóth et al., 2013). The LUCAS dataset is a sample of $N = 19{,}790$ georeferenced topsoil(0–30 cm) measurements of thirteen soil properties spanning 23 European countries. The sampling density varies between 11 and 77 measurements per 10,000 km^2 with an average of 48. The sample was collected by a two-stage systematic sampling design (Gallego and Delincé, 2010) using a stratification based on seven land cover classes. The resulting sample is spread fairly uniformly in space and within the different land cover classes. A map of the sampling locations is provided in Orgiazzi et al. (2018, Figure 1a). We used as target soil property the soil organic carbon (SOC) concentration in g kg^{-1} as measured by an automated vario MAX CN analyzer (Elementar Analysensysteme GmbH, Germany) (Tóth et al., 2013). In this study, we treat the *N* LUCAS topsoil SOC measurements as our population of interest. This means that we ignore that the LUCAS units are a sample from the true area of interest, in our case the European countries included in the survey.

In addition to the LUCAS SOC sample, we used a set of 197 readily available continuous environmental variables at 1 km × 1 km resolution as covariates. The list of covariates is given in Hengl et al. (2017).

2.2. Random forest

Random forest (RF) is an ensemble machine learning method based on decision trees (Breiman, 2001). A single decision tree is built by repeating a binary recursive partitioning of the input training data. In the root node, the training data are grouped into a single partition. All possible binary partitions of the training data are evaluated using a splitting metric (Louppe, 2014). The binary split that has the smallest metric is selected. The newly created partitions undergo the same procedure, until a stopping criterion, the minimum node size, is met. The final prediction for continuous variable is taken as the average of the values at the end of nodes of the decision tree.

Breiman (1996) introduced the bagging technique. Bagging stands for bootstrap and aggregating, and aims at reducing the prediction error variance by building an ensemble of regression trees. A large number of trees is built based on bootstrap samples of the training data. All tree predictions are aggregated through averaging, and these averages are taken as the final predictions. The RF algorithm elaborates on this and introduces an additional random perturbation during the splitting of a tree (Breiman, 2001). In each split, the partitioning considers only a subset of size mtry from the original set of covariates.

The calibration of the RF model is therefore based on three user-defined parameters. The first is the number of trees ntree. To avoid computational load in fine-tuning ntree for each model, we fixed ntree = 200, as a compromise between accuracy and computational efficiency. Lopes (2015) showed that in many cases 150 trees is sufficient to obtain stable results, in particular when the number of covariates is smaller than the calibration sample size. The second parameter, mtry, is the number of covariates to randomly select at each split. By default, we used mtry as set to the rounded down square root of the total number of covariates. The third parameter is the minimal terminal node size (nodesize), which controls the minimum number of training data required to continue the process of tree growth. Parameter nodesize was set to its default value of 5.

2.3. Sampling designs

We compared five common spatial sampling designs.

Random: Simple random sampling without replacement (Cochran, 1977) is the simplest form of random sampling technique which does not require any prior knowledge on the soil property spatial variation. In simple random sampling, every unit of the population has equal probability of being selected and sampling units are selected independently. We used the sample function from the base package in the R language (R Core Team, 2018) for selecting simple random samples.

Spatial Coverage Sampling (SCS): A SCS design aims at dispersing the units throughout the study area as uniformly as possible (Royle and Nychka, 1998). Coverage designs are created by minimization of a criterion that is a function of the distance between sampling and prediction locations. Brus et al. (1999) proposed to compute the Mean of the Squared Shortest Distance (MSSD), denoted MSSD_G hereafter, between sampling locations and the centre cells of a fine prediction grid as criterion to obtain a spatial coverage design. This criterion can be minimized by the fast *k*-means clustering algorithm. We implemented it with the R base function kmeans, using the spatial coordinates of the nodes of a fine discretization grid of the whole study area as clustering variables. Since our population of interest is the LUCAS dataset, the selected sampling units are the LUCAS points closest (in geographic distance) to the centres of the geographic clusters.

Feature Space Coverage Sampling (FSCS): A FSCS design follows the same principle as a spatial coverage design. However, in this case distances are measured in feature space instead of geographic space. Since covariates can have very different scales, it is important to standardize them (zero mean and unit variance) so that the criterion to be minimized becomes the Mean Squared Shortest Standardized Distance (MSSSD) (Brus, 2019), denoted MSSD_F hereafter. Sampling the centre of the *k*-means clusters ensures a uniform spread of the units in the multi-dimensional space of the covariates. We derive a FSCS design using the base R function kmeans. Similar to SCS, the LUCAS points closest (in standardized features space) to the centres of the clusters are used as sampling points.

Conditioned Latin Hypercube Sampling (cLHS): cLHS (Minasny and McBratney, 2006) is a stratified random sampling procedure. For each covariate, *n* marginal strata are defined using the quantiles of the cumulative frequency distribution, with *n* being the sample size. Next, an optimization procedure minimizes the weighted sum of two components (O_1 and O_3) so that each covariate contains one unit per stratum in the multi-dimensional feature space (O_1) and the correlation between the covariate values in the sample and in the population is preserved (O_3). Note that we do not use component O_2 because our case study has no categorical covariates. In cLHS, the covariate marginal distribution of the sample is close to that of the population (Brus, 2019). Note that in this study cLHS designs were based on the 20 most important covariates for RF. These covariates were derived from a RF model calibrated using all LUCAS topsoil OC data (about 20,000 units). We refer to this design as the cLHS (20) design, and compare it to the FSCS optimized on the same 20 most important RF covariates (FSCS (20)). The most important covariates of the RF model are defined using the Gini impurity index (Nembrini et al., 2018) as implemented in the ranger package (Wright et al., 2017) in R. We used the R package clhs (Roudier, 2018) to obtain a cLHS sample from the population. The default implementation in the clhs package assigns equal weights to the O_1 and O_3 components.

MSE optimized: In this case an optimized design is obtained by minimization of the MSE between the predicted and measured SOC in the independent test set, from a RF model whose parameters are estimated using a calibration set. The choice of the minimization criterion is discussed more extensively in the Discussion. The minimization is achieved by Spatial Simulated Annealing (SSA) (Van Groenigen and Stein, 1998; Wadoux et al., 2019). For each iteration in SSA, a RF model is built, which is subsequently used to predict at the test set locations. If the MSE becomes smaller, the proposed calibration sample is accepted, otherwise it is accepted with a probability that decreases during the optimization. We used the function optimUSER from the R package spsann (Samuel-Rosa, 2017). The total number of SSA iterations was set to 50 times the sample size.

Each of the five sampling designs described above is evaluated by computing the MSE between the SOC prediction and observation for an independent validation set. Computation of a population MSE value is detailed in the next section.

2.4. Estimation of the population MSE

The procedure for estimating the population MSE, for a given sampling design and sample size n = 100, 200, 500 and 1000, is given as follows:

```
for $r = 1$ to $R$ do
    Split the LUCAS dataset fully randomly into $K$ disjoint subsets of equal
      size (validation subsets);
    for $k = 1$ to $K$ do
        Define the $k$-th subset as the validation subset. Merge the remaining
          $K - 1$ subsets and split the merged set fully randomly into $L$ disjoint
          subsets of equal size (test subsets).;
        for $l = 1$ to $L$ do
            Define the $l$-th subset as the test dataset. Merge the remaining
              $L - 1$ subsets (calibration sampling subset).;
            for $m = 1$ to $M$ do
                1. Select a sample of size $n$ from the merged $L - 1$ subsets. This is the
                   calibration dataset. In case of the MSE optimized sampling design, the
                   sample is selected such that it minimizes the MSE of the test dataset. In
                   case of the other four designs the test dataset is not used but a sample is
                   selected according to the criterion of the design (i.e. simple random
                   sampling, SCS, FSCS, cLHS).
                2. Calibrate the RF model using the sample selected by the design.
                3. Predict at the locations of the validation dataset and compute the
                   squared prediction errors for all validation locations.
            end
            Average the $M$ squared prediction errors at each validation location.
        end
        Average the $L$ averaged squared prediction errors at each validation
          location.
    end
    Average the final averaged squared prediction errors over all LUCAS
      locations, the outcome is a single estimate of the population MSE
      associated with the selected calibration sample.
end
Plot the distribution and print summary statistics of the $R$ estimates of the
  population MSE.
```

Hereafter, the distribution of the R estimates of the population MSE as obtained using this procedure for a given sampling design type and calibration sample size n is referred to as the "MSE sampling distribution". Two sources of randomness are involved in the generation of the random variable. The first source of randomness is the repeated selection of calibration samples with a given sampling design. With each of the five sampling design types multiple calibration samples of a given size are selected. Each calibration sample is associated with a population MSE. So the population MSE is not a fixed quantity but a random variable. Simple random sampling is a random sampling design, so it is evident that with this design type multiple samples can be selected. SCS and FSCS using k-means are not random sampling designs. K-means is a deterministic algorithm, which means that, given an initial sample the final optimal sample is fixed. However, if we select the initial sample randomly, repeated selection of initial samples will result in different final samples. In practice, multiple initial samples are selected, and the sample with the smallest MSSD is kept. With a finite number of initial samples, repeated sampling using different seeds still will result in different optimized samples. For similar reasons, the cLHS and MSE optimized designs are also not fully deterministic and optimizing the calibration sample for these sampling design types will not always yield the same result

The second source is the random splitting of the LUCAS dataset into a validation subset and a subset from which the calibration sample is selected (calibration sampling subset). For a given calibration sample, the population MSE associated with this calibration sample is not known without error, but is estimated from the validation sample, introducing a random sampling error. The R estimates are independent and identically distributed (iid) realizations of the random variable "population MSE". Thus, the mean of the R estimates is an unbiased estimate of the expected value of the population MSE, while their standard error characterizes how close this mean is to the true mean (expectation) of the population MSE.

The values of R, K, L and M were chosen based on the computational load and degree of randomness of each design. When a design is more random, larger values of R, L and M are preferred. We chose $R = 10$ for all designs except for the MSE optimized design, where we used $R = 5$. K and L were set to 5 for all designs while $M = 20$ for the random and SCS designs, $M = 10$ for the FSCS design and $M = 1$ for the cLHS and

MSE optimized designs. We used $M = 1$ for the cLHS and MSE optimized design because these designs have a high computational load, while their randomness is modest.

If the K, L and M MSE estimates are not averaged and all $R \times K \times L \times M$ individual estimates of the population MSE are kept, the procedure above yields the distribution of the MSE that is obtained when one uses simple data splitting to select a single calibration dataset and a single validation dataset (a commonly used approach in DSM). The width of this distribution shows how uncertain the outcome of a data splitting validation procedure is, for a given calibration sample size and a given validation sample size.

2.5. Statistical hypothesis testing

Given the R population MSE estimates for each design, we tested for all pairs of designs and all calibration sample sizes n whether the medians of the distributions are significantly different using the Mann-Whitney U test (Wilcoxon rank-sum test) (Mann and Whitney, 1947). The Mann-Whitney U test is a non-parametric test of the null-hypothesis that two distributions have the same median. Thus, under the null hypothesis a randomly selected value from one of the distributions has 50% chance of being smaller or greater than a randomly selected value from the other distribution. Contrary to the two independent samples t-test the Mann-Whitney U test does not require the normality assumption of the distributions that are compared. Significant differences between MSE sampling distributions are characterized by a significance threshold fixed at a p-value smaller or equal than 0.05.

2.6. Diagnostics of the designs

Sampling designs are not only evaluated by the resulting MSE, but also by the spread of the samples in the geographic and feature space. This is done with the aim to reveal the characteristics of the designs, in particular the MSE optimized design, which may help to design future surveys. Thus, all sampling designs are evaluated in terms of all criteria, not just MSE, but also MSSD_G, MSSD_F and $O_1 + O_3$ as minimized in cLHS.

3. Results

3.1. MSE sampling distribution

Fig. 1 shows the estimated expectation of the population MSE (estimated by the average of the R estimates of population MSEs) with its standard error for all combinations of sampling designs and sample sizes. As expected, the MSE of the optimized design is smaller than for the other designs. This is particularly true for small sample sizes (e.g. 100 units) where the MSE optimized design has an expected MSE that is about 10% smaller than that of a simple random sampling design. For small sample sizes, simple random sampling and cLHS have the largest expected MSE (the medians are 7208 and 7174 $(\text{g kg}^{-1})^2$, respectively) and FSCS has a somewhat smaller expected MSE (median is 7090 $(\text{g kg}^{-1})^2$). This pattern is preserved with larger sample sizes, but the differences in expected MSEs become negligible as the sample size increases. For instance, the difference in expected MSE between designs is smaller than 100 $(\text{g kg}^{-1})^2$ for a sample size of 1000. Note that with cLHS for all sample sizes tested, the sampling distribution of MSE has a large median value, about equal to that of simple random sampling.

Fig. 2 shows the individual estimated MSE values ($R \times K \times L \times M$ MSE values) derived from the experimental design of Section 2.4, for all combinations of sampling designs and sample sizes. Fig. 2 shows how variable the outcome of a validation analysis can be in data splitting if this is done only once. The variability of the estimated MSE values is large and the MSE distributions for a given sample size overlap for all sampling designs tested. The MSE distribution is clearly narrower for MSE optimized designs and slightly more narrow for cLHS, compared to that of simple random sampling, SCS and FSCS. Note that the variability also depends on the validation sample size, which was about 4000 in

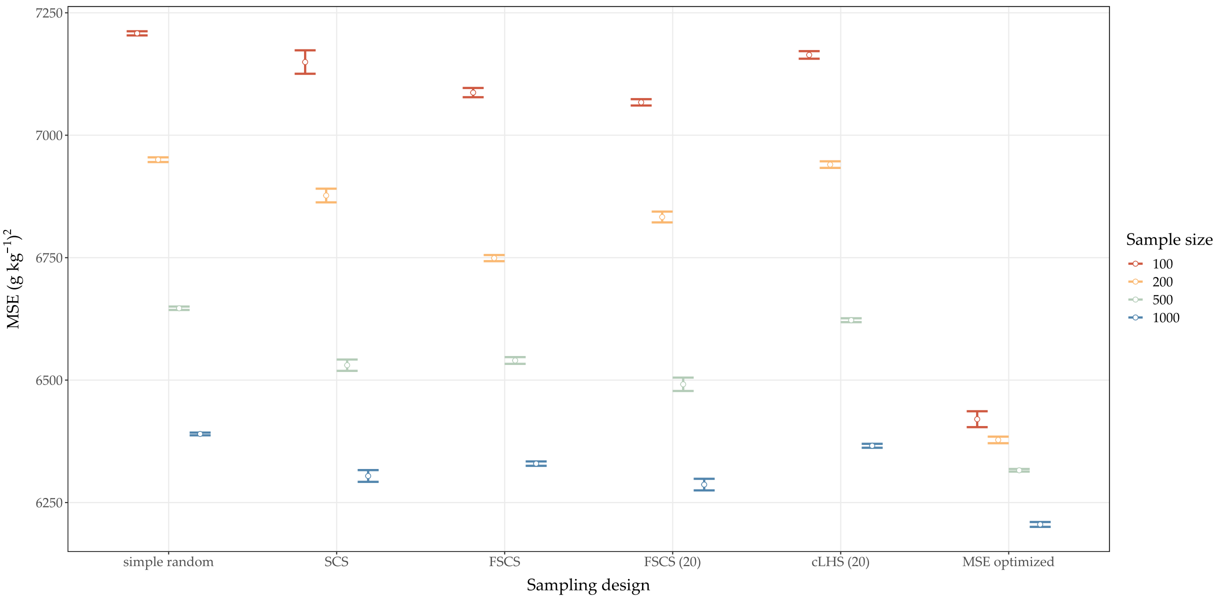

Fig. 1. Estimated expectation and standard error of the population MSE as derived with the experimental design of Section 2.4 but without averaging over K, L and M, for each of the tested sampling design types and for different calibration sample sizes. FSCS (20) and cLHS (20) refer to designs computed on the 20 most important covariates for the RF model, calibrated using all LUCAS topsoil OC data (about 20,000 units).

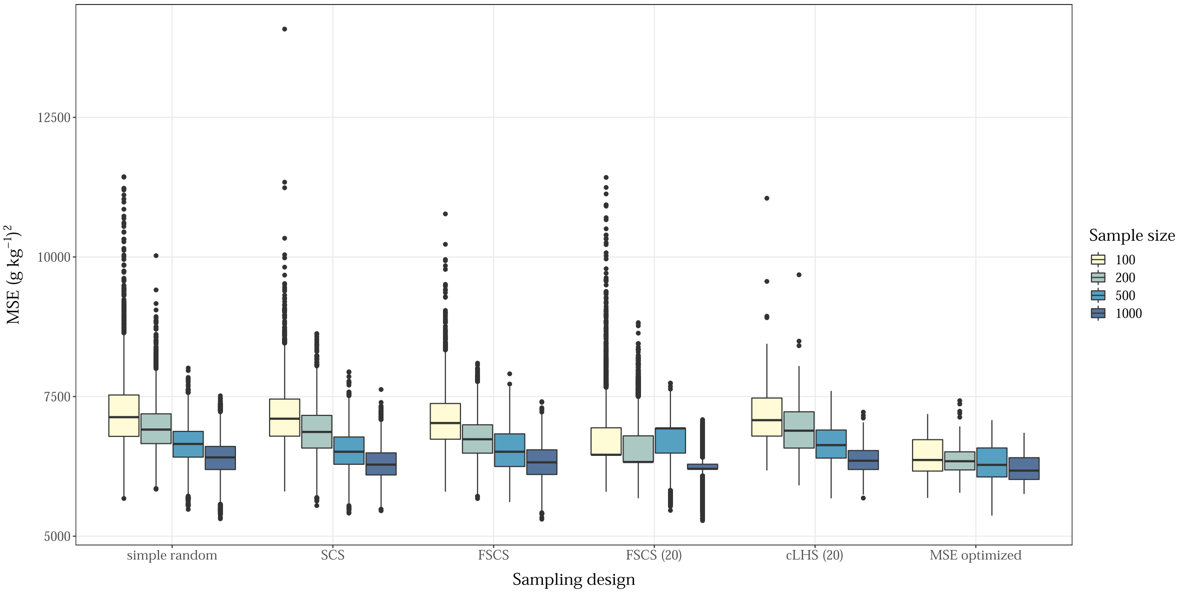

Fig. 2. Boxplots of individual population MSE estimates ($R \times K \times L \times M$ MSE estimates per boxplot) derived from the experimental design of Section 2.4, for each of the tested sampling designs and for different calibration sample sizes. FSCS (20) and cLHS (20) refer to designs computed on the 20 most important covariates for the RF model, calibrated using all LUCAS topsoil OC data (about 20,000 units).

Table 1
Mann-Whitney *U* test results for differences in median MSE obtained with random forest models calibrated with samples of various designs and sample sizes. Common letters indicate non-significant differences at significance level α of 0.05.

	Sample size			
	100	200	500	1000
Simple random	a	a	a	a
cLHS (20)	a b	a	a	a
SCS	b	b	b	b
FSCS	c	c	b	c
FSCS (20)	c	d	c	d
MSE optimized	d	e	d	e

this study. In many studies, the validation sample size will be much smaller than that, and this will increase variability even more. This is discussed more extensively in the Discussion.

3.2. Statistical hypothesis testing

Table 1 shows the result of the statistical hypothesis testing. Sampling designs with median MSE that are not significantly different at $\alpha = 0.05$ have the same letter. The median MSE of the cLHS design is, for all sample sizes tested, not significantly different from the median MSE of the simple random design. In contrast, the median MSE based on the MSE optimized design is always significantly different from those of other designs. This is an expected result given that the corresponding MSE standard errors shown in Fig. 1 do not overlap. For sample size 100, the median MSE of the cLHS design is not significantly

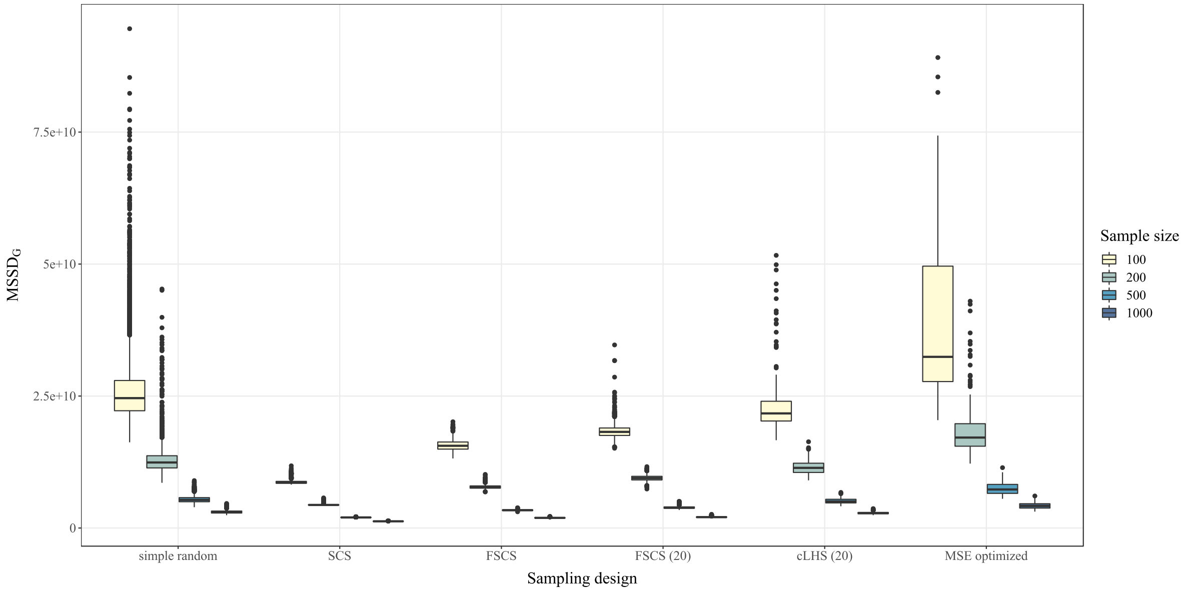

Fig. 3. Boxplots of MSSD$_G$ for all sampling designs and sample sizes.

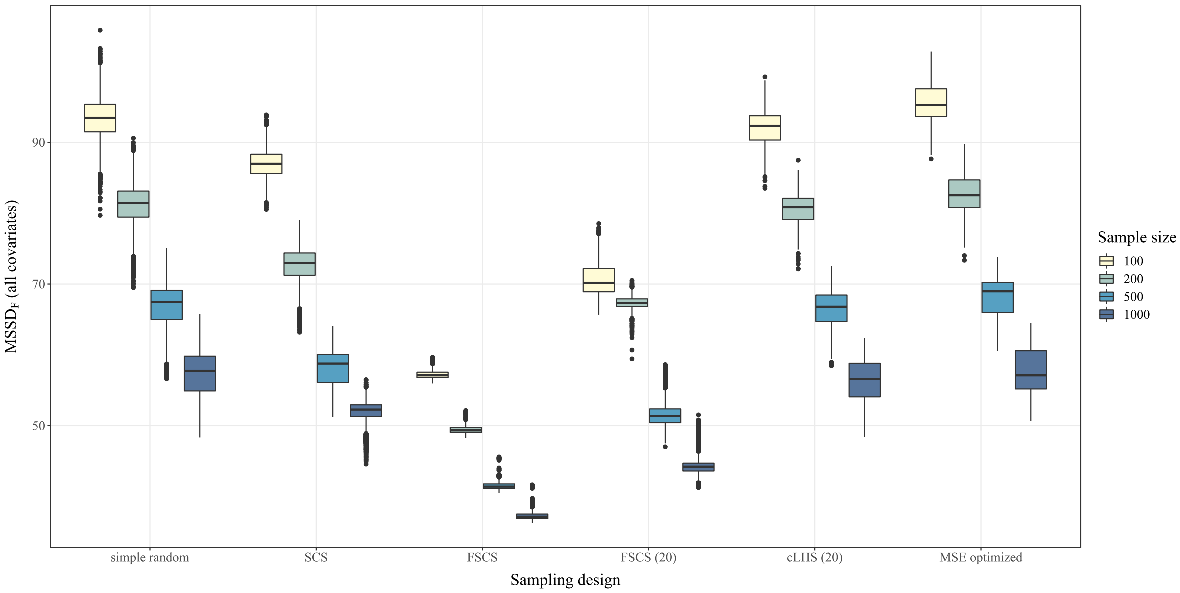

Fig. 4. Boxplots of the MSSD$_F$ for different sample sizes and sampling designs.

different from that of the SCS design, and the median MSE between FSCS designs using all or the 20 most important covariates are not significantly different. For sample size 500, the median MSE of the SCS design is not significantly different from that of the FSCS design using all covariates. Overall, it appears that parameters *R*, *L* and *M* were large enough to detect significant differences between designs.

3.3. Diagnostics of the designs

Fig. 3 shows the distribution of the MSSD$_G$ for all designs and sample sizes. Because the SCS design is optimized for this criterion it has always the smallest median MSSD$_G$ compared to other designs, for the same sample size. FSCS designs (optimized on all or the 20 most important covariates) have relatively small MSSD$_G$ values. This may be because the spatial coordinates are also included as covariates and hence used to optimize these designs. The simple random and MSE optimized designs have the largest MSSD$_G$ values and also the largest MSSD$_G$ variability (standard deviation of 7.28^9 and $1.51^{10}\,m^2$ for a sample size of 100, respectively). The MSE optimized design has on average the least uniform spread in geographic space, as shown by the median MSSD$_G$. This is the case for all sample sizes, even though the differences in MSSD$_G$ among designs are negligible for large sample sizes.

Figs. 4 and 5 show the MSSD$_F$ distributions, computed using all the covariates or a subset containing the 10 most important covariates of the RF models, respectively. Both figures show, as expected, that FSCS designs have the smallest MSSD$_F$ compared to other designs. All other sampling designs have similar MSSD$_F$ distributions. A similar pattern is

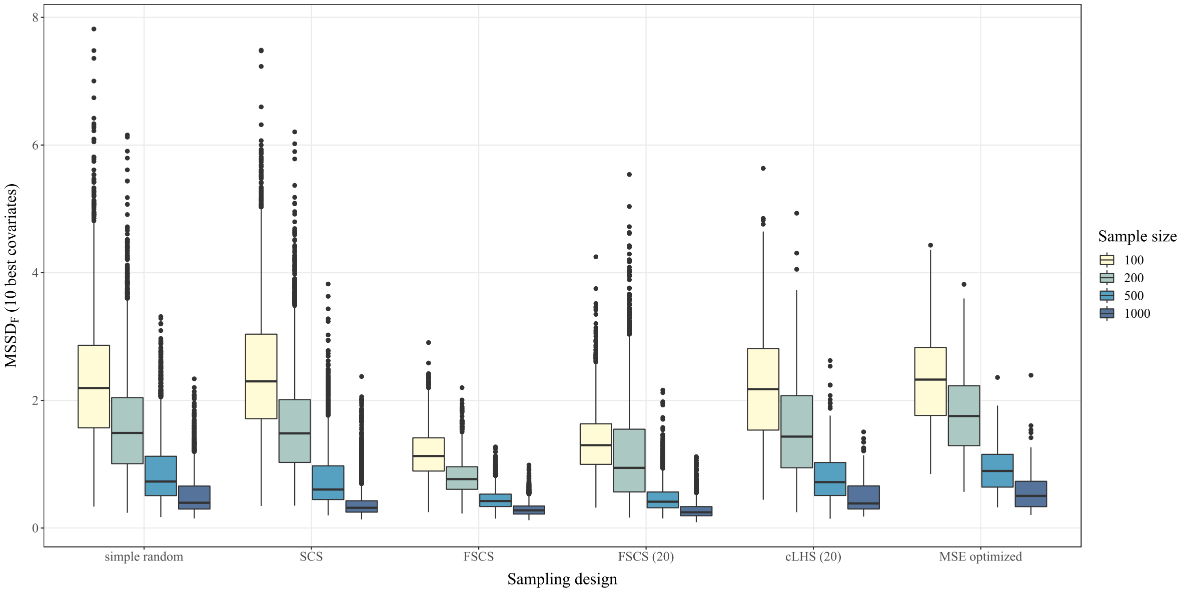

Fig. 5. Boxplots of the MSSD$_F$, based on the ten most important covariates of each design.

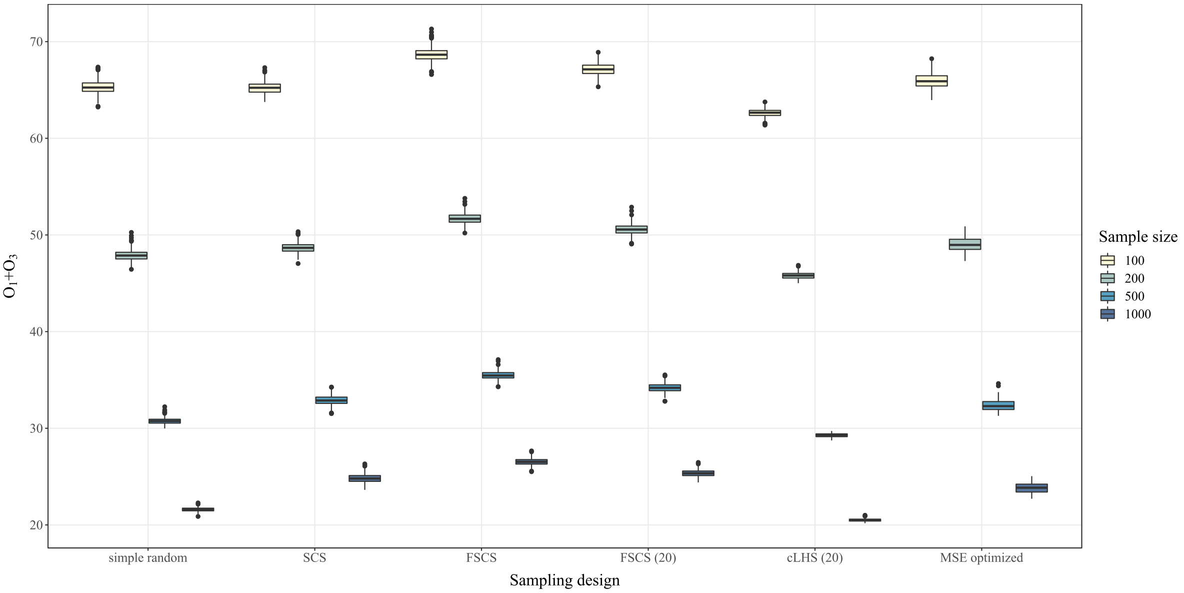

Fig. 6. Boxplots of the $O_1 + O_3$ cLHS criterion for each of the sampling designs and sample sizes. The elements $O_1 + O_3$ are computed as means.

observed in Fig. 5: all designs (except FSCS designs) have similar $MSSD_F$ distributions. Note that simple random and SCS designs have a very large spread in the $MSSD_F$, while the MSE optimized design has narrower $MSSD_F$ distributions and very few outliers.

Fig. 6 shows the distribution of the $O_1 + O_3$ cLHS criterion computed for each of the designs and sample sizes. Note that in Fig. 6 the elements $O_1 + O_3$ are not computed as sums but as means. This is discussed more extensively in the Discussion. Since a cLHS design is optimized for this criterion, it has the smallest values for all sample sizes. For large sample size, the simple random sampling design is almost equivalent in terms of the cLHS criterion. The FSCS designs (using all or the 20 most important covariates) have always the largest value of the cLHS criterion.

Fig. 7a indicates how often a sampling location is selected in the MSE optimized design, where red colours correspond to a case where a sampling location is selected more often than would be expected under a simple random sampling design and blue colours indicate the opposite. Fig. 7b shows the proportion of sampling locations used for calibration of the MSE optimized design. Red colours indicate that locally, a relatively large number of sites were used for calibration, while blue colours indicate that relatively few sites in the local neighbourhood were used for calibration of the MSE optimized design. Areas with fewer than five LUCAS sites within the local neighbourhood (100 km circular radius) were masked out. Fig. 7b shows that the MSE optimized design leads to a fairly high relative density in a geographic band spanning from France to Poland. Germany and Denmark have a high relative density across their entire country. Great Britain, Ireland, southern and northern Europe tend to have a lower relative density of sampling units included in the MSE optimized design, even though they might locally have a very high *absolute* density of sampling locations (e.g. North of Madrid).

4. Discussion

4.1. Impact of sampling designs on prediction accuracy

The sampling design had a significant impact on the accuracy of random forest predictions. In the case study mapping topsoil OC using RF in Europe, the MSE optimized design had the smallest mean squared prediction error, as shown in Fig. 1. This is because the MSE optimized design was optimized for this purpose, by minimizing the MSE of the test set. All other designs reach substantially higher MSE value than the MSE optimized design. However, the MSE optimized design can be used only when subsampling an existing dataset with known values of the target soil property at all locations. In other words, it may be used in a case where thinning of an existing sampling network is required, but not in a case where one needs to design a sampling scheme from scratch, such as in a reconnaissance survey. In this case, it is best to use a FSCS design which, for the case study, had the smallest prediction MSE of all other designs tested. This is not surprising because predictions made by machine learning methods rely on non-linear relationships with covariates, and estimation of these relationships benefits from a spread of the sampling units in feature space, as noted by Brus (2019). To our surprise the MSE values obtained with cLHS design were large (Fig. 1) and not statistically different from those obtained using simple random sampling (Table 1). This is discussed more extensively later in this Discussion. In spite of the differences in MSE between designs for small sample sizes, the MSE between designs for large sample sizes (in our case study larger than 1 unit per 4159 km^2) are negligible. This result applies to our case study using the LUCAS dataset as the population of interest, but is likely also valid more generally: increasing the sample size reduces the MSE differences between designs because the selected sample covers all cases sufficiently well.

4.2. How to compare calibration sampling designs?

Several studies (e.g. Schmidt et al., 2014; Ng et al., 2018) have investigated the effect of the sampling design on the resulting map accuracy or calibrated model. In these studies one sample was selected, which was randomly split into a validation sample and a calibration sample. While this is a common approach in soil mapping studies, Fig. 2 shows that it is delicate to draw conclusions based on a single data split because of the large variability of MSE estimates and the large overlap of the MSE distributions of the various sampling design types for a given sample size. In other words, the conclusion that calibration design type A is better than design type B obtained with a given data split is very sensitive to the data split that happened to have been used. Fig. 2 shows that a different conclusion might well have been obtained if a different split had been used. Note also that the results obtained in Fig. 2 depend on the size of the calibration and validation sets. In our case study, the validation set was large compared to the population size. One can expect that in most practical cases, where a smaller validation set is used, the distribution of the population MSE will be even wider than shown in Fig. 2. The effect of the validation set size on the distribution of the population MSE has not been investigated in this study.

So, in studies comparing calibration sampling designs, it is more useful to compare these designs on the basis of the MSE sampling distribution obtained by repeated selection of calibration samples, instead of on the estimated MSE obtained with a single calibration and a single validation sample. However, if the validation set size is sufficiently large, the MSE estimate computed on a validation set from a single split will still be close to the true MSE.

4.3. Design diagnostics

In practice we cannot obtain an MSE optimized design because this requires that the target property is known at all locations in the study area. This is why it is useful to interpret and diagnose the MSE optimized designs obtained in the case study, because if general patterns can be derived then these may be used to design spatial sampling designs for DSM using RF. Diagnostics on the MSE optimized designs reveal that RF does not benefit much from a spread of the sampling units in geographic space (Fig. 3). One possible reason is that spatial location is ignored during the RF modelling process (Hengl et al., 2018) and in other machine learning techniques (Behrens et al., 2018). Figs. 4 and 6 show that, in addition, RF neither benefits much from a spread of the sampling units in the feature (i.e. covariate) space, nor from reproducing the marginal distributions of the covariates. This is unexpected because many studies (e.g. Castro-Franco et al., 2015; Domenech et al., 2017; Brus, 2019) suggested that spread in the feature space is crucial. In fact, it is more subtle than that. We learn from Fig. 5 that the importance of the covariates used in the RF model must be taken into account as well. This is an important finding of this study: the predictions made by a RF model benefit from a design spread uniformly in the space spanned by the most important covariates. We acknowledge that this finding is based on a single case study and needs to be tested in further research. If this finding is confirmed by future studies, one can derive practical recommendations to design a soil survey for mapping with RF: (i) determine what are the most important covariates, either using a legacy sample, previous studies, pedological expertise or a two-stage sampling approach; and (ii) optimize the design using coverage sampling in covariate space for the important covariates (possibly using weights derived from the importance).

4.4. Conditioned Latin Hypercube sampling design

While it was shown above that RF benefits from a uniform spread of the sampling locations in the feature space of the most important covariates, predictions based on the cLHS design were on average not more accurate than those based on a simple random sampling design, and even worse than predictions obtained using all other designs. Sampling the marginal distribution of the covariates as implemented in cLHS was not a useful strategy for mapping with RF in this study. The criterion values in Figs. 4

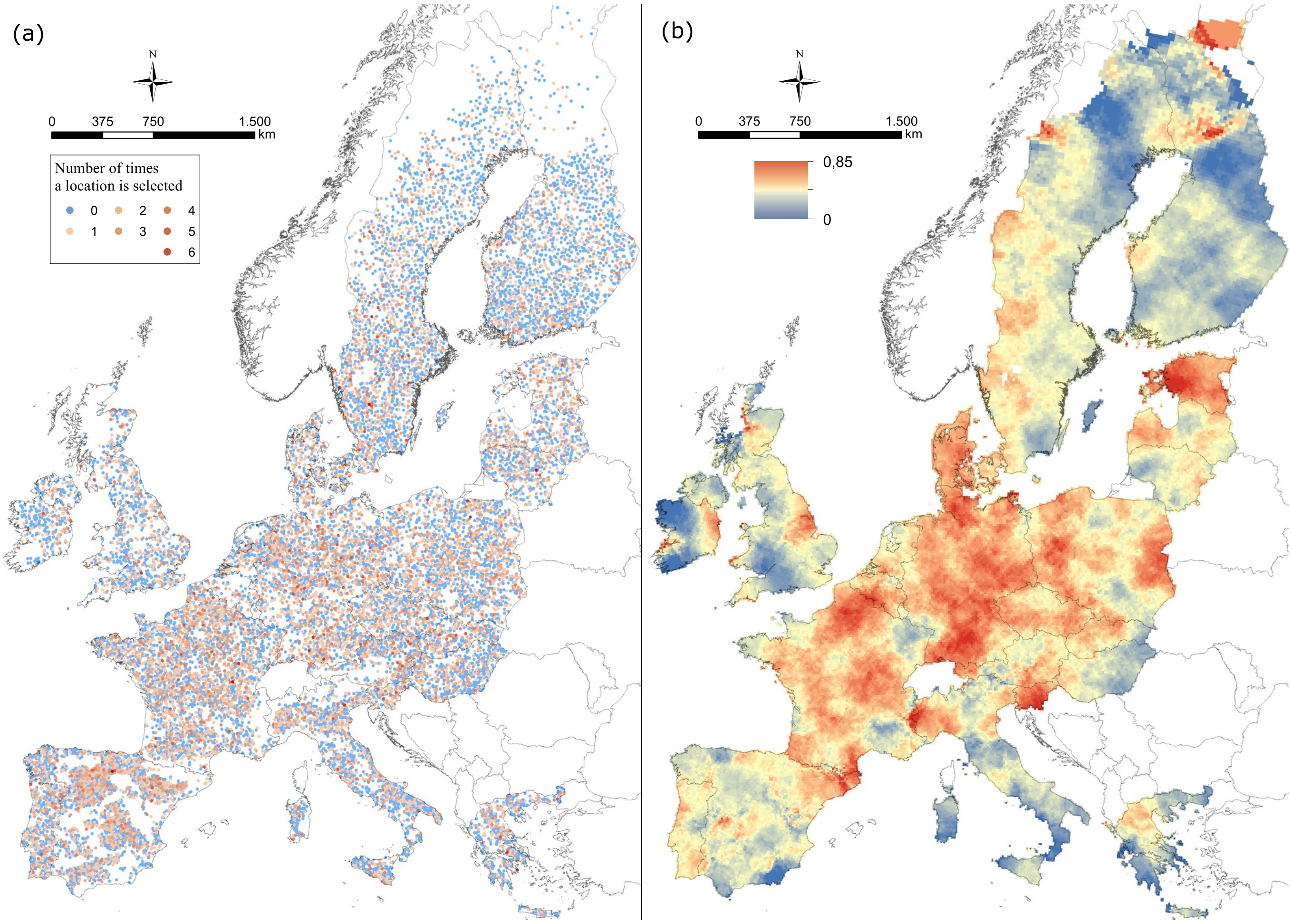

Fig. 7. Number of times a sampling location is selected by the MSE optimized design (a). Red colours indicate that the location is selected more often than one would expect under a simple random sampling design, blue colours indicate that it is selected less often than expected under simple random sampling. Ratio of number of sampling sites used for calibration of the MSE optimized model and total number of sampling sites, as computed in a circular neighbourhood with radius 100 km (b). Red colours indicate regions for which sampling units are often included in the MSE optimized design, blue colours refer to regions for which sampling units are less often included in the MSE optimized design.

and 6 show that the cLHS and FSCS designs are very different in the way they spread the sampling units in feature space. This had a major impact on the resulting prediction accuracy in this study. While several studies (e.g. Schmidt et al., 2014; Contreras et al., 2019) showed that using RF in combination with cLHS gives the most accurate prediction, we showed that in our case cLHS performed worse than other sampling designs exploiting covariates for mapping with RF. While the results obtained by Schmidt et al. (2014) and Contreras et al. (2019) are possible outcomes (as shown by Fig. 2), these could have been incidental results if their validation sample size was small. To judge whether the validation metrics are sufficiently accurate, it is best to compute confidence intervals of the validation metrics, which is possible only if the validation sample is collected using probability sampling (Brus et al., 2011). Confidence intervals are needed to be able to interpret the true value of the validation statistics and evaluate whether differences in prediction accuracy between different sampling designs for mapping are statistically significant. Note that in this study we used the cLHS implementation from the R package of Roudier (2018) following the Minasny and McBratney (2006) paper, where the O_1 and O_3 components are computed as sums, not as means. The resulting criterion is therefore affected by the magnitude of the O_1 and O_3 components. This may cause an unbalance between the relative importance of O_1 and O_3. To solve this problem, other implementations (e.g. Samuel-Rosa, 2017) compute O_1 as the mean of the absolute deviations between the marginal strata sample size and targeted sample size, while O_3 is computed as the mean of the deviations over all off-diagonal entries of the correlation matrix. Taking the latter into account might improve the performance of the cLHS design. However, we did not consider it in this study.

4.5. Optimization criteria

In our case study, the MSE optimized design was derived based on the MSE between predicted and measured SOC values in the test dataset. The MSE is a universal criterion which can be computed for any mapping method, also in a case where we do not have a model-based estimate of the prediction error variance. If a model-based estimate of the prediction error variance is available, we can use a function of the prediction error variance as minimization criterion. Obvious candidates for such function are the spatial mean (Brus and Heuvelink, 2007) and maximum (Van Groenigen et al., 1999) prediction error variance. For the RF model used in the case study, the prediction error can be quantified by Quantile Regression Forest (QRF) (Meinshausen, 2006), for instance using the width of the 90% prediction interval. We explored this and used the average width of the QRF 90% prediction interval over the study area (i.e. the 23 EU countries included in this study) as a minimization criterion. However, we observed that the sampling units of the optimized design had a narrow SOC distribution and small SOC variance. These sampling units were selected because this resulted in narrow QRF predicting intervals and hence a small criterion value. As a result, validation of the quantified uncertainty (e.g. using accuracy plots Deutsch, 1997; Wadoux et al., 2018) showed that the uncertainty was systematically and severely underestimated. Thus, we did not pursue this any further.

4.6. Sampling for other machine learning techniques

Finally, there is a need to further investigate whether a design that is optimal for RF modelling is also optimal for other machine learning models. Our results were obtained for a tree-based model. We hypothesize that a design that is optimal for RF may also be efficient for modelling and predicting using other tree-based models (e.g. CART; Breiman, 2017), because they are comparable in their basic structure and splitting metrics. Note also that in our study we investigated sampling design for mapping a single soil property. This could be extended to sampling design optimization for multivariate soil mapping using random forests. Sampling to support other machine learning models (e.g. support vector machine or deep neural network) introduces additional considerations and also deserves further investigation. For example, Pozdnoukhov and Kanevski (2006) and Tuia et al. (2013) optimized a network for mapping using support vector machine. They specifically aimed at minimizing the “risk” of selecting new sampling units that do not have a valuable contribution to the model (by becoming support vectors). Recently, Wadoux (2019) showed how a deep neural network can be used for soil mapping, and how the minimized loss function can be modified to include additional information (e.g. to quantify the prediction uncertainty). Formulating a loss function that searches for optimal units to be measured using the feature (i.e. covariates) space has been tackled by MacKay (1992). How much a design optimal for a neural network model would differ from that of a RF model requires further study. This would certainly make a valuable contribution to future DSM studies.

5. Conclusion

We computed an MSE optimized design for mapping with RF and compared it to several commonly used sampling designs. We compared the designs in terms of both prediction accuracy and spread of sampling units in geographic and feature space. In a case study, we used the LUCAS topsoil OC measurements as our population of interest, from which subsamples were collected. From the Results andDiscussion we draw the following conclusions:

- An MSE optimized design provides the smallest mean squared prediction error. However this is feasible only in case of subsampling an existing dataset with known values of the target soil property at all locations.
- In terms of accuracy, a sample selected by feature space coverage sampling of the most important covariates had the closest match with the MSE optimized design.
- Comparison of calibration sampling designs on the basis of the estimated population MSE obtained by splitting the data only once into a calibration and validation subset is prone to incidental results if the validation sample size is small. One should compute confidence intervals of the validation metrics and verify that these are sufficiently narrow. Narrow confidence intervals can be obtained by repeatedly splitting the data into a calibration and validation subset (this study) or by using a sufficiently large validation set.
- Preferably calibration sampling designs are compared on the basis of estimates of the “expectation” of the population MSE. Performance differences between sampling design types have no real meaning until these are shown to be statistically significant.
- For large sample sizes, the differences between prediction accuracies of different designs become negligible. In our continental scale case study, this was for a sampling density greater than 1 sampling unit per about 4000 km^2.
- In our case study, predictions based on a cLH sample had the poorest prediction accuracy, similar to that of predictions based on a simple random sample. There is need for further observational research to investigate whether conditioned Latin Hypercube sampling (cLHS) design is efficient for mapping using RF.
- Diagnostics on the MSE optimized design showed that for RF the optimal sampling design is not achieved by a uniform spread of the sampling units in the geographic and/or feature (i.e. covariate) space, nor from reproducing the marginal distributions of the whole set of covariates.
- Further diagnostics of the MSE optimized design showed that the importance of the covariates used in the RF model must be taken into account when optimizing the spatial sampling design. RF benefits from a spread of the sampling units uniformly in the feature space of the most important covariates of the RF model. The most important covariates can be selected using a sample from a reconnaissance survey, by pedological expertise or by a two-stage sampling strategy.

Acknowledgments

We thank ISRIC - World Soil Information and Tom Hengl, Envirometrix Ltd., Wageningen, for providing the covariates used in the case study.

References

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168, 21–33.

Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., MacMillan, R.A., 2018. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci. 69, 757–770.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge, New York.

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

Brus, D.J., De Gruijter, J.J., Van Groenigen, J.W., 2007. Designing spatial coverage samples using the k-means clustering algorithm. Dev. Soil Sci. 31, 183–192.

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. Eur. J. Soil Sci. 62, 394–407.

Brus, D.J., Spätjens, L.E.E.M., De Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89, 129–148.

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 Sci. 180, 74–85.

Cochran, W.G., 1977. Sampling Techniques, 3rd edition. John Wiley & Sons, New York.

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. Int. J. Climatol. 39, 2209–2226.

Deutsch, C., 1997. Direct assessment of local accuracy and precision. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics Wollongong'96, pp. 115–125.

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.

Gallego, J., Delincé, J., 2010. The European land use and cover area-frame statistical survey. In: Library, W.O. (Ed.), Agricultural Survey Methods, pp. 149–168.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado island-digital soil mapping using random forests analysis. Geoderma 146, 102–113.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C. Appl. Stat. 28, 100–108.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al., 2017. Soilgrids250m: global gridded soil information based on machine learning. PLoS one 12, e0169748.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Heuvelink, G.B.M., Brus, D.J., de Gruijter, J.J., 2006. Optimization of sample configurations for digital mapping of soil properties with universal kriging. Dev. Soil Sci. 31, 137–151.

Lopes, M.E., 2015. Measuring the algorithmic convergence of random forests via bootstrap extrapolation. In: Technical Report. Department of Statistics, University of California, Davis CA.

Louppe, G., 2014. Understanding Random Forests: From Theory to Practice. Ph.D. thesis. University of Liège.

MacKay, D.J.C., 1992. Information-based objective functions for active data selection. Neural Comput. 4, 590–604.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 50–60.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Minasny, B., McBratney, A.B., 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388.

Nembrini, S., König, I.R., Wright, M.N., 2018. The revival of the Gini importance? Bioinformatics 34, 3711–3718.

Ng, W., Minasny, B., Malone, B., Filippi, P., 2018. In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. PeerJ 6, e5722.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153.

Pozdnoukhov, A., Kanevski, M., 2006. Monitoring network optimisation for spatial data classification using support vector machines. Int. J. Environ. Pollut. 28, 465–484.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roudier, P., 2018. Package "clhs". R package version 0.7-0. https://CRAN.R-project.org/package=clhs.

Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Comput. Geosci. 24, 479–488.

Samuel-Rosa, A., 2017. spsann: Optimization of Sample Configurations using Spatial Simulated Annealing. R package version 2.1-0. https://CRAN.R-project.org/package=spsann.

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.

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

Tuia, D., Pozdnoukhov, A., Foresti, L., Kanevski, M., 2013. Active learning for monitoring network optimization. In: Spatio-Temporal Design: Advances in Efficient Data Acquisition. Wiley Online Library, Chichester, pp. 285–318.

Van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

Van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. J. Environ. Qual. 27, 1078–1086.

Wadoux, A.M.J.-C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.

Wadoux, A.M.J.-C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

Wadoux, A.M.J.-C., Marchant, B.P., Lark, R.M., 2019. Efficient sampling for geostatistical surveys. Eur. J. Soil Sci (In Press).

Walvoort, D.J.J., Brus, D.J., De Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36, 1261–1267.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons, Chichester.

Wright, M.N., Ziegler, A., et al., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. JJ. Stat. Softw. 77.