Geoderma 370 (2020) 114366

Contents lists available at ScienceDirect

Geoderma

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

Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps

Tianwu Ma[a,b,c], Dick J. Brus[a,b,g,*], A-Xing Zhu[a,b,c,d,e,f], Lei Zhang[a,b,c], Thomas Scholten[h,i]

[a] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[b] School of Geography, Nanjing Normal University, Nanjing, 210023, China
[c] Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, 210023, China
[d] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[e] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA
[f] Center for Social Sciences, Southern University of Science and Technology, Shenzhen 518055, China
[g] Biometris, Wageningen University and Research, PO Box 16, Wageningen 6700 AA, Netherlands
[h] Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Rümelinstr. 19-23, Tübingen, Germany
[i] DFG Cluster of Excellence "Machine Learning", University of Tübingen, AI Research Building, Maria-von-Linden-Str. 6, 72076 Tübingen, Germany

ARTICLE INFO

Keywords:
Soil sampling
Random forest
Similarity-based predictive soil mapping
K-means
Simulated annealing
Calibration sampling

ABSTRACT

This study investigates sampling design for mapping soil classes based on multiple environmental features associated with the soil classes. Two types of sampling design for calibrating the prediction models are compared: conditioned Latin hypercube sampling (CLHS) and feature space coverage sampling (FSCS). Simple random sampling (SRS), which does not utilize the environmental features, is added as a reference design. The sample sizes used are 20, 30, 40, 50, 75, and 100 points, and at each sample size 100 sample sets were drawn using each of the three types of design. Each of these sample sets was then used to calibrate three prediction models: random forest (RF), individual predictive soil mapping (iPSM), and multinomial logistic regression (MLR). These sampling designs were compared based on the overall accuracy of predicted soil class maps obtained by these three prediction methods. The comparison was conducted in two study areas: Ammertal (Germany) and Raffelson (USA). For each of these two areas a detailed legacy soil class map is available. These soil class maps were used as references in a simulation study for the comparison. Results of both study areas show that on average FSCS outperforms CLHS and SRS for all three prediction methods. The difference in estimated medians of overall accuracy with CLHS and SRS was marginal. Moreover, the variation in overall accuracy among sample sets of the same size was considerably smaller for FSCS than that for CLHS. These results in the two study areas suggest that FSCS is a more effective sampling design.

1. Introduction

Information on the spatial distribution of soil classes is of great importance for, amongst others, agriculture management and watershed process simulation (Cook et al., 2008; Lagacherie, 2008; McBratney et al., 2003; Sanchez et al., 2009). As the soil cannot be observed everywhere, we need to predict the soil classes at unvisited locations from a finite set of observations at other locations. These observations can then be used to calibrate a model relating the soil classes to environmental features whose spatial variation are readily available. Subsequently the calibrated model can be used to predict the soil classes at any unvisited location in an area. The selection of observation locations (sample locations) for calibration is a particularly important step (Brus, 2019) because they directly impact the calibration of the prediction model. Clearly, sampling design, which determines the calibration sampling locations, directly affects the prediction accuracy of the spatial distribution of soil classes. However, sampling is labor intensive and resources demanding (Webster and Oliver, 1990). Therefore, how to improve the sampling efficiency, achieving a highly accurate and detailed soil class map with limited samples, is an important issue in soil sampling.

To calibrate a model that relates soil classes to environmental features we must select points (locations) from the multidimensional space defined by the environmental features. Brus (2019) suggests amongst

* Corresponding author at: Biometris, Wageningen University and Research, PO Box 16, Wageningen 6700 AA, Netherlands.
E-mail address: dick.brus@wur.nl (D.J. Brus).

https://doi.org/10.1016/j.geoderma.2020.114366
Received 16 January 2020; Received in revised form 25 March 2020; Accepted 1 April 2020
Available online 10 April 2020

others conditional Latin hypercube sampling (CLHS) and feature space coverage sampling (FSCS) for predictive mapping with environmental features. In this paper, the term “sample” and “sample set” are the same, they both refer to a subset from a population and are used interchangeably in this paper. The term “sampling points” refers to the individual units in a sample or a sample set, and “sample size” refers to the number of sampling points in a sample or sample set.

CLHS is a popular design in soil mapping, see for instance (Lin et al., 2011; Roudier et al., 2012; Schmidt et al., 2014; Stumpf et al., 2016; Yang et al., 2020). CLHS is done by minimizing a criterion which is a function of the number of sampling points in the marginal strata and the correlation matrix of the environmental features. The criterion is minimized through simulated annealing (Minasny and McBratney, 2006). Several studies have investigated the impacts of this sampling design on the accuracy of digital soil mapping. Worsham et al. (2012) compared simple random sampling, stratified random sampling and CLHS for mapping soil carbon content in a forest plot using regression kriging with sample sizes of 40, 100 and 300 points. The root mean squared error of stratified random sampling and CLHS were about equal. Schmidt et al. (2014) compared weighted CLHS, fuzzy *k*-means sampling and response surface sampling for mapping several quantitative soil properties, using multiple linear regression and random forest, in a real-world case study and in two simulated fields. Their study (Table 4 in that paper) showed that overall weighted CLHS and fuzzy *k*-means performed comparable. However, the calibration samples were also used for validation, so that no reliable conclusions can be drawn from this study.

With FSCS the minimization criterion is a function of the squared shortest distances in the feature space of the sampling points to the prediction points. The criterion is minimized by the *k*-means algorithm (Brus, 2019). Recently, Wadoux et al. (2019) compared CLHS, FSCS and Simple Random Sampling (SRS) for mapping a continuous soil property. They did so in a simulation study, so that for each sampling design many samples could be drawn. They concluded that the median mean squared error (MSE) with FSCS was significantly smaller than with CLHS for all sample sizes.

So far the comparisons are only conducted for mapping continuous soil property and there is a need to compare these sampling approaches for mapping soil classes. The aim of this research was to compare CLHS, FSCS and SRS for predicting soil classes, using several commonly applied prediction methods. The focus is on the performance of the sampling designs, not on the performance of the individual prediction methods, but we were interested in seeing whether the relative performance of the sampling designs was consistent among the prediction methods or not.

2. Materials and methods

2.1. Experimental design

The purpose of this paper is to compare the sampling efficiency of CLHS and FSCS for predicting soil classes in a simulation study. The sampling efficiency is quantified by the overall accuracy of spatial predictions of soil classes. SRS is added to our study as a reference sampling design. Comparison of CLHS and FSCS with SRS gives insight in how useful the environmental features are at the sampling stage. For prediction, we selected three prediction methods: random forest (RF), multinomial logistic regression (MLR) and individual predictive soil mapping (iPSM) (Zhu et al., 2018, 2015).

To examine the impacts of sampling design types on prediction model calibration a range of sample sizes was used: n = 20, 30, 40, 50, 75 and 100. For each sample size, 100 repeats were conducted for each sampling design type, CLHS, FSCS and SRS. In other words, for each of the sampling design types, we generated 100 sample sets at the sample size of 20 points, 100 sample sets at the size of 30 points, and so on. Each of these sample sets was subsequently used to calibrate (train) the prediction models. The calibrated models were then used to predict the soil classes. This results in 3 (design types) × 6 (sample sizes) × 100 (number of sample sets) × 3 (prediction methods) = 5400 models for each study area in this experiment.

We conducted the simulation study in two study areas. A detailed soil class map is available for each of the two study areas. The predicted soil classes using the calibrated models were compared with the soil classes as depicted on the soil class maps. The difference between the soil classes as depicted on the legacy soil class map and the predicted soil classes are used as the prediction errors. We are aware that the legacy soil class maps are not error free, so that the estimated overall accuracy of the predictions can be biased. However, we were not interested in the overall accuracies *per se*, but in the differences of the overall accuracies between calibration sampling designs. So, we think that despite the errors in the soil classes as depicted on the legacy soil class maps, these maps are still useful for comparing calibration sampling designs. Our assumption is that sampling design A does not profit more from ignoring the errors in the legacy map than sampling design B. In other words, we assumed that the errors in the map do not favor one calibration sampling design over another.

2.2. Study areas and soil class maps

We selected two study areas of which a detailed soil class map and data on multiple environmental features were available. The first study area, Ammertal (Fig. 1), covers the catchment of the river Ammer, Germany. The Ammertal is a first order tributary to the river Neckar and as part of the upper river Neckar basin that drains into the river Rhine which is the second largest river in Germany. The Ammertal is embedded in a hilly topography between the north-eastern foot of the Black Forest in the west and the Swabian Alb in the southeast. Elevations range between 300 and 500 m above sea level. The bedrock consists mainly of a sequence of evaporites, sandstones, and claystones of the Upper Triassic and the Lower and Middle Jurassic (Grathwohl et al., 2013). Land use is mainly agriculture (71%), followed by forest (12%) and urban areas (17%). Mean annual precipitation is 920 mm,

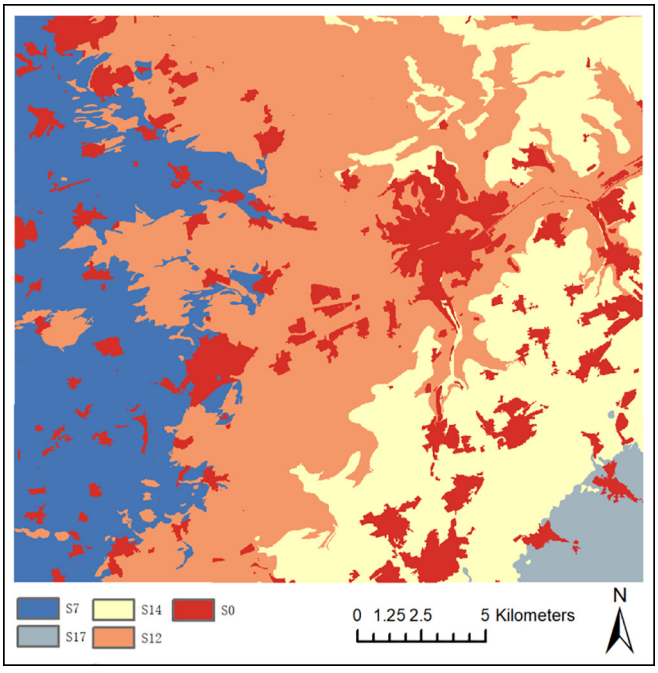

Fig. 1. The detailed soil class map of the Ammertal study area showing soil landscapes (in German: 'Bodengroßlandschaft') as a specific spatial context of soil classes according to the German Soil Classification (Ad-hoc-AG Boden, 2005).

Table 1
Description and category of the environmental features in Ammertal.

Category	Description, abbreviation and references
Local terrain attribute	Elevation (EL), Aspect (AS)(Horn, 1981), Mean slope (MS)(Dietrich and Montgomery, 1998), Steepest slope (SS)(Tarboton, 1997), Plan curvature (PLC), Profile curvature (PRC), Horizontal curvature(HOC)(Shary et al., 2002), and Relative elevation (RE)
Regional terrain attribute	Distance to stream (DTS), Local elevation above channel network (LECN)(MacMillan et al., 2000), Elevation below culmination line for mountain areas (EBCLM), Elevation below culmination line for lowland (EBCLL), Relative hillslope position for mountain areas (RHPM), Relative hillslope position for lowlands (RHPL)(Hatfield, 1999), Crest index for mountain areas (CIM), and Crest index for lowlands (CIL)
Combined terrain attribute	Terrain classification index (TCI)
Other	Soil moisture (SM), Diffuse radiation (DFR), and Direct radiation (DRR)

mean annual temperature 8 °C.

The soil class map is a polygon map made by traditional soil survey showing soil landscapes as a specific spatial context of soil types according to the German Soil Classification (Ad-hoc-AG Boden, 2005). There are in total 5 soil classes shown on the soil map in Fig. 1. It was produced on the basis of the German soil mapping regulations. The entire map was processed by one experienced soil surveyor and there were no changes in the taxonomic system during the mapping work. We assume that the overall accuracy is acceptable.

Table 1 lists the 20 environmental features of Ammertal, grouped into four categories. The first three groups are terrain attributes and some other environmental features included in this study area are soil moisture, diffuse radiation and direct radiation in the last group. The resolution of all environmental features is 10 m. The terrain attributes were derived from the DEM with the common terrain analysis tools provided in SAGA GIS 6.0 (Conrad et al., 2015).

The second study area is located in the Raffelson watershed, east of La Crosse County, Wisconsin, USA (Fig. 2). This area is in the so-called “driftless area” of southwestern Wisconsin, which has remained free of direct impact from late Pleistocene era continental glaciers. The area is approximately 4 km^2 with relatively flat narrow ridges and broad flat valleys. The elevation of the study area ranges from about 250 m to 420 m and the slope gradient ranges from 0% to 60%. Most of the mountaintops and valleys are cultivated. The slopes are dominated by woodlands. Only a small part is a pasture that has been transformed by human activities (Cheng et al., 2019; Qi et al., 2008).

There are in total 16 soil classes depicted on the soil map (Fig. 2). This map was generated with expert knowledge and fuzzy logic (Zhu et al., 2001), with an accuracy of 83.8%. The map is a raster map with a resolution of 10 m. In total there were more than 34,760 raster cells.

Many studies on sampling and digital soil mapping have been conducted in this study area. Therefore, based on previous studies, we selected seven environmental features for sampling and spatial prediction (Cheng et al., 2019; Qi and Zhu, 2011; Zhu et al., 2001; Zhu and Mackay, 2001). The selected features are elevation (EL), slope gradient (MS), slope aspect (AS), profile curvature (PRC), planform curvature (PLC), topographic wetness index (TWI) (Qin et al., 2007; Wang et al., 2019), and alluvial composition ratio (ACR). The resolution of all environmental feature is 10 m.

2.3. Sampling designs

For sampling we discretised the polygon map of Ammertal by a fine raster with a resolution of 10 m, which created about 5.5 million raster cells in total for the study area. To reduce the computational complexity and memory requirement for running it on a personal computer, we used stratified simple random sampling without replacement to select 10,000 raster cells from the about 5.5 million raster cells. The five soil classes were used as strata. The sample sizes per stratum were proportional to the total number of raster cells of the soil classes, so that each soil class is equally represented in the stratified random sample. This master sample of 10,000 raster cells was used as the population representing the Ammertal. We believe that such substitute (using the 10,000 raster cells in place of the 5.5 million raster cells) is reasonable. First, 10,000 raster cells are a very large number, and all covariates and soil classes are very well represented in the sample. We believe that there is no effect of pre-selection. Second, both FSCS samples and CLHS samples were selected from this master sample of 10,000 raster cells and we expect that if there were negative effects on the sampling, the effects should be the same to both sampling designs.

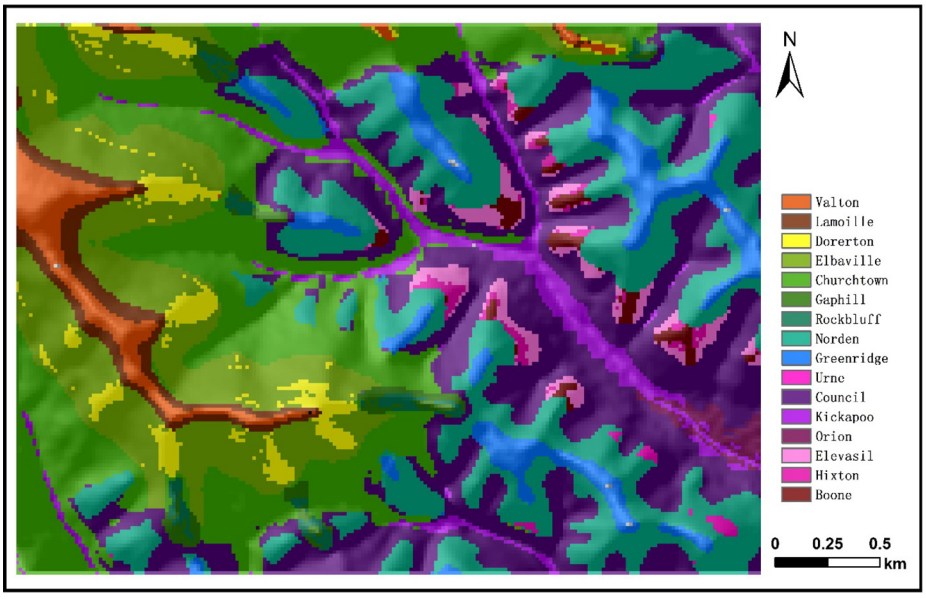

Fig. 2. The detailed soil class map of the Raffelson study area.

For Raffelson we similarily used the above strategy to select a master sample of 10,000 raster cells out of the total 34,760 raster cells, using the sixteen soil classes as strata. These two sample sets were treated as the master samples from which the calibration samples were then selected using the respective strategies: CLHS FSCS and SRS.

2.3.1. Conditioned Latin hypercube sampling (CLHS)

CLHS selects samples in a way that the marginal frequency distributions of the quantitative features (or environmental features) in the population are reproduced by the sample as closely as possible (Minasny and McBratney, 2006). To this end for each feature, marginal strata are constructed using equally spaced quantiles of the marginal distribution, so that all marginal strata contain an equal number of pixels. The number of marginal strata per feature is equal to the sample size. The fitness of a sample is evaluated by the sum over all n^p (n is sample size, p is number of quantitative features) marginal strata of the absolute difference between the marginal stratum sample size and the targeted sample size of 1 raster cell per marginal stratum. This is fitness *O1*. Besides CLHS aims at reproducing the population correlation matrix. The fitness of a sample for this second aim is evaluated by the sum of the absolute difference of the off-diagonal entries in the population and sample correlation matrices (fitness *O3*). Finally, with qualitative features (categorical covariates) CLHS aims at sample sizes per category that are proportional to the area of the categories (*O2* fitness). A weighted average of the three fitnesses is used as a minimization criterion. Commonly, the criterion is minimized by simulated annealing (Minasny and McBratney, 2006). To select samples with this design we used R package spsann, version 2.2.0 (Samuel-Rosa, 2019). The initial temperature was chosen such that in the first chain the acceptance probability was about 90%. The temperature decrease parameter was 0.95. The number of iterations per chain was four times the number of sampling points. The annealing stopped when there were ten successive chains without change of the minimization criterion. We set the argument *clhs.version* of the function *optim.CLHS* to "paper", so that the criterion is computed as those in the paper by Minasny and McBratney (2006).

2.3.2. Feature space coverage sampling (FSCS)

FSCS aims at optimal coverage of the space spanned by the quantitative features in such way that the average distance of the population units (raster cells) to the nearest sampling unit is as small as possible in the feature space as defined by the prescribed set of environmental covariates. By taking the squares of the shortest distance, the criterion can be minimized by the *k*-means clustering algorithm. Brus et al. (2007) proposed this procedure to optimize the coverage of a sample in the geographic space (spatial coverage sampling). In feature space coverage sampling as in this study, the feature values need to be standardized due to the fact that the ranges of these features are largely different from each other. The standardization is done by dividing the values by their standard deviation. The mean squared shortest standardized distance (*MSSSD*) was then used as a minimization criterion (Brus, 2019):

$$MSSSD = \frac{1}{N}\sum_{i=1}^{N} min_j(d_{ij}^2) \quad (1)$$

where N is the number of raster cells in the population and $min_j(d_{ij}^2)$ the minimum of the squared distance between the i^{th} raster cell and all cluster centroids in the standardized feature space. The iterative process is stopped when the *MSSSD* cannot be further reduced.

The *k*-means algorithm is a deterministic procedure, which entails that the final clustering is fully determined by the initial clustering. With an unlucky start (initial clustering), it is possible to end in a sub-optimal clustering (Berkhin, 2006). The risk of a sub-optimal clustering can be reduced in two ways. The first solution is to repeat the clustering many times with different initial clusters, and to save the best clustering (Hartigan and Wong, 2006). The second solution is to select an initial random sample that is closer to the optimal sample than a fully random sample. This is implemented in the *k*-means++ algorithm (Arthur and Vassilvitskii, 2007). The *k*-means++ algorithm consists of two parts, namely the selection of the optimized initial sample and the standard *k*-means. The algorithm is as follows:

(1) Choose one sampling location (raster cell) uniformly at random.
(2) For each raster cell j, compute d_{ij}, the distance in standardized feature space between j and the nearest raster cell i that has already been chosen.
(3) Choose one new raster cell at random as a new sampling location with probabilities proportional to d_{ij}^2.
(4) Repeat Steps 2 and 3 until n centers have been chosen.
(5) Now that the initial centers have been chosen, proceed using standard k-means.

We used the function *kmeanspp* of the R package LICORS (Goerg, 2013) to select FSCS samples with the *k*-means++ algorithm described above. The argument *nstart* for the number of initial clusterings was set to 100.

2.3.3. Simple random sampling (SRS)

Simple random sampling (SRS) without replacement is the simplest random sampling method which selects independently locations from the population with equal probability (Cochran, 1977). It does not exploit any environmental features associated to the soil classes. So by comparing CLHS and FSCS with SRS we can quantify the gain in accuracy due to the use of the features at the sampling stage.

2.4. Prediction methods

2.4.1. Random forest (RF)

Random forest is an ensemble learning method based on decision trees for classification or regression. This method constructs a large number of decision trees, and outputs the class that is the mode of the predicted classes (classification) or predicted mean (regression) of the individual decision trees (Breiman, 2001).

An individual decision tree is built by repeating a binary recursive partitioning of the input training data (model calibration sample). 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 a categorical variable is taken as the mode of the class frequencies in the end nodes of the decision tree.

Subsequently, the bagging technique was introduced to reduce the prediction error variance by establishing an ensemble of classification trees (Breiman, 2001). A large number of trees is created based on bootstrap samples of the training data. All tree predictions are aggregated resulting in a discrete probability distribution of the classes for each unvisited location. The class with the largest probability is used as the predicted class.

We used function *randomForest* of R package randomForest (Liaw and Wiener, 2002) for the calibration of the RF model. Three parameters need to be customized. The first parameter, *ntree*, is the number of decision trees. To reduce computation time, we set *ntree* to a fixed value of 500 as a compromise between accuracy and computational efficiency. The second parameter is *mtry*, which is the number of environmental features randomly selected at each split. It is set to the rounded value after the square root of the total number of environmental features. The third parameter is the minimum terminal node size (*nodesize*). It controls the minimum amount of training data required to continue the tree growth process. This parameter generally takes the default value of 5 (which we used in the case studies).

2.4.2. Individual predictive soil mapping (iPSM)

The basic idea of iPSM is that similar environment conditions will result in similar soils (the Third Law of Geography, Zhu et al., 2018, 2015). The iPSM method, also referred to as similarity-based predictive soil mapping, includes two main steps. The first step is to measure environmental similarity values. Environmental similarity between an unvisited location j and a sampling location i is first evaluated for the individual environmental features, and then similarities based on all environmental features are integrated to represent the overall similarity between an unvisited location j and a sampling location i.

The environmental similarity between an unvisited location j and a sampling location i with regard to the v^{th} environmental feature is calculated by:

$$S_{i,j}^{v} = exp\left[-\frac{(x_i^v - x_j^v)^2}{2 \times \left(\frac{\sigma^v}{\sigma_j^v} \times \sigma^v\right)^2}\right] \quad (2)$$

where x_i^v and x_j^v are the value of the v^{th} environmental feature at sampling location i and unvisited location j, σ^v is the standard deviation of the v^{th} environmental feature in the study area, and σ_j^v is the square root of the mean deviation of the value of v^{th} environmental feature at all unvisited locations ($j = 1, 2,\ldots,k$) from the value at sampling location i, defined as follows:

$$\sigma_j^v = \sqrt{\frac{\sum_{j=1}^{k}(X_i^v - X_j^v)^2}{k}} \quad (3)$$

The overall environmental similarity between an unvisited location j and a sampling location i considering all selected environmental features, $S_{i,j}$ is then determined following a limiting factor approach based on the assumption that the least favorite environmental feature, that is the feature with the smallest similarity value, limits the development of soil at the prediction point to the level of that at the sampling location (c). In this paper, a minimum operator based on the Liebig's law of the minimum (van der Ploeg et al., 1999) are applied to take the minimum value among all environmental similarities (i.e., $S_{i,j}^1, S_{i,j}^2, \ldots, S_{i,j}^v$) as the overall environmental similarity at the soil sample level (Zhu et al., 1997; Shi et al., 2004):

$$S_{i,j} = min(S_{i,j}^1, S_{i,j}^2, \cdots, S_{i,j}^v) \quad (4)$$

The overall environmental similarity between an unvisited location j and each of the n sampling locations is determined using Eq. (2) through Eq. (4).

The second step is to predict the soil class at unvisited location j, based on its environmental similarities to the sampling locations. In this paper, the maximum operator is used for this step. The soil class at the sampling location with the highest similarity to the unvisited location is used as the predicted soil class at the unvisited location (Shi et al., 2004).

2.4.3. Multinomial logistic regression (MLR)

In contrast to RF and iPSM, MLR starts with a statistical model of the data. MLR assume that the data come from a multinomial distribution. It is an extension of binomial logistic regression (BLR) assuming a binomial distribution for the data. In BLR we have two classes only, and the log ratio of the two classes, referred to as the logit, is modelled as a linear combination of the environmental features:

$$\text{logit}(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X}'\boldsymbol{\beta} \quad (5)$$

where $\boldsymbol{X}$ is a vector of environmental features, and $\boldsymbol{\beta}$ is a vector of model coefficients that are typically estimated by maximum likelihood. The Eq. (6) can be rewritten as follows:

$$\frac{\pi}{1-\pi} = exp(\boldsymbol{X}'\boldsymbol{\beta}) = exp(\eta) \quad (6)$$

In MLR one arbitrarily chosen class is taken as a reference (Kempen et al., 2009). For each of the other classes, the log ratio of the probability of that class to the probability of the reference class is modeled as a linear combination of the features. Given the estimated regression coefficients, the probability of any class can be computed by Eq. (7):

$$\pi_i = \frac{exp(\eta_i)}{exp(\eta_1) + exp(\eta_2) + \cdots + exp(\eta_k)} \quad (7)$$

where $\eta_k = 0$. So this model ensures that all probabilities are in the interval [0, 1] and the probabilities sum to 1.

2.5. Evaluation

The performance of sampling designs and their impacts on prediction models were evaluated by the overall accuracy (Brus et al., 2011). This overall accuracy (*OA*) is defined as the population mean of an indicator that has value 1 if the predicted soil class equals the true class, and 0 else:

$$OA = \frac{1}{N}\sum_{i=1}^{N} a_i \quad (8)$$

where N is the total number of population units, and a_i is an indicator defined as follows:

$$a_i = \begin{cases} 1, & \widehat{c_i} = c_i \\ 0, & \widehat{c_i} \neq c_i \end{cases} \quad (9)$$

with $\widehat{c_i}$ the predicted soil class for unit i, and $c_i$the true soil class at that unit.

For Ammertal *OA* was estimated from a validation sample of 10,000 points, which was selected by stratified simple random sampling from the 5.5 million raster cells minus the master sample of 10,000 raster cells used for selecting the calibration samples. This ensures that none of the model calibration sampling locations is included in the validation sample. The validation sample was allocated proportionally to the size of the strata. Consequently, the unweighted sample average is an unbiased estimate of the population *OA*. For Raffelson *OA* was estimated by predicting for a total of 34,760 – 10,000 raster cells that were not used for the selection of calibration samples. Differences in the estimated median of the calibration sampling distribution of *OA* were tested using the Mann-Whitney *U* test (Corder and Foreman, 2014).

2.6. Selection of environmental features for sampling and prediction

Environmental feature selection is an important step both for sampling and prediction. First, the relative importance of all environmental features is computed by RF. The importance is quantified by two measures: mean decrease in accuracy (*MDA*) (Genuer et al., 2010) and mean decrease in Gini (*MDG*) (Calle and Urrea, 2011; Jiang et al., 2009). We chose the set of the environmental features that are of high importance under both *MDA* and *MDG*.

In Ammertal, Relative elevation (RE), Elevation (EL), Diffuse radiation (DFR), Mean slope (MS), Elevation below culmination line for mountain areas (EBCLM), Relative hillslope position for mountain areas (RHPM), Elevation below culmination line for lowland (EBCLL), and Soil moisture (SM) were selected as variables for soil sampling *and* predicting by RF and iPSM. For predicting with MLR on average 4 to 5 features were selected by stepwise regression.

For the Raffelson study area, we selected elevation (EL), slope gradient (MS), slope aspect (AS), planform curvature (PLC), topographic wetness index (TWI), and alluvial composition ratio (ACR) for sampling and predicting using RF and iPSM. For MLR on average 2 ($n = 20$) to 4 ($n = 100$) features were selected by stepwise regression. With the larger sample sizes more predictors become significant,

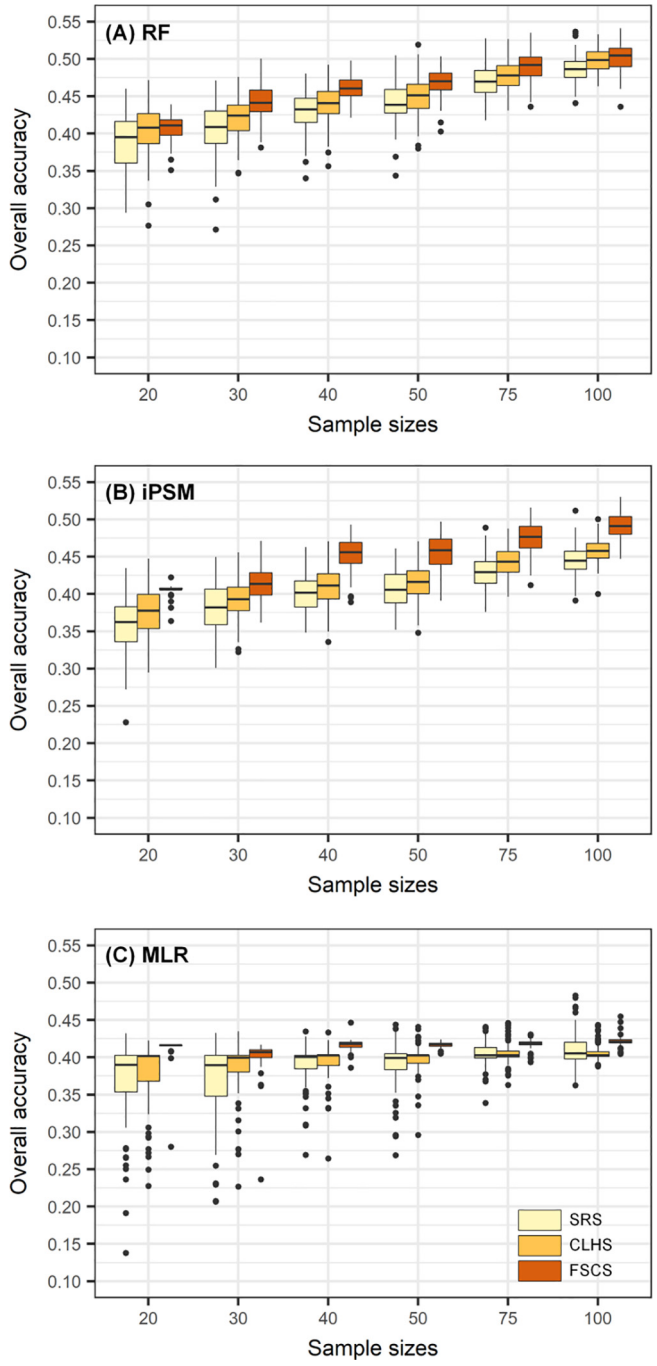

Fig. 3. Boxplots of overall accuracies for 3 * 3 combinations of sampling design and prediction method in Ammertal. (A) RF; (B) iPSM and (C) MLR.

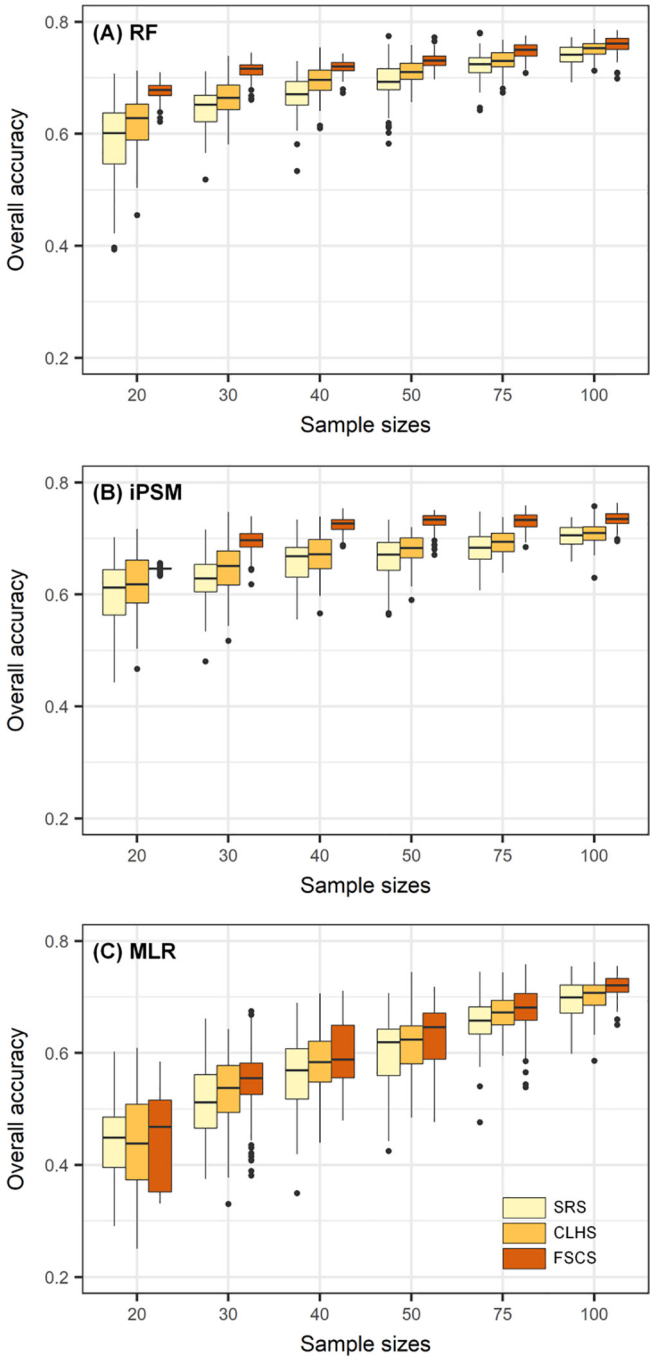

Fig. 4. Boxplots of overall accuracies for 3 * 3 combinations of sampling design and prediction method in Raffelson. (A) RF; (B) iPSM and (C) MLR.

explaining the larger number of selected features.

3. Results

3.1. Overall accuracies

For both study areas, for all sample sizes and all three prediction methods, the order of the three sampling designs in terms of the estimated median or estimated mean of *OA* is the same, which is FSCS, CLHS, and SRS from high to low (Figs. 3 and 4, Tables 2 and 3). The differences between the estimated medians and estimated means were very small, so when we refer to the estimated median hereafter, the statement also holds for the estimated mean. The differences between the estimated medians with FSCS and CLHS were relatively large compared to the difference between the estimated medians with CLHS and SRS. The estimated median with CLHS was in most cases only marginally larger than the estimated median with SRS. For Ammertal, the difference in estimated medians with SRS and CLHS was significant at the level of 0.05 for all sample sizes when using RF and iPSM as a prediction method, but with MLR these differences were not significant (Table 4). Contrarily, for Raffelson only for RF all differences were significant, whereas for iPSM and MLR the estimated medians with CLHS and SRS were not significantly different for quite a few sample sizes. The standard deviation in *OA* among samples decreased with the sample size, in both study areas, for all nine combinations of sampling design and prediction method (Tables 2 and 3). The coefficient of

Table 2
Descriptive statistics (including mean, standard deviation (SD), coefficient of variation (CV), minimum (Min), and maximum (Max)) of the overall accuracy for each sample size in the Ammertal study area. The highest overall accuracy is in boldface in column Mean and Max, while in column SD, CV and Min the lowest overall accuracy is in boldface.

Sizes	Methods	RF					iPSM					MLR				
		Mean	SD	CV	Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV	Min	Max
20	SRS	0.387	0.039	0.100	0.294	0.460	0.361	0.036	0.099	**0.228**	0.435	0.370	0.051	0.139	**0.138**	**0.432**
	CLHS	0.402	0.035	0.086	**0.277**	**0.472**	0.378	0.034	0.090	0.295	**0.447**	0.378	0.040	0.105	0.228	0.422
	FSCS	**0.408**	**0.016**	**0.038**	0.351	0.439	**0.405**	**0.007**	**0.016**	0.364	0.422	**0.414**	**0.014**	**0.033**	0.280	0.416
30	SRS	0.406	0.038	0.093	**0.272**	0.471	0.381	0.032	0.085	**0.301**	0.449	0.369	0.050	0.135	**0.207**	0.433
	CLHS	0.420	0.027	0.064	0.347	0.476	0.394	0.026	0.066	0.322	0.456	0.386	0.033	0.086	0.227	**0.435**
	FSCS	**0.443**	**0.024**	**0.055**	0.381	**0.500**	**0.413**	**0.023**	**0.055**	0.362	**0.471**	**0.402**	**0.020**	**0.049**	0.237	0.417
40	SRS	0.431	0.026	0.060	**0.340**	0.480	0.400	0.025	0.063	0.348	0.463	0.391	0.025	0.064	0.269	0.435
	CLHS	0.440	0.025	0.056	0.356	0.492	0.411	0.027	0.066	**0.336**	0.470	0.395	0.021	0.053	**0.264**	0.433
	FSCS	**0.461**	**0.016**	**0.035**	0.421	**0.498**	**0.453**	**0.023**	**0.050**	0.389	**0.493**	**0.417**	**0.007**	**0.016**	0.386	**0.446**
50	SRS	0.440	0.026	0.058	0.344	0.505	0.408	0.025	0.061	0.352	0.461	0.392	0.030	0.077	**0.269**	**0.444**
	CLHS	0.450	0.026	0.057	0.380	0.519	0.414	**0.022**	0.054	**0.348**	0.471	0.398	0.018	0.046	0.296	0.441
	FSCS	**0.468**	**0.018**	**0.037**	0.403	0.503	**0.455**	0.025	**0.053**	0.391	**0.497**	**0.417**	**0.003**	**0.008**	0.405	0.424
75	SRS	0.470	0.022	0.046	**0.418**	0.528	0.430	0.022	0.052	**0.376**	0.489	0.406	0.016	0.040	**0.339**	0.441
	CLHS	0.478	0.020	0.041	0.431	0.527	0.444	**0.019**	**0.043**	0.396	0.488	0.405	0.014	0.036	0.363	**0.446**
	FSCS	**0.491**	**0.019**	**0.039**	0.436	**0.535**	**0.475**	0.022	0.045	0.412	**0.516**	**0.418**	**0.006**	**0.015**	0.394	0.431
100	SRS	0.488	0.018	0.038	0.441	0.537	0.446	0.020	0.044	**0.392**	0.512	0.411	0.022	0.054	**0.362**	**0.483**
	CLHS	0.498	**0.016**	**0.033**	0.463	0.533	0.458	**0.017**	**0.036**	0.400	0.500	0.407	0.012	0.028	0.388	0.443
	FSCS	**0.503**	0.018	0.036	**0.436**	**0.552**	**0.491**	0.018	0.037	0.447	**0.530**	**0.421**	**0.007**	**0.016**	0.405	0.455

variation, i.e. the standard deviation of *OA* divided by the mean *OA*, was smallest for FSCS for both study areas, all prediction methods and all sample sizes, except for $n = 100$ with RF and iPSM in Ammertal (Tables 2 and 3).

3.2. Relation between minimization criterion and overall accuracy

The *MSSSD* values were clearly the smallest for the FSCS samples in both study areas, which is not surprising because FSCS minimize *MSSSD* (Fig. 5). The clouds of the CLHS samples and SRS samples largely overlap, but the average *MSSSD* of the CLHS samples was somewhat smaller than that of the SRS samples.

The values of *O1 + O3* are smallest for the CLHS samples in both study areas (Fig. 6), which is again not surprising because CLHS optimizes on *O1 + O3*.

There are weak but significant negative linear relationships between *OA* from the RF predictions and the minimized *MSSSD* for both study areas (Fig. 5, Table 5 and Table 6). The results from iPSM and MLR were similar. For Ammertal the coefficient of determination, R^2, of the simple linear models fitted on the minimized *MSSSD* and *OA* for all samples of a given size, varied from 7.11% ($n = 20$) to 23.02% ($n = 40$) (Table 5). The estimated residual standard deviation decreased with the sample size from 0.031 to 0.018. For Raffelson the linear relation between the minimized *MSSSD* and *OA* was stronger (Table 6). R^2 varied from to 17.61% ($n = 100$) to 48.52% ($n = 30$). The residual standard deviation deceased with the sample size from 0.044 to 0.016.

There are almost no linear relationships between the minimized *O1 + O3* criterion and *OA* as obtained with the prediction results from RF as shown by the nearly horizontal lines with regression slopes close to zero (Fig. 6, Table 5 and Table 6). The two significant slopes (Ammertal, n = 40 and Raffelson, n = 30) were even positive, wherad these should be negative. The results from iPSM and MLR were similar. The coefficient of determination was about zero for all sample sizes (Table 5 and Table 6), and the residual standard deviations were somewhat larger compared with those of the models for *OA* with *MSSSD* as an independent variable.

The above results and analysis show that the minimized *MSSSD*

Table 3
Descriptive statistics (including mean, standard deviation (SD), coefficient of variation (CV), minimum (Min), and maximum (Max)) of the overall accuracy for each sample size in the Raffelson study area. The highest overall accuracy is in boldface in column Mean and Max, while in column SD, CV and Min the lowest overall accuracy is in boldface.

Sizes	Methods	RF					iPSM					MLR				
		Mean	SD	CV	Min	Max	Mean	SD	CV	Min	Max	Mean	SD	CV	Min	Max
20	SRS	0.589	0.066	0.112	**0.393**	0.707	0.601	0.060	0.100	**0.443**	0.703	0.440	**0.078**	**0.177**	**0.181**	0.603
	CLHS	0.619	0.045	0.073	0.455	**0.713**	0.619	0.052	0.085	0.467	**0.717**	0.437	0.082	0.187	0.251	**0.609**
	FSCS	**0.676**	**0.016**	**0.023**	0.622	0.710	**0.647**	**0.004**	**0.007**	0.634	0.656	**0.442**	0.083	0.187	0.331	0.585
30	SRS	0.643	0.039	0.060	**0.518**	0.712	0.627	0.042	0.068	**0.481**	0.716	0.512	0.069	0.135	0.375	0.662
	CLHS	0.665	0.035	0.052	0.581	0.739	0.648	0.042	0.064	0.517	**0.747**	0.529	**0.063**	0.119	**0.331**	0.643
	FSCS	**0.713**	**0.017**	**0.023**	0.661	**0.745**	**0.695**	**0.021**	**0.030**	0.618	0.740	**0.550**	0.064	**0.117**	0.382	**0.675**
40	SRS	0.670	0.034	0.051	**0.533**	0.730	0.658	0.041	0.063	**0.555**	0.734	0.564	0.066	0.117	**0.350**	0.689
	CLHS	0.695	0.028	0.040	0.610	**0.754**	0.669	0.034	0.051	0.567	0.739	0.583	**0.058**	**0.100**	0.440	0.707
	FSCS	**0.719**	**0.013**	**0.018**	0.673	0.743	**0.725**	**0.014**	**0.019**	0.686	**0.754**	**0.600**	0.061	0.102	0.479	**0.711**
50	SRS	0.692	0.035	0.050	**0.583**	**0.775**	0.666	0.037	0.056	**0.564**	0.733	0.602	0.061	0.102	**0.425**	0.707
	CLHS	0.711	0.020	0.029	0.656	0.758	0.681	0.026	0.038	0.590	0.721	0.614	**0.056**	**0.090**	0.485	**0.744**
	FSCS	**0.731**	**0.013**	**0.018**	0.697	0.772	**0.730**	**0.016**	**0.022**	0.671	**0.751**	**0.630**	0.059	0.093	0.477	0.718
75	SRS	0.721	0.025	0.034	**0.642**	**0.781**	0.682	0.029	0.042	**0.607**	0.748	0.656	0.041	0.062	**0.417**	0.745
	CLHS	0.731	0.018	0.025	0.674	0.768	0.693	0.022	0.032	0.639	0.738	0.673	**0.031**	**0.047**	0.595	0.744
	FSCS	**0.748**	**0.015**	**0.020**	0.709	0.775	**0.730**	**0.016**	**0.021**	0.685	**0.759**	**0.677**	0.042	0.062	0.539	**0.758**
100	SRS	0.741	0.018	0.024	**0.692**	0.772	0.703	0.020	0.028	0.659	0.738	0.696	0.033	0.048	0.598	0.755
	CLHS	0.752	**0.014**	**0.018**	0.713	**0.787**	0.708	0.020	0.029	**0.630**	0.758	0.702	0.030	0.042	**0.586**	**0.762**
	FSCS	**0.759**	0.017	0.022	0.699	0.785	**0.735**	**0.014**	**0.019**	0.695	**0.764**	**0.719**	**0.022**	**0.030**	0.651	0.756

Table 4
Results of Mann-Whitney *U* test for differences in estimated median overall accuracy with various types of sampling design, prediction methods, and sample sizes. Same letters within rows indicate non-significant differences at significance level α of 0.05.

Sizes	Ammertal									Raffelson								
	RF			iPSM			MLR			RF			iPSM			MLR		
	SRS	CLHS	FSCS	SRS	CLHS	FSCS	SRS	CLHS	FSCS	SRS	CLHS	FSCS	SRS	CLHS	FSCS	SRS	CLHS	FSCS
20	a	b	c	a	b	c	a	a	b	a	b	c	a	a	b	a	a	a
30	a	b	c	a	b	c	a	a	b	a	b	c	a	b	c	a	a	b
40	a	b	c	a	b	c	a	a	b	a	b	c	a	a	b	a	a	a
50	a	b	c	a	b	c	a	a	b	a	b	c	a	b	c	a	a	b
75	a	b	c	a	b	c	a	a	b	a	b	c	a	b	c	a	b	b
100	a	b	c	a	b	c	a	a	b	a	b	c	a	a	b	a	a	b

values have stronger linear relationships with *OA* than the minimized *O1 + O3* values do. This indicates that *MSSSD* is a more effective minimization criterion than *O1 + O3*. This contributes to the fact that FSCS behaves better than CLHS because FSCS is based on *MSSSD*.

4. Discussion

The most striking results of this research are the smaller mean and median *OA* and the smaller variation (expressed as standard deviation

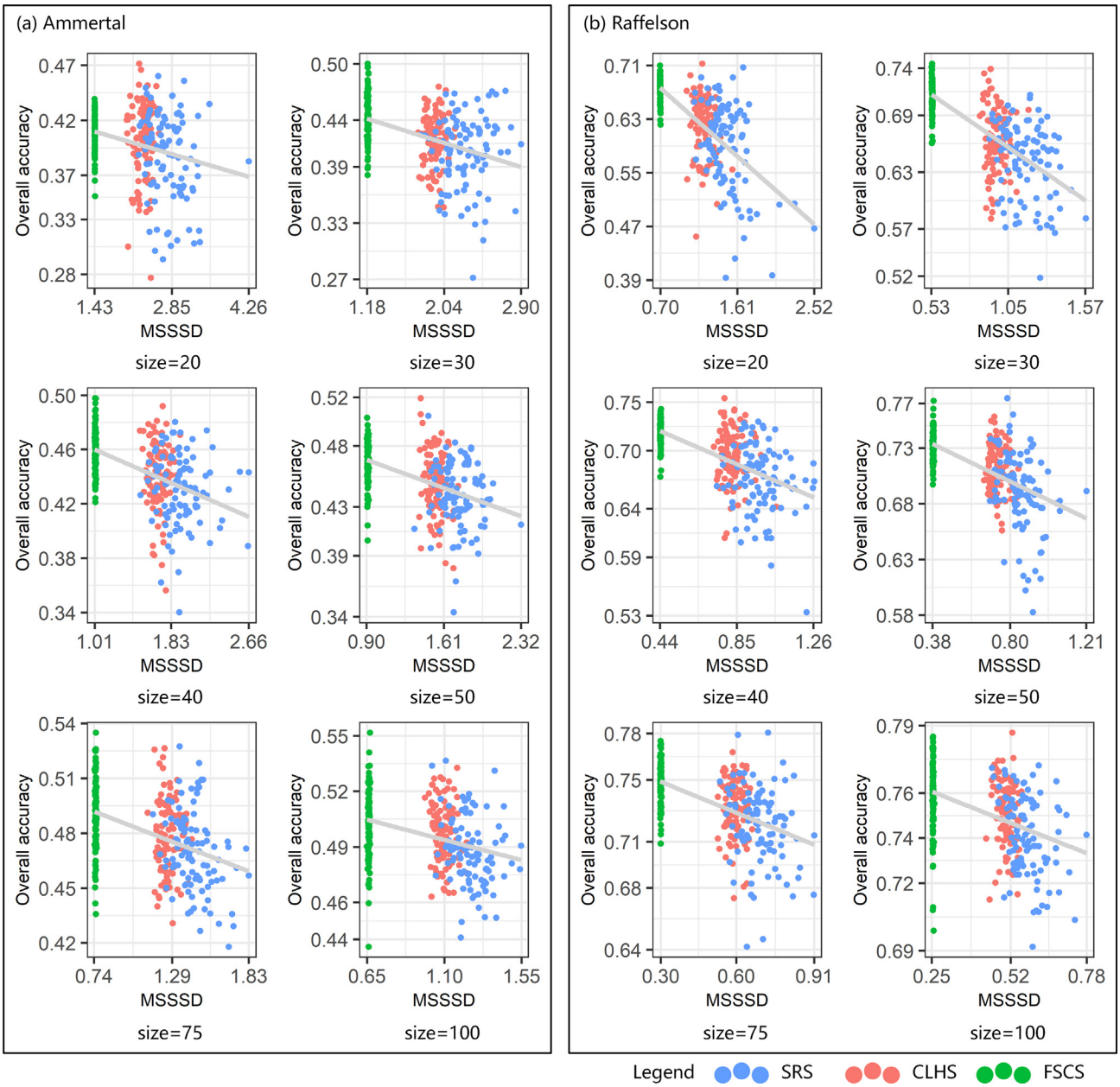

Fig. 5. The relationship between minimization criterion (*O1 + O3* and *MSSSD*) and overall accuracy produced by RF for each sample size in the Ammertal study area.

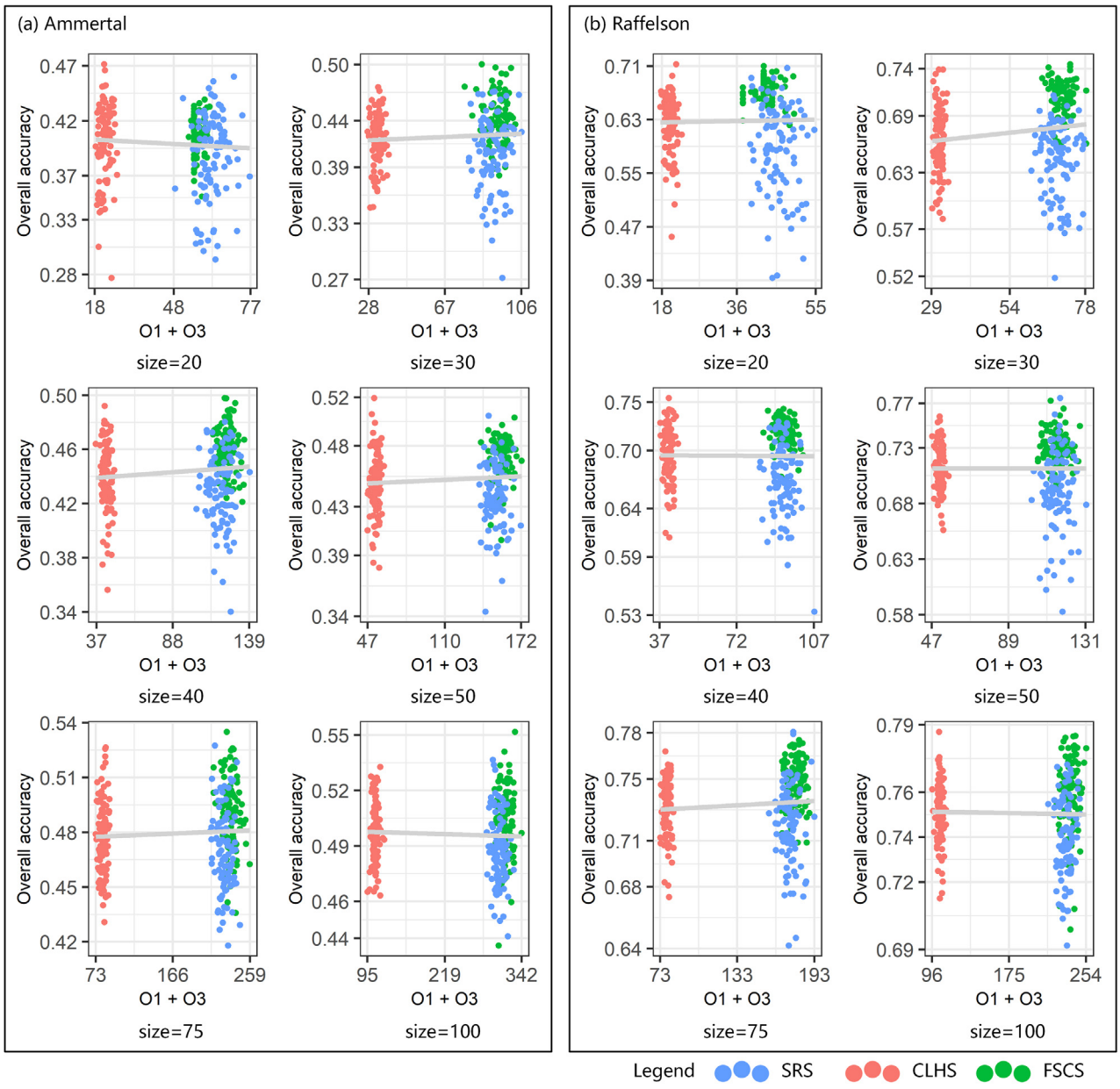

Fig. 6. The relationship between minimization criterion (*O1 + O3* and *MSSSD*) and overall accuracy as obtained with RF prediction, for each sample size in the Raffelson study area.

and coefficient of variation) of *OA* over repeated calibration samples for FSCS compared to those for CLHS. Apparently, the spreading of the calibration samples in multivariate feature space through minimization of the *MSSSD* criterion is better compared to the spreading in feature space through minimization of a weighted combination of the *O1* and *O3* criterion. A possible reason for this is that in CLHS the spreading is largely enforced through spreading along the separate, univariate axes spanning the multivariate space, and not directly in multivariate space

Table 5
Quality and estimated slopes of the simple linear regression models for the overall accuracy as obtained with RF prediction (*OA*), using the minimization criterion (*O1 + O3* or *MSSSD*) as an independent variable, for each sample size in the Ammertal study area.

Size	Residual standard deviation		R^2		Slope		Pearson's *r*	
	O1 + O3	*MSSSD*	*O1 + O3*	*MSSSD*	*O1 + O3*	*MSSSD*	*O1 + O3*	*MSSSD*
20	0.032	0.031	0.0055	0.0711	-1.34×10^{-4}	-0.0145 ***	-0.0743	-0.2668 ***
30	0.034	0.031	0.0058	0.1775	8.96×10^{-5}	-0.0299 ***	0.0762	-0.4213 ***
40	0.025	0.023	0.0137	0.2302	8.10×10^{-5} *	-0.0300 ***	0.1170 *	-0.4798 ***
50	0.026	0.023	0.0072	0.2112	4.67×10^{-5}	-0.0326 ***	0.0851	-0.4596 ***
75	0.022	0.019	0.0038	0.1730	1.90×10^{-5}	-0.0300 ***	0.0620	-0.4159 ***
100	0.018	0.018	0.0027	0.1140	-1.02×10^{-5}	-0.0244 ***	-0.0521	-0.3377 ***

Slopes and Pearson's *r* significant at 0.001, 0.01, 0.05 levels are marked by ***, **, *, respectively

Table 6
Quality and estimated slopes of the simple linear regression models for the overall accuracy as obtained with RF prediction (*OA*), using the minimization criterion (*O1 + O3* or *MSSSD*) as an independent variable in the Raffelson study area.

Size	Residual standard deviation		R^2		Slope		Pearson's *r*	
	O1 + O3	*MSSSD*	*O1 + O3*	*MSSSD*	*O1 + O3*	*MSSSD*	*O1 + O3*	*MSSSD*
20	0.059	0.044	4.63×10^{-4}	0.4532	1.07×10^{-4}	-0.1116^{***}	0.0215	-0.6732^{***}
30	0.042	0.031	2.35×10^{-2}	0.4852	$3.60 \times 10^{-4}{}^{**}$	-0.1077^{***}	0.1535^{**}	-0.6965^{***}
40	0.033	0.027	6.91×10^{-5}	0.3540	-1.10×10^{-5}	-0.0832^{***}	−0.0083	-0.5950^{***}
50	0.029	0.024	3.17×10^{-9}	0.3304	-5.38×10^{-8}	-0.0804^{***}	−0.0001	-0.5748^{***}
75	0.023	0.020	9.08×10^{-3}	0.2657	4.65×10^{-5}	-0.0669^{***}	0.0953	-0.5155^{***}
100	0.018	0.016	7.81×10^{-4}	0.1761	-7.93×10^{-6}	-0.0518^{***}	−0.0279	-0.4196^{***}

Slopes and Pearson's *r* significant at 0.001, 0.01, 0.05 levels are marked by ***, **, *, respectively.

as in FSCS. Besides, the *O1* criterion is a rather rough criterion for enforcing the spreading in univariate space. A stratified sample with points close to the boundaries of the marginal strata of some feature (so the values are close to the equally spaced quantiles of that feature) is not as well-spread as a stratified sample with values half-way the stratum boundaries, but both samples have the same value for *O1*.

The large variation in *OA* with CLHS can possibly be explained by the random selection of points within the strata (combinations of the marginal strata). Given the strata, multiple samples may exist for the same value of *O1* and nearly for the same value of *O3*. In that case the CLHS algorithm does not have a clear preference for any of these samples: the selection probabilities of these samples are about equal. However, the spreading of these samples along the axes spanning the multivariate space can differ substantially, which may lead to large differences in the calibrated models, and as a consequence a large variation in *OA*. The hypothesis that the same value of *O1 + O3* can lead to a largely different *OA* is supported by the vertical series of points for FSCS samples in various scatter plots of *OA* against of *O1 + O3* (Figs. 5 and 6).

The *O1 + O3* value of the optimized CLHS samples shows strong variation, indicating that for most samples the global optimum was not reached. However, as there is no relation between *O1 + O3* and *OA* we do not think that the performance of CLHS can be further improved by increasing the values of the simulated annealing parameters and/or using multiple starts (initial samples).

The relatively small variation of *OA* with FSCS compared to that with CLHS can be explained by the small variation in *MSSSD* among FSCS samples. As there is a relation between *MSSSD* and *OA*, a small variation in *MSSSD* is a necessary condition, but not a sufficient condition. A constant *MSSSD* is no guarantee for a constant *OA*. This depends on how unique the sample with the minimized *MSSSD* value is. Various samples may exist with exactly or nearly the same minimal *MSSSD* value but with rather different *OA* values. Apparently, the difference in spreading in feature space among these FSCS samples with the same *MSSSD* value is considerably smaller than that of CLHS samples with the same *O1 + O3* value.

In practice we do not have the budget to select many calibration samples with the sampling designs for comparison so that we lack information about the distribution of *OA* over repeated calibration sampling. Besides, for a given map the *OA* is not known, but must be estimated, either by cross-validation, or preferably from a probability sample of limited size. In that case there will be an error in the estimated *OA* that cannot be ignored. Thus, in real-world case studies it is impossible to draw general conclusions about the relative performance of the calibration sampling designs. This study gives an idea of the variation in *OA* that can be expected if the calibration sampling would be repeated. If one accepts the premise that a type of sampling design, which can provide digital soil mapping with *OA* values which are high with a small variation, would be a good sampling design, then we could reach a conclusion in these study areas FSCS is a better and more efficient sampling design than both CLHS and SRS.

Although our interest is the differences among the calibration sampling design types, we note that in Raffelson the *OA*s were considerably larger than in Ammertal, despite that in Raffelson the number of soil classes was much larger than in Ammertal: sixteen in Raffelson and only five in Ammertal. This relatively high accuracy of the predicted soil maps of Raffelson can possibly be explained by the procedure that is used to construct the soil class map. This map was constructed with expert knowledge and fuzzy logic in the SoLIM approach (Zhu et al., 2001), using partly the same environmental features as we used here for sampling and prediction. On the contrary the soil class map of Ammertal was made using the traditional soil survey approach.

5. Conclusions

This study compared conditioned Latin hypercube sampling (CLHS), feature space coverage sampling (FSCS) and simple random sampling (SRS) in a simulation study using the legacy soil class maps of two study areas as references. The comparison of the three sampling design types was based on the values and the variation of the overall accuracy of predicted soil classes using the samples from these three sampling design types. The models used to predict the soil classes are random forest (RF), individual predictive soil mapping (iPSM) and multinomial logistic regression (MLR). We conclude that:

- In both study areas the median overall accuracy with FSCS was higher than those with CLHS and SRS over all sample sizes and cross all three prediction methods. The median overall accuracy with CLHS was only marginally larger than with SRS.
- There was a significant negative correlation between *MSSSD* and overall accuracy, whereas no such correlation was found between *O1 + O3* and overall accuracy.
- The coefficient of variation in overall accuracy among samples selected using FSCS was smaller than these using CLHS and SRS at the same sample sizes.
- With CLHS the variation in overall accuracy among samples was large, so that there is a serious risk that a particular sample might lead to a low overall accuracy.
- FSCS-RF is the most accurate combination of sampling and prediction for both study areas.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work reported here was supported by grants from National Natural Science Foundation of China (Project No.: 41871300, 41431177), National Basic Research Program of China (Project No.: 2015CB954102), PAPD, and Outstanding Innovation Team in Colleges and Universities in Jiangsu Province. Supports to A-Xing Zhu through

the Vilas Associate Award, the Hammel Faculty Fellow Award, and the Manasse Chair Professorship from the University of Wisconsin-Madison are greatly appreciated. Thomas Scholten thanks the German Research Foundation (DFG) for supporting this research through the DFG Cluster of Excellence “Machine Learning - New Perspectives for Science”, EXC 2064/1, project number 390727645. We also thank the Landesamt für Geologie, Rohstoffe und Bergbau Baden-Württemberg, Freiburg, Germany, for providing the Ammertal data.

References

Ad-hoc-AG Boden, 2005. Bodenkundliche Kartieranleitung. KA5, 5, verb. und, erw. Aufl. Schweizerbart, Stuttgart, Germany.

Arthur, D., Vassilvitskii, S., 2007. K-Means + +: the Advantages of Careful Seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, pp. 1027–1035. https://doi.org/10.1145/1283383.1283494.

Berkhin, P., 2006. A survey of clustering data mining techniques, in: Grouping Multidimensional Data: Recent Advances in Clustering. pp. 25–71. https://doi.org/10.1007/3-540-28349-8_2.

Breiman, L., 2001. Random Forest. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464–480. https://doi.org/10.1016/j.geoderma.2018.07.036.

Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2007. Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. Developments in Soil Science. Elsevier, pp. 183–192. https://doi.org/10.1016/S0166-2481(06)31014-8.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62, 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x.

Calle, M.L., Urrea, V., 2011. Letter to the editor: Stability of Random Forest importance measures. Briefings Bioinf. 12, 86–89. https://doi.org/10.1093/bib/bbq011.

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v.2.1.4. Geosci. Model Dev. 8, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015.

Cook, S.E., Jarvis, A., Gonzalez, J.P., 2008. In: Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 31–41. https://doi.org/10.1007/978-1-4020-8592-5_3.

Corder, G.W., Foreman, D.I., 2014. Nonparametric Statistics: A Step-by-Step. Approach. Wiley.

Dietrich, W.E.Montgomery, D.R. A digital terrain model for mapping shallow landslide potential. http://socrates.berkeley.edu/~geomorph/shalstab.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31, 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Goerg, G.M., 2013. LICORS: Light Cone Reconstruction of States - Predictive State Estimation. From Spatio-Temporal Data.

Grathwohl, P., Rügner, H., Wöhling, T., Osenbrück, K., Schwientek, M., Gayler, S., Wollschläger, U., Selle, B., Pause, M., Delfs, J.O., Grzeschik, M., Weller, U., Ivanov, M., Cirpka, O.A., Maier, U., Kuch, B., Nowak, W., Wulfmeyer, V., Warrach-Sagi, K., Streck, T., Attinger, S., Bilke, L., Dietrich, P., Fleckenstein, J.H., Kalbacher, T., Kolditz, O., Rink, K., Samaniego, L., Vogel, H.J., Werban, U., Teutsch, G., 2013. Catchments as reactors: a comprehensive approach for water fluxes and solute turnover. Environ. Earth Sci. 69, 317–333. https://doi.org/10.1007/s12665-013-2281-7.

Hartigan, J.A., Wong, M.A., 2006. Algorithm AS 136: A K-Means Clustering Algorithm. Appl. Stat. 28, 100. https://doi.org/10.2307/2346830.

Hatfield, D.C., 1999. TopoTools: A Collection of Topographic Modeling Tools for ArcINFO. http://www.giscafe.com/GISVision/TechPaper/TopoTools.html.

Horn, B.K.P., 1981. Hill Shading and the Reflectance Map. Proc. IEEE 69, 14–47. https://doi.org/10.1109/PROC.1981.11918.

Cheng, W., Zhu, A.X., Qin, C.Z., Qi, F., 2019. Updating conventional soil maps by mining soil–environment relationships from individual soil polygons. J. Integr. Agric. 18, 265–278. https://doi.org/10.1016/S2095-3119(18)61938-0.

Jiang, R., Tang, W., Wu, X., Fu, W., 2009. A random forest approach to the detection of epistatic interactions in case-control studies, in: BMC Bioinformatics. https://doi.org/10.1186/1471-2105-10-S1-S65.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311–326. https://doi.org/10.1016/j.geoderma.2009.04.023.

Lagacherie, P., 2008. Digital Soil Mapping: A State of the Art. In: Hartemink, A.E., McBratney, A., Mendonça-Santosde, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 3–14. https://doi.org/10.1007/978-1-4020-8592-5_1.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Lin, Y.P., Chu, H.-J., Huang, Y.L., Tang, C.H., Rouhani, S., 2011. Monitoring and identification of spatiotemporal landscape changes in multiple remote sensing images by using a stratified conditional Latin hypercube sampling approach and geostatistical simulation. Environ. Monit. Assess. 177, 353–373. https://doi.org/10.1007/s10661-010-1639-5.

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

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113, 81–109. https://doi.org/10.1016/S0165-0114(99)00014-7.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

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

Qi, F., Pei, T., Zhu, A.X., Qin, C., Burt, J.E., 2008. Knowledge discovery from area-class resource maps: capturing prototype effects. Cartogr. Geogr. Info. Sci. 35, 223–237. https://doi.org/10.1559/152304008786140533.

Qi, F., Zhu, A.X., 2011. Comparing three methods for modeling the uncertainty in knowledge discovery from area-class soil maps. Comput. Geosci. 37, 1425–1436. https://doi.org/10.1016/j.cageo.2010.10.016.

Qin, C., Zhu, A., Pei, T., Li, B., Zhou, C., Yang, L., 2007. An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. Int. J. Geogr. Info. Sci. 21, 443–458. https://doi.org/10.1080/13658810601073240.

Roudier, P., Hewitt, A., Beaudette, D., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Digital Soil Assessments and Beyond. Proceedings of the Fifth Global Workshop on Digital Soil Mapping, pp. 227–231. https://doi.org/10.1201/b12728-46.

Samuel-Rosa, A., 2019. spsann: Optimization of Sample Configurations using Spatial Simulated Annealing. R package version 2.2.0.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M. de L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital Soil Map of the World. Science 325, 680–681. https://doi.org/10.1126/science.1175084.

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 232–234, 243–256. https://doi.org/10.1016/j.geoderma.2014.05.013.

Shary, P.A., Sharaya, L.S., Mitusov, A.V., 2002. Fundamental quantitative methods of land surface analysis. Geoderma 107, 1–32. https://doi.org/10.1016/S0016-7061(01)00136-7.

Shi, X., Zhu, A.X., Burt, J.E., Qi, F., Simonson, D., 2004. A Case-based Reasoning Approach to Fuzzy Soil Mapping. Soil Sci. Soc. Am. J. 68, 885–894. https://doi.org/10.2136/sssaj2004.8850.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., Scholten, T., 2016. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. J. Plant Nutr. Soil Sci. 179, 499–509. https://doi.org/10.1002/jpln.201500313.

Tarboton, D.G., 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resour. Res. 33, 309–319. https://doi.org/10.1029/96WR03137.

van der Ploeg, R.R., Böhm, W., Kirkham, M.B., 1999. On the Origin of the Theory of Mineral Nutrition of Plants and the Law of the Minimum. Soil Sci. Soc. Am. J. 63, 1055–1062. https://doi.org/10.2136/sssaj1999.6351055x.

Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B.M., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913. https://doi.org/10.1016/j.geoderma.2019.113913.

Wang, Y.J., Qin, C.Z., Zhu, A.X., 2019. Review on algorithms of dealing with depressions in grid DEM. Ann. Gis 25, 83–97. https://doi.org/10.1080/19475683.2019.1604571.

Webster, R., Oliver, M.A., 1990. Statistical methods in soil and land resource survey. Oxford University Press, Oxford.

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. https://doi.org/10.5849/forsci.11-084.

Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B., Chen, Z., Zhu, A.X., Zhou, C., 2020. Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. Geoderma 369, 114337. https://doi.org/10.1016/j.geoderma.2020.114337.

Zhu, A.X., Band, L., Vertessy, R., Dutton, B., 1997. Derivation of Soil Properties Using a Soil Land Inference Model (SoLIM). Soil Sci. Soc. Am. J. 61, 523–533. https://doi.org/10.2136/sssaj1997.03615995006100020022x.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. Soil Sci. Soc. Am. J. 65, 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66, 535–547. https://doi.org/10.1111/ejss.12244.

Zhu, A.X., Lu, G., Liu, J., Qin, C.Z., Zhou, C., 2018. Spatial prediction based on Third Law of Geography. Ann. Gis 24, 225–240. https://doi.org/10.1080/19475683.2018.1534890.

Zhu, A.X., Mackay, S.D., 2001. Effects of spatial detail of soil information on watershed modeling. J. Hydrol. 248, 54–77. https://doi.org/10.1016/S0022-1694(01)00390-0.