Geoderma 253–254 (2015) 111–121

Contents lists available at ScienceDirect

Geoderma

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

Balanced sampling: A versatile sampling approach for statistical soil surveys

D.J. Brus

Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 18 December 2014
Received in revised form 16 March 2015
Accepted 8 April 2015
Available online xxxx

Keywords:
Probability sampling
Latin hypercube sampling
Digital soil mapping
Design-based estimation
Random forests

ABSTRACT

In balanced sampling a linear relation between the soil property of interest and one or more covariates with known means is exploited in selecting the sampling locations. Recent developments make this sampling design attractive for statistical soil surveys. This paper introduces balanced sampling and demonstrates its potential utility and versatility. Latin hypercube sampling appears to be a special case of balanced sampling. When implemented as a balanced sampling design, the inclusion probabilities of the population units are known. Population parameters can then be estimated by design-based, model-assisted or model-based inference. In a simulation study balanced (b) random sampling, balanced coverage (bc) random sampling, and latin hypercube (lh) random sampling were compared in terms of the sampling distributions of number of unsampled marginal strata (*U*) measuring coverage of feature space, Mean Squared Shortest Distance (*MSSD*) measuring spatial coverage, and error in the estimated mean *e*. In designs b and bc four covariates were used as balancing variables. In bc the four covariates and the spatial coordinates were used as spreading variables. With lh the total sample size was random, but the size fluctuations were acceptable. Design lh clearly scored best with regard to *U*, but had by far the largest variance of *e*. Based on *U* and *MSSD* design bc outperformed design b, which can be explained by the use of spreading variables in bc. The variance of *e* for these designs was about 0.8 times the variance for simple random sampling. Using the ratio estimator for lh this variance ratio was about 1.8, showing the poor estimation performance of lh.

1. Introduction

The important role of sampling design in soil survey cannot be overemphasized. The sampling design largely determines the quality of the survey result. Also, problems with statistical inference can often be avoided by a proper sampling design. Not merely the number of sampling locations, also the spatial coordinates of the sampling locations are of importance.

In the past decades various sampling approaches have been developed for mapping natural resources. At a high level these sampling approaches can be divided into designs with optimized coverage of geographic space, feature space or both. Optimization of coverage of geographic space leads to spatial coverage sampling, also referred to as spatially balanced sampling. Quite a few sampling designs/algorithms have been proposed for designing such samples, among which are spatial simulated annealing (van Groenigen and Stein, 1998), k-means (Brus et al., 1999), Generalized Random-Tesselation Stratified sampling (Stevens and Olson, 2004), the local pivotal algorithm (Grafström et al., 2012) and balanced acceptance sampling (Robertson et al., 2013).

For optimization of coverage of the space spanned by the covariates, referred to as feature or attribute space, I would like to mention response surface sampling (Lesch et al., 1995; Lesch, 2005) and constrained latin hypercube sampling (Minasny and McBratney, 2006). The first design is of specific interest for mapping using linear models (linear regression models, linear mixed models), the second for mapping with classification and regression tree (CART) or random forest models. Both designs are adaptations of experimental designs for observational studies.

For optimization of coverage of both geographic and feature space, Brus and Heuvelink (2007) proposed a model-based sampling design. In this approach a sample is searched for with minimal error variance of predictions obtained by kriging with an external drift. This variance has two components, that of the error in the estimated trend and that of the interpolation error. By minimizing this variance a balance between spreading in feature space and geographic space is automatically obtained.

In this paper I introduce a sampling approach unnoticed so far by soil scientist: balanced sampling. Despite its long successful history in other branches of research such as socio-economic research, I am not aware of applications of balanced sampling in soil survey. There are several recent developments in balanced sampling which makes this sampling design of even more interest for statistical soil surveys. First, Falorsi and Righi (2008) showed how multi-way stratified designs can be implemented as a balanced design. As a latin hypercube sample is a special case of a multi-way stratified sample, balanced sampling is an alternative for the sampling algorithm proposed by Minasny and McBratney (2006). Second, recently an adaptation of the algorithm has been developed improving the spread of the balanced samples in geographic and/

E-mail address: dick.brus@wur.nl.

http://dx.doi.org/10.1016/j.geoderma.2015.04.009
0016-7061/© 2015 Elsevier B.V. All rights reserved.

or feature space (Grafström and Tillé, 2013). Finally, a fast algorithm for designing balanced samples has been developed (Deville and Tillé, 2004). This algorithm has been implemented in the R packages sampling available at http://CRAN.R-project.org/ (Tillé and Matei, 2012) and BalancedSampling, available at http://www.antongrafstrom.se/balancedsampling/, so that implementation of balanced designs in practice has become feasible.

The main advantage of latin hypercube sampling implemented as balanced sampling over the constrained latin hypercube sampling approach proposed by Minasny and McBratney (2006) is that the generated samples are probability samples, i.e., random samples with known inclusion probabilities larger than 0 for all population units. In the selection procedure proposed by Minasny and McBratney (2006) there is also randomness, but I am not aware of publications showing that all units have a positive probability of being selected, and how these inclusion probabilities can be computed. As a consequence, contrary to constrained latin hypercube samples selected by the approach of Minasny and McBratney (2006), latin hypercube *random* samples proposed in this paper can be used for design-based or model-assisted inference (Särndal et al., 1992), such as estimation of spatial means or totals. Work on latin hypercube sampling with known inclusion probabilities for non-rectangular regions has been done before, see for instance Hung et al. (2010). The suitability of this design for observational studies as in digital soil mapping still must be explored.

By generating samples that are suitable both for mapping and for design-based estimation of means or totals, we have increased flexibility. An example of where this flexibility can be advantageous is, for instance, soil organic carbon (SOC) mapping. In many cases we do not only want maps of SOC, but also an estimate of the SOC stocks in the study area, or the SOC stocks in several subareas such as land use units or soil map units. For estimation of these stocks model-free, design-based statistical inference can be advantageous as then the quality of the estimated mean (total) and its variance is independent of the quality of a spatial model (Brus and de Gruijter, 1997).

The aim of this paper is to introduce balanced sampling to the soil science community, and to demonstrate its potential utility and versatility. I first explain what balanced sampling is, how the “cube” algorithm and the adaptation of this algorithm as proposed by Grafström and Tillé (2013) for selecting balanced sampling work, and how latin hypercube random samples can be selected by balanced sampling. Then balanced sampling is demonstrated in a simulation study on SOC-mapping and estimation of SOC stocks in three woredas of Ethiopia.

2. Balanced sampling

The idea of balanced sampling is that if we know the mean of some covariate that is linearly related to the variable of interest, then it will be efficient to select a sample for which the average of the covariate equals the population mean. In the context of probability sampling, a sampling design is balanced on variable x when

$$\sum_{k=1}^{n}\frac{x_k}{\pi_k}=\sum_{k=1}^{N}x_k \tag{1}$$

for all samples that can be drawn with the sampling design (Deville and Tillé, 2004). In this equation n is the sample size, π_k is the inclusion probability of unit k, and N is the total number of units in the population. In words, for a balanced sampling design for *all possible* samples the Horvitz–Thompson estimate of the total of the variable used in balancing the sample (shortly referred to as the balancing variable hereafter) equals the true total. In other words, the sampling variance of the estimated total of x equals zero. Notice that, for instance, with simple random sampling, only the *expectation* of the Horvitz–Thompson estimator of the total of the balancing variable equals the true total, i.e., the estimator is p-unbiased, but for individual samples the estimated total will differ from the true total.

Balanced sampling is illustrated now with a simple example. The top-left subfigure in Fig. 1 shows a sample balanced on the variable Easting. The average of the Easting coordinate equals the population mean, which is 10. The simulated values show a clear spatial trend from West to East. By selecting samples that are balanced on the Easting coordinate, a more precise estimate of the mean of the variable of interest is obtained compared to simple random sampling. Fig. 2 shows the squared error in the estimated mean plotted against the difference between the sample average and the population mean of Easting for 100 simple random samples of size four. The larger the absolute value of this difference, the larger the squared error. In this simple example the sampling variance of the estimated mean for balanced sampling equals 10.6, whereas this variance for simple random sampling is 39.7.

The simulated field of Fig. 1 also shows a slight trend from South to North. We may therefore balance the sample both on Easting and Northing. This is shown in the top-right subfigure of Fig. 1. By using Northing as a second balancing variable, the sampling variance is further reduced to 4.5.

Note that when a sample is balanced on some covariates, this does not necessarily imply a good spread along the marginal axes of the space spanned by the covariates. This is nicely illustrated in Fig. 1. The sample in the top-right figure is balanced on Easting and Northing, but the spread along the horizontal axis for Easting is very poor. Balancing and spreading of sampling locations are different things. For that reason, I think the term spatially balanced samples for samples well-spread in geographic space is confusing. I prefer to refer to such designs as spatial coverage sampling.

Similar to the regression estimator, balanced sampling exploits the linear relation between the variable of interest and one or more covariates. In the regression estimator this is done at the estimation stage, see Brus (2000) for an application in soil survey. Balanced sampling does so at the sampling stage. For a single covariate the regression estimator equals

$$\widehat{\bar{y}}_{\mathrm{r}}=\widehat{\bar{y}}_{\mathrm{HT}}+b\left(\bar{x}-\widehat{\bar{x}}_{\mathrm{HT}}\right), \tag{2}$$

with $\widehat{\bar{y}}_{\mathrm{HT}}$ and $\widehat{\bar{x}}_{\mathrm{HT}}$ the Horvitz–Thompson estimators of the mean of the variable of interest y and the covariate x, respectively, $\bar{x}$ the known population mean (spatial mean) of the covariate, and b the estimated slope of the regression line. With a perfectly balanced sample the adjustment term in the regression estimator (the second term) equals zero.

2.1. The cube method for selecting balanced samples

Balanced samples can be selected by the “cube method” (Deville and Tillé, 2004). The name of this method is derived from the geometrical representation of the set of all possible samples that can be selected from a finite population. In sampling without replacement each population unit is either selected or not selected. The total number of samples that can be selected by sampling without replacement from a population of size N therefore equals 2^N. These samples can geometrically be represented by the vertices of an N-dimensional hypercube C. Fig. 3 shows a hypercube for a population of $N=3$ units. A sample from a finite population with N units is denoted as a vector $\mathbf{s}=(s_1,s_2\cdots s_N)'$, with $s_k=1$ if unit k is selected, 0 else. So, for example, in Fig. 3 the vertex (1,0,0) represents the sample of size 1 containing the first population unit.

A sampling design $p(\cdot)$ assigns a selection probability to each of the 2^N samples. The probability that a unit k is included in the sample is determined by the sample selection probabilities:

$$\pi_k=\sum_{\mathbf{s}\in\mathcal{S}_k}p(\mathbf{s}), \tag{3}$$

with $\mathcal{S}_k$ all samples containing unit k, and $p(\mathbf{s})$ the probability that sample $\mathbf{s}$ is selected. A sampling design $p(\cdot)$ determines the inclusion

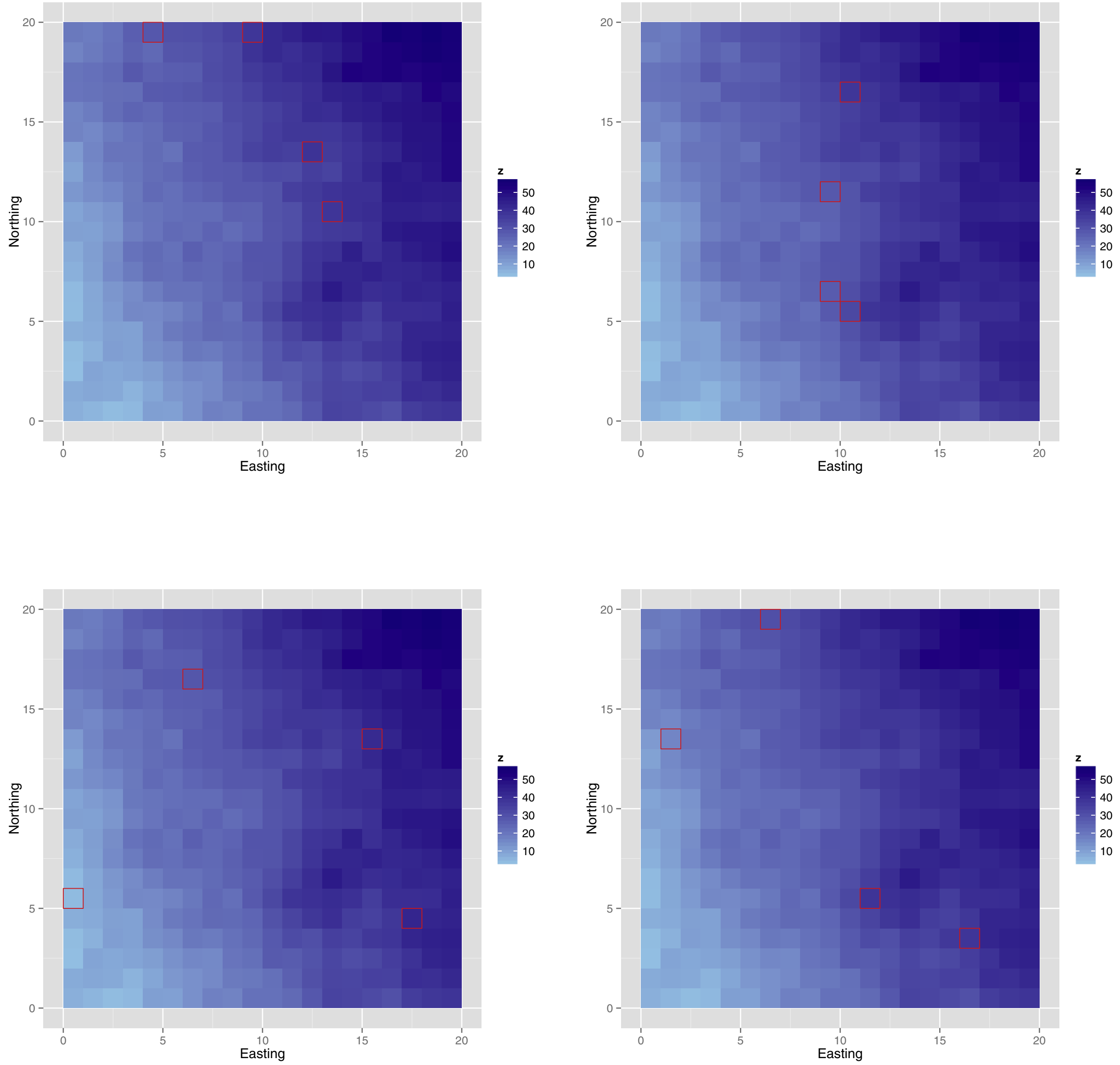

Fig. 1. Sample balanced on Easting (top-left), balanced on Easting and Northing (top-right), balanced coverage sample, balanced on Easting and Northing (bottom-left), and latin hypercube sample (bottom right).

probabilities for all population units $k = 1 \cdots N$, which can be represented as a vector within the N-dimensional cube. In Fig. 3 (left subfigure) this vector is labeled by $\boldsymbol{\pi}\,(0)$.

Tillé (2006) describes various sampling algorithms that define a random path through the N-dimensional hypercube C, starting from the vector with the inclusion probabilities, and ending at one of the vertices of C. The cube method is one of these algorithms (Tillé, 2011). It consists of two phases, the flight phase and the landing phase. In the flight phase the inclusion probabilities are sequentially updated until no changes can be made anymore. The algorithm of the flight phase is as follows:

1. Generate any vector $\mathbf{u}\,(t)$ such that $\check{\mathbf{x}}'\ \mathbf{u}(t) = 0$ and $u_k(t) = 0$ if $\pi_k(t)$ is an integer number.
2. Compute the largest values for $\lambda_1(t)$ and $\lambda_2(t)$ such that
   - $0 \leq \pi(t) + \lambda_1(t)\mathbf{u}(t) \leq 1$
   - $0 \leq \pi(t) - \lambda_2(t)\mathbf{u}(t) \leq 1$
3. Compute

$$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } q_1(t) \\ \pi(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } 1 - q_1(t) \end{cases},$$

where $\lambda_1^*(t)$ and $\lambda_2^*(t)$ are the largest value for $\lambda_1(t)$ and $\lambda_2(t)$ computed in step 2, and $q_1(t) = \lambda_2^*(t)\mathbf{u}(t)/(\lambda_1^*(t) + \lambda_2^*(t))$.

In this algorithm $\check{\mathbf{x}}$ is the vector of length N with the π-expanded values of the balancing variable, i.e., the values of the balancing variable x divided by the inclusion probabilities (Eq. (1)). The three steps of the above algorithm are repeated until no changes can be made anymore. In each iteration the inclusion probability of at least one unit is updated to either 0 (not included in sample) or 1 (included).

Geometrically, updating is done by adding a vector $\mathbf{u}$, multiplied by a scalar λ, to the current vector with inclusion probabilities. Fig. 3 illustrates this updating for a population of size three. In this case the vector $\mathbf{u}$ is in a two-dimensional plane Q perpendicular to the vector $\check{\mathbf{x}}$, the

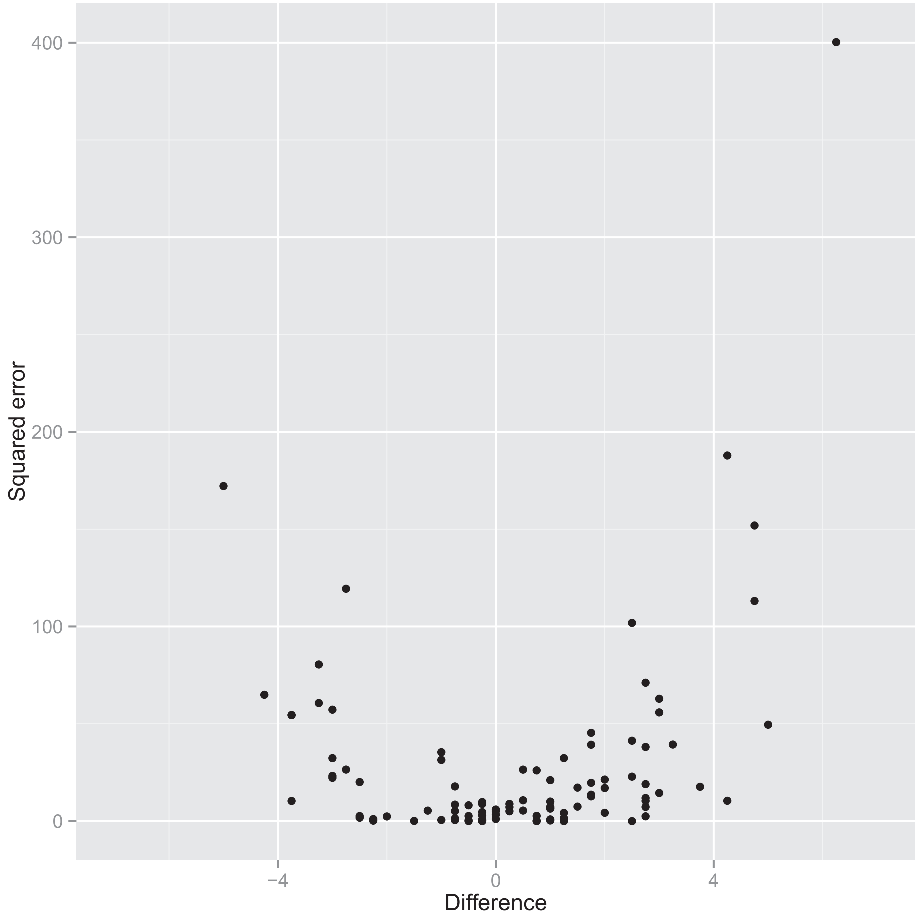

Fig. 2. Scatter of squared error in estimated mean for simple random sampling without replacement of size four, plotted against the difference between the sample average and population average of the ancillary variable (Easting).

shaded plane in Fig. 3. This plane Q is defined by the balancing equation, Eq. (1). If in Eq. (1) $\frac{x_k}{\pi_k}$ is multiplied by the sample membership indicator s_k, the sum over the n sampling units can be replaced by a sum over all N population units ($s_k = 0$ for units not selected):

$$\sum_{k=1}^{N} \frac{x_k}{\pi_k} s_k = \sum_{k=1}^{N} x_k. \tag{4}$$

For $N = 3$ this equation defines a two-dimensional plane. As long as the endpoint of the updated vector with inclusion probabilities stays in this plane, the balancing equation is satisfied, i.e., the estimated total of the covariate x equals the population total. The dashed arrows in the left subfigure are the vectors $\lambda_1^*(t)\mathbf{u}(t)$ and $-\lambda_2^*(t)\mathbf{u}(t)$ in the above algorithm for $t = 0$. The direction of $\mathbf{u}(t)$ in plane Q is randomly selected. One of the two vectors is randomly selected, with probability inversely proportional to their length. In Fig. 3 after the first iteration the endpoint of the updated vector is either in the top-side or the back-side of

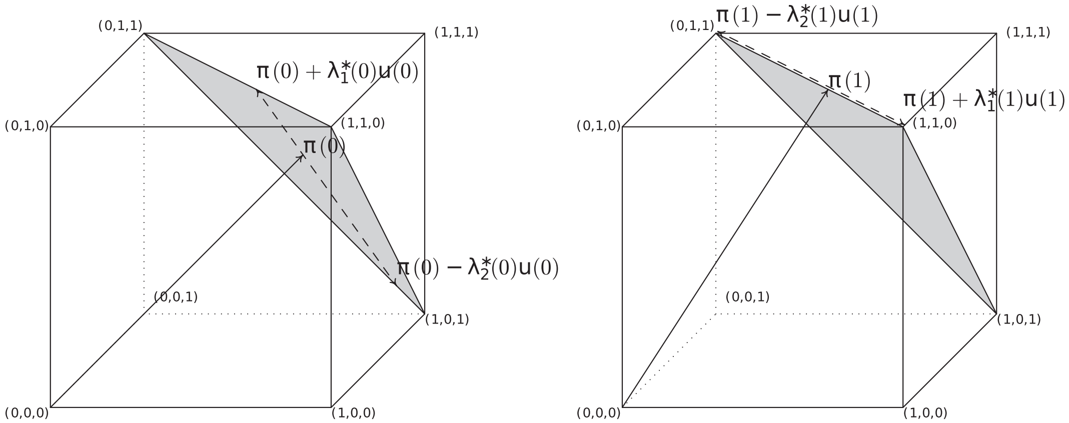

Fig. 3. Updating of vector with inclusion probabilities for a population of size 3 and a single balancing variable. Left: first iteration; Right: second iteration. After Tillé (2006).

the cube. If the first vector is selected, the updated vector is in the top-side and the second unit is selected ($s_2 = 1$); if the second vector is selected, the updated vector is in the back-side and the third unit is selected ($s_3 = 1$). Suppose the first vector is selected. Now a new vector $\mathbf{u}(1)$ is randomly selected in Q but with $s_2 = 0$. This means that $\mathbf{u}(1)$ is randomly selected along the diagonal (0,1,1)–(1,1,0) in the top-side of the cube. Again one of the two vectors constructed with $\mathbf{u}(1)$ is randomly selected, after which the endpoint of the vector with updated inclusion probabilities is in vertex (0,1,1) or (1,1,0), and a sample is selected. The algorithm is easily extended to multiple balancing variables by replacing vector $\check{\mathbf{x}}$ by an $N \times p$ matrix $\check{\mathbf{X}}$, with p the number of balancing variables. With multiple balancing variables the vector $\mathbf{u}$ is perpendicular to all columns in the matrix $\check{\mathbf{X}}$. Plane Q then is a hyperplane with dimension $N - p$, with p the number of balancing variables.

The sample of four units balanced on Easting shown in the top-left subfigure of Fig. 1 was selected with

$$\mathbf{X} = \begin{bmatrix} \pi_1 & x_1 \\ \pi_2 & x_2 \\ \vdots & \vdots \\ \pi_{400} & x_{400} \end{bmatrix}. \tag{5}$$

In the second column we have the Easting coordinates of all $N = 400$ sampling units. The first column with inclusion probabilities is added to achieve a fixed sample size. This can be explained as follows. If we insert $x_k = \pi_k, k = 1 \cdots N$ in Eq. (4), we obtain $\sum_{k=1}^{N} s_k = \sum_{k=1}^{N} \pi_k$. Let us take a vector with inclusion probabilities that sum to an integer n. In the example of Fig. 1 equal inclusion probabilities of 0.01 were assigned to all 400 units ($\pi_k = 0.01, k = 1 \cdots 400$), so that the sum of the inclusion probabilities equals four. So for samples of which the sum of the sample membership indicator s_k over all N population units equals four, i.e., samples of size four, the balancing equation is satisfied. Fig. 3 also illustrates this balancing on the inclusion probabilities, to achieve a fixed sample size of two, $\sum_{k=1}^{3} \pi_k = 2$. Eq. (4) then becomes

$$\sum_{k=1}^{3} s_k = 2. \tag{6}$$

Note that the cube vertices (1, 1, 0), (1, 0, 1) and (0, 1, 1), representing the three possible samples of size two, are in the plane Q defined by this equation.

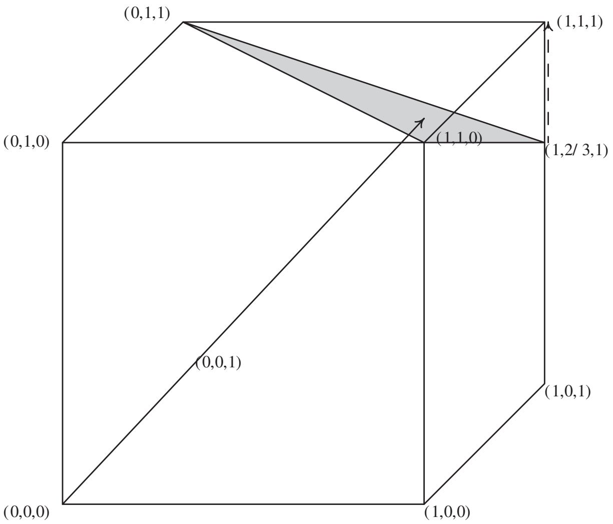

Fig. 4. Example of situation in which not for all samples the balancing equation is satisfied. $\boldsymbol{\pi} = (0.8, 0.8, 0.8)$; $\check{\mathbf{x}} = (1, 3, 1)$; $\sum_{i=1}^{3} x_i = 4$. The dashed arrow indicates the final updating in the landing phase when at the end of the flight phase vertex (1, 2/3, 1) is reached.
After Tillé (2006).

Following Tillé (2006), the selection of a probability sample was described as a random path through C, starting from the vector of inclusion probabilities, and ending at one of the vertices of C. The balancing equations impose constraints on this random path. As explained above, the random path through the hypercube C is constrained to a hyperplane Q that is perpendicular to the vectors with π-expanded values of the balancing variables in $\check{\mathbf{X}}$. At the end of the flight phase a vertex of the polytope K that is the intersection of the hyperplane Q and the hypercube C ($K = Q \cap C$), is reached. Two situations can be distinguished now. In the first situation all vertices of K are also vertices of C, as in Fig. 3. In this case all possible samples are perfectly balanced for all covariates. In the second situation some or all vertices of K do not coincide with vertices of C. In Fig. 4 the constraint plane Q passes through two vertices of C, but one vertex of K is not a vertex of C. This means that only two samples are perfectly balanced. If at the end of the flight phase a vertex of K is reached that is not a vertex of C (point (1, 2/3, 1) in Fig. 4), the algorithm enters the landing phase, in which a vertex of C is selected. Two methods are available for landing, linear programming, and by suppression of variables (Tillé, 2006, p. 163–164). Suppose q elements of the vector $\boldsymbol{\pi}$ at the end of the flight phase are unequal to either 0 or 1. Then there are 2^q samples that are equal to $\boldsymbol{\pi}$ at the end of the flight phase apart from the non-integer entries in this vector. Landing by linear programming is an enumerative search, through this set of 2^q samples, for the sample at shortest standardized Euclidian distance from the constraint plane Q (Tillé, 2006, p. 163–164). In Fig. 4 with $q = 1$ vertex (1,1,1) is closer to the shaded plane Q than vertex (1,0,1), and so sample (1,1,1) will be selected in the landing phase. For large q, say $q > 20$, an enumerative search becomes unfeasible, and landing by suppression of variables is the only option. In this landing method the constraints imposed by the covariates are relaxed one-by-one.

2.2. Balanced coverage random sampling

Balanced coverage random sampling, referred to as doubly balanced sampling, was proposed by Grafström and Tillé (2013). The method builds on the result found by Chauvet and Tillé (2006) that in each step of the flight phase it is not needed to use all N population units. It suffices to use a subset of units containing at least $p + 1$ units not updated to 0 (not included in sample) or 1 (included in sample) in previous steps.

Grafström and Tillé (2013) proposed to select in each step $p + 1$ neighboring units, i.e., a cluster of units. What are neighboring units depends on what variables are used to measure the Euclidian distance between the units. In general the spatial coordinates will be used for this, so that a cluster is a spatial cluster. However, we may also use other variables such as environmental covariates (and I actually will do this in the case study hereafter). The variables used for defining the Euclidian distance between the units are referred to as the spreading variables.

By updating the inclusion probabilities of the units in a cluster, for at least one unit the updated inclusion probability will be either 0 or 1. In updating the sum of the inclusion probabilities is respected, so if for one unit the inclusion probability is updated to 0, the inclusion probabilities for the other units in the cluster will increase. Similarly, when one unit is updated to 1, the inclusion probabilities of the other units of the cluster will decrease. So by updating the inclusion probabilities locally, the joint inclusion probabilities for neighboring units will be small. Fig. 1, bottom-left subfigure, shows a balanced coverage random sample, using Easting and Northing both for balancing and for spreading. The spatial coverage (spreading in geographic space) is much better than the sample only balanced on

Easting and Northing (top-right figure of Fig. 1). In practice often other variables than the spatial coordinates will be used for balancing such as environmental covariates.

2.3. Stratified random sampling as balanced sampling

The alternative for balanced sampling in the simple example above with Easting as a single balancing variable (top-left subfigure of Fig. 1) would be stratified random sampling, using vertical stripes of equal width as strata. I show now that stratified random sampling is a special case of balanced sampling. Basically it is balanced sampling using a categorical variable as a balancing variable. The design matrix **X** consists of as many columns as there are strata. For instance, with four strata **X** becomes

$$\mathbf{X} = \begin{bmatrix} \pi_{1,1} & 0 & 0 & 0 \\ \pi_{2,1} & 0 & 0 & 0 \\ \pi_{3,1} & 0 & 0 & 0 \\ \pi_{4,1} & 0 & 0 & 0 \\ \pi_{5,1} & 0 & 0 & 0 \\ 0 & \pi_{6,2} & 0 & 0 \\ 0 & \pi_{7,2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \pi_{400,4} \end{bmatrix}. \tag{7}$$

The first five rows refer to the five leftmost bottom row sampling units in Fig. 1. These units belong to stratum 1, which explains that the first column for these units contain non-zeroes. These non-zeroes are the inclusion probabilities of the units in stratum 1. The other three columns for these rows contain all zeroes. The sixth to 10*th* unit belongs to stratum 2, so that the second column for these rows contain the inclusion probabilities for stratum 2, and so on. The final row is the upper-right sampling unit in stratum 4, so the first three columns contain zeroes, and the fourth column is filled with the inclusion probability of this stratum. The sum of the inclusion probabilities in the first column is the sample size of stratum 1. Or reversely, if, for instance, we want to select n_h units from stratum h with equal probability for all units in this stratum, then the inclusion probabilities should equal n_h/N_h, with N_h the total number of units in this stratum.

If we want to use both Easting and Northing as categorical balancing variables, then the number of columns of the design matrix becomes the number of Easting-strata times the number of Northing-strata. So with four marginal strata for both variables the minimum sample size already equals 16 (one point per stratum). In digital soil mapping we often have more covariates, and we would like to have more marginal strata per covariate. The required sample size then readily becomes prohibitive. The alternative would then be a latin hypercube sample. This design can also be implemented as a balanced sampling design, as shown now.

2.4. Latin hypercube random sampling

Falorsi and Righi (2008) describe a balanced sampling approach for multi-way stratification designs. This approach is of interest when one has multiple stratification variables, each stratification variable leading to several strata, so that the total number of cross-classification strata becomes so large that not all cross-classification strata can be sampled separately. In the sampling approach of Falorsi and Righi (2008) as many covariates are defined as there are cross-classification strata. Let us denote the total number of stratification variables by P, and the total number of marginal strata by M. The vector of length M with covariate values for population unit k consists of P non-zero values in entries corresponding with the cross-classification stratum unit k belongs to, and zeroes in the remaining entries. For instance, for the population shown in Fig. 1 with two balancing variables and four marginal strata per variable, design matrix **X** becomes

$$\mathbf{X} = \begin{bmatrix} \pi_{1,1} & 0 & 0 & 0 & \pi_{1,1} & 0 & 0 & 0 \\ \pi_{2,1} & 0 & 0 & 0 & \pi_{2,1} & 0 & 0 & 0 \\ \pi_{3,1} & 0 & 0 & 0 & \pi_{3,1} & 0 & 0 & 0 \\ \pi_{4,1} & 0 & 0 & 0 & \pi_{4,1} & 0 & 0 & 0 \\ \pi_{5,1} & 0 & 0 & 0 & \pi_{5,1} & 0 & 0 & 0 \\ 0 & \pi_{6,2} & 0 & 0 & \pi_{6,2} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \pi_{101,5} & 0 & 0 & 0 & 0 & \pi_{101,5} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \pi_{400,16} & 0 & 0 & 0 & \pi_{400,16} \end{bmatrix}. \tag{8}$$

Similar to constrained latin hypercube sampling as proposed by Minasny and McBratney (2006), the marginal strata are constructed by computing quantiles of the cumulative spatial distribution function of the covariates. If a marginal sample size of one is targeted for, the quantiles $1/n$, $2/n$, $\cdots$, $(n-1)/n$ are computed, leading to n marginal strata per covariate with equal number of population units in it. If the targeted marginal sample size is larger than 1, say m, then the quantiles m/n, $2m/n$, $\cdots$, $(n-m)/n$ are computed. Strictly speaking, with $m > 1$ the design is not a latin hypercube design, but a specific case of a multi-way stratified design. The quantiles are used as the boundaries of the marginal strata. The number of units in the cross-classification strata, N_h, can then be computed. The inclusion probabilities are then computed as

$$\pi_{kh} = \frac{n}{LN_h}, \tag{9}$$

with L the number of non-empty ($N_h > 0$) cross-classification strata.

Fig. 1 (bottom-right subfigure) shows a latin hypercube random sample. The marginal strata are four vertical and four horizontal stripes of five unit width. In each of the eight stripes there is exactly one selected sampling unit.

When all cross-classification strata contain at least one population unit (pixel), $N_h \geq 1$ for all strata, all realized marginal sample sizes will be equal to the targeted marginal sample size m for any latin hypercube random sample selected with the design. This is easy to see for two covariates. With all cross-classification strata non-empty we have $L = n^2$. Then the marginal sample sizes are:

$$\sum_{h=1}^{\sqrt{L}} N_h \frac{n}{L \cdot N_h} = \sqrt{L}\frac{n}{L} = \frac{n}{\sqrt{L}} = 1 \tag{10}$$

with $L < n^P$ the marginal sample sizes may differ from m, and also the total sample size n becomes random.

2.5. Design-based estimation of spatial mean

With probability sampling, the mean of the variable of interest can be estimated unbiasedly by the Horvitz–Thompson estimator:

$$\widehat{\bar{y}}_{\mathrm{HT}} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}, \tag{11}$$

with N total number of population units (raster cells).

With latin hypercube random sampling I also estimated the mean by the ratio estimator

$$\widehat{\bar{y}}_{\mathrm{ratio}} = \frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}}{\sum_{i=1}^{n}\frac{1}{\pi_i}}. \tag{12}$$

The numerator in this estimator is an estimate of the population total of the variable of interest y, the denominator is an estimate of the population size N (total number of raster cells). Although this population size is known, in the ratio estimator it is estimated from the sample. When one or several units are selected with very small inclusion probabilities the estimated population total will be large, as in the Horvitz–Thompson estimator the observation are expanded by these inclusion probabilities (see Eq. (11)). Also the estimated population size, $\hat{N}$, will then be large. So by dividing through $\hat{N}$ instead of N, the estimator will correct for the selection of units with very small inclusion probabilities. As a consequence, I expect a smaller sampling variance of the estimated mean for this ratio estimator.

3. Case study

The case study is on soil organic carbon (SOC) mapping in three woredas in Ethiopia. Legacy data were used to fit a random forest for natural log-transformed SOC using ten covariates and the spatial coordinates. The random forest was used to predict lnSOC at the nodes of a fine grid. The residual variogram of the random forest predictions showed no spatial structure, so I added white noise to the predictions, to obtain a realistic field of lnSOC.

In selecting random samples I used as balancing variables landsurface temperature (*lst*), elevation (*dem*), near-infrared reflectance (*NIR*), and enhanced vegetation index (*evi*). These four covariates were important in the random forest. Note that less covariates were used in sampling as in simulation. I selected balanced (b), balanced coverage (bc), and latin hypercube (lh) random samples. In bc random sampling I used both the spatial coordinates and the four covariates for spreading, so that the samples have good coverage of both geographic and feature space. The sample sizes were 20 and 100.

With all four probability sampling designs 1000 samples were selected. We used R package `BalancedSampling`, available at http://www.antongrafstrom.se/balancedsampling/, for selecting the random samples.

In lh with $n = 20$ I used $m = 1$ as a targeted marginal sample size, with $n = 100$ this sample size was five. So in both cases the number of marginal strata per covariate was 20. With $n = 100$ and $m = 1$ ties in the quantiles for covariate *lst* were obtained (not enough different values for *lst*). With $m = 2$ these ties were avoided, but computing time was prohibitively large due to the large size of matrix $\mathbf{X}$, which then has $4 \times 50 = 200$ columns. Computing time was less than a second per sample for all balanced samples, except for lh with $n = 100$. For this sampling design computing time was about 7 min per sample.

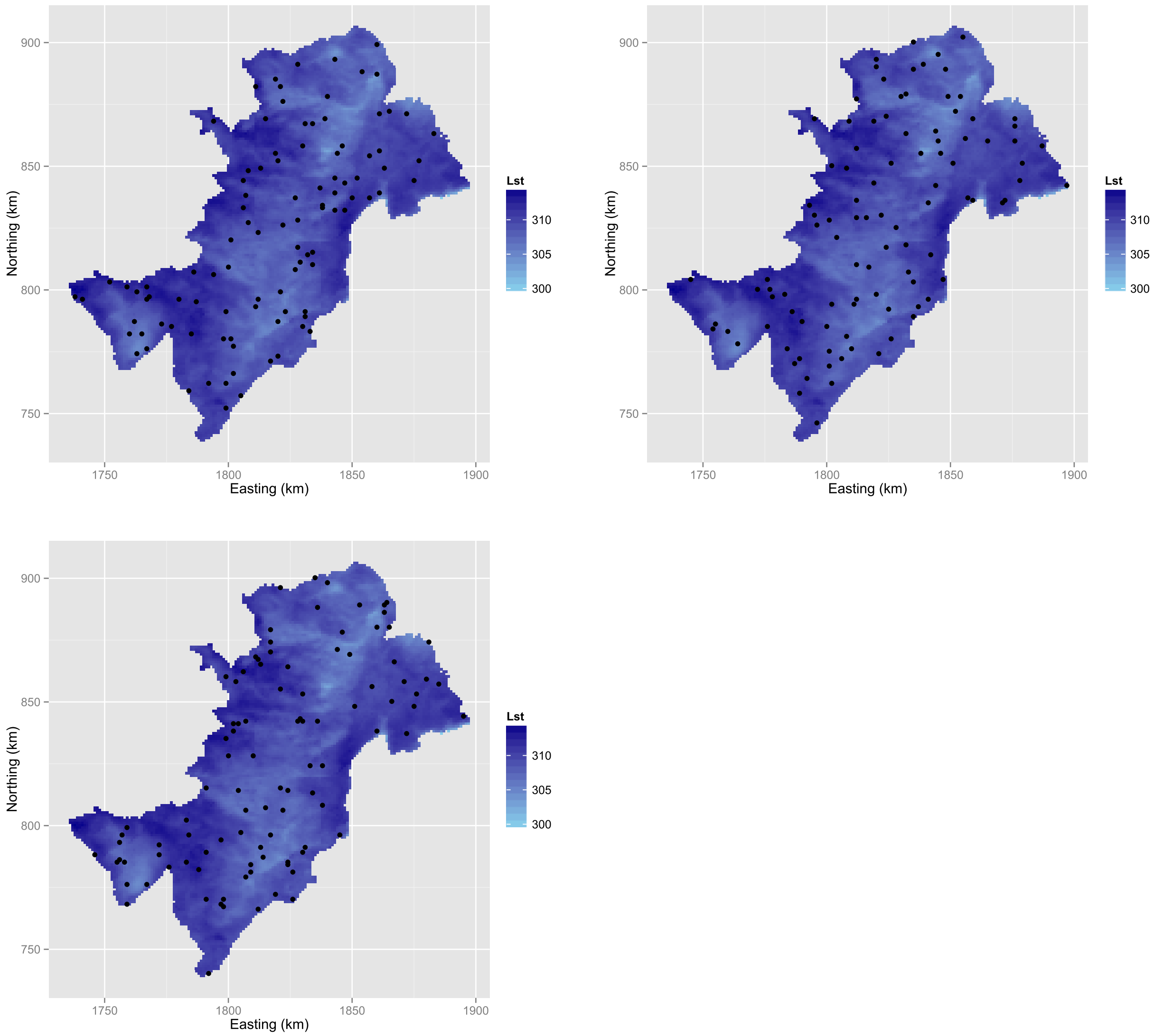

Fig. 5. A balanced (top-left), balanced coverage (top-right), and latin hypercube (bottom-left) of size 100 plotted on the covariate landsurface temperature.

In b and bc I used equal inclusion probabilities. In lh the inclusion probabilities were not constant. Units in small cross-classification strata had larger inclusion probability than units in large cross-classification strata, see Eq. (9). A total of 6100 out of 7742 non-empty cross-classification strata contained one unit only. The largest cross-classification stratum contained 47 units. As a consequence, the inclusion probabilities of these 47 units was 47 times smaller than those of the majority of the units.

Fig. 5 shows a balanced, balanced coverage, and a latin hypercube random sample of size 100 from the study area.

3.1. Evaluation of the sampling designs

I will evaluate the four sampling designs on the basis of the sampling distribution of

- number of unsampled marginal LHS strata

$$U = \sum_{p}^{P} \sum_{h}^{H} i_{ph} \tag{13}$$

with P the number of covariates ($P = 4$), H the number of marginal strata per covariate ($H = 20$), and $i_{ph} = 1$ if sample size in marginal stratum h of covariate p is zero, and 0 else.

- spatial coverage as quantified by the Mean Squared Shortest Distance:

$$MSSD = \frac{1}{N} \sum_{i=1}^{N} \min_j D_{ji}^2 \tag{14}$$

with D_{ij} the distance between node (raster cell) i and all sampling points. This criterion is minimized in spatial coverage random sampling as proposed by Brus et al. (1999) and implemented in the R package `spcosa` (Walvoort et al., 2010).

- error in estimated spatial mean of lnSOC

$$e_{HT} = \widehat{\bar{y}}_{\mathrm{HT}} - \bar{y}, \tag{15}$$

with $\bar{y}$ the true spatial mean of y.

For lh also the error for the ratio estimator was computed.

Small values for U are important for growing regression trees and random forests. The cutpoints of the covariates used as splitting variables are unknown, so it is important to have a sample with a good coverage of the marginal distributions of the covariates. Small values for $MSSD$ are profitable in estimating the spatial mean and for mapping by spatial interpolation (ordinary kriging or kriging with an external drift)

3.2. Results

With $n = 20$ the number of unsampled marginal strata U was by far the smallest for lh (Fig. 6). The median number of unsampled marginal strata was 15 (total number of marginal strata was 80). Design bc scored somewhat better on this criterion than design b. This can be explained by the use of the covariates (and spatial coordinates) as spreading variables. Also for $n = 100$ lh scored best on U, but the differences with the other two balanced designs were smaller. With lh of size 100 15% of the

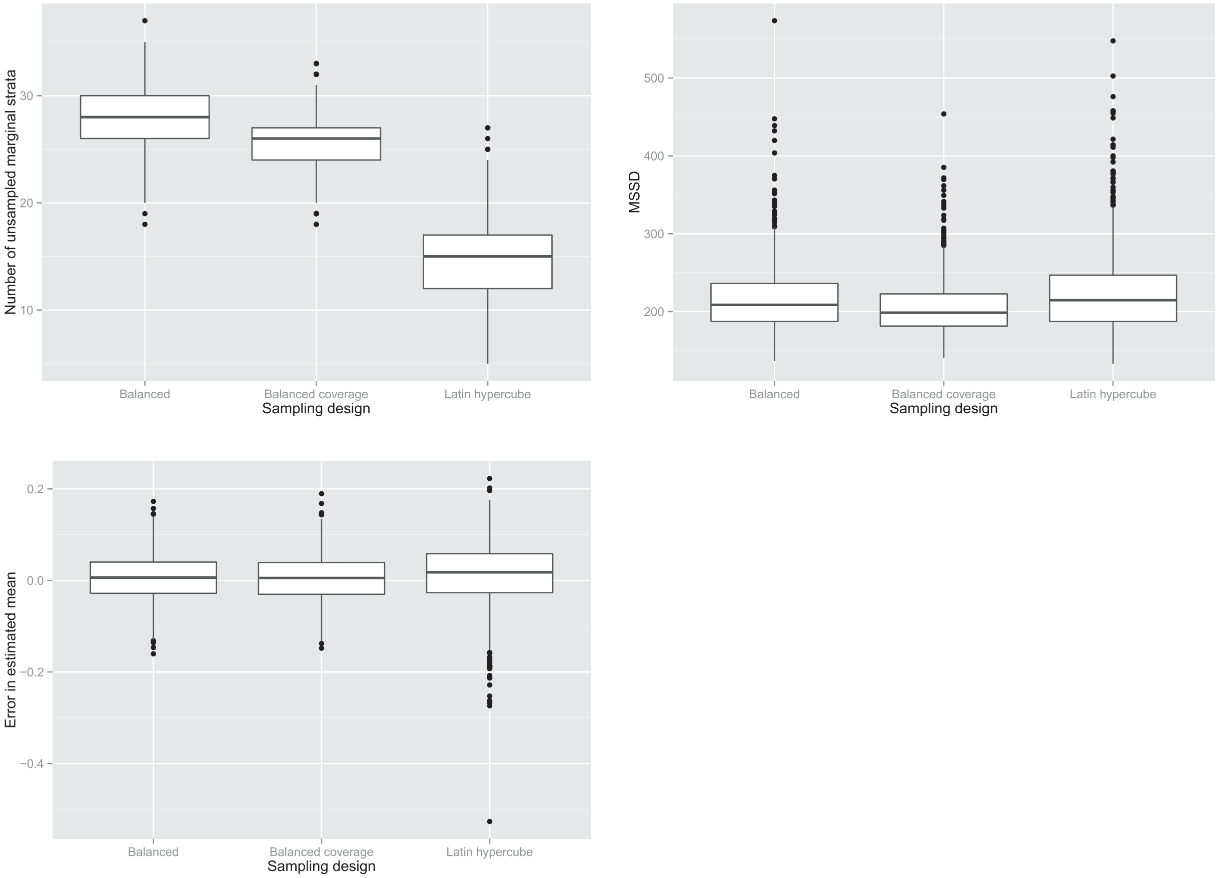

Fig. 6. Distribution of number of unsampled marginal strata, Mean Squared Shortest Distance and error in estimated mean for three balanced random sampling designs of size 20.

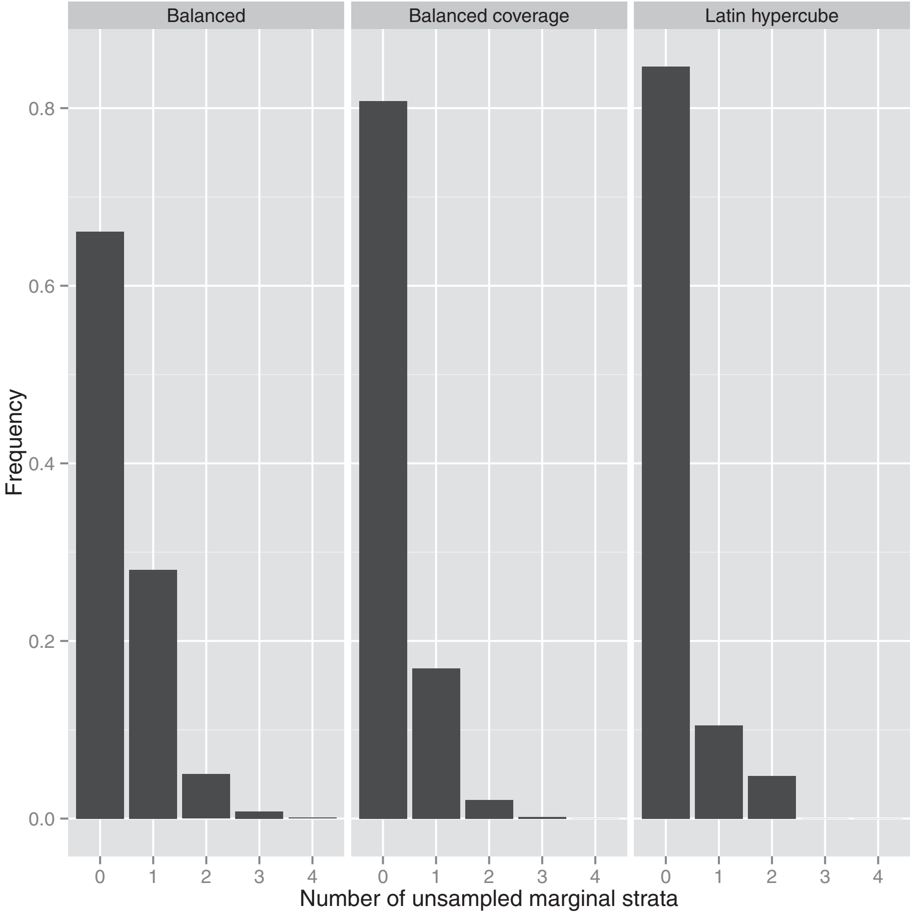

Fig. 7. Distribution of number of unsampled marginal strata for three balanced random sampling designs of size 100.

samples did not cover all marginal strata. Again design bc scored somewhat better than design b (Fig. 7).

In bc also the spatial coordinates were used as spreading variables. This explains why with respect to criterion *MSSD* sampling design bc scored somewhat better than designs b and lh, for both sample sizes (Figs. 6, 8).

With respect to the error in the estimated mean, bc scored best (smallest variance) for both sample sizes, very closely followed by design b. The sampling variance of the ratio estimator of the mean with lh was about twice the variances with the other two designs. The Horvitz–Thompson estimator of the mean performed even more poor with lh. The sampling variance of this estimator of the mean for $n = 20$ was ten times larger than the variance of the ratio estimator: 0.0539 versus 0.00554. This is due to the large differences in inclusion probabilities that were not related to the lnSOC values.

Designs b and bc were more precise than simple random sampling (SI). The variance of the estimated mean for these designs divided by the variance with SI ranged from 0.78 to 0.85 (Table 1). The modest gain in precision can be explained by the low R-square value of 14.6% of the linear regression model for lnSOC using the four covariates as

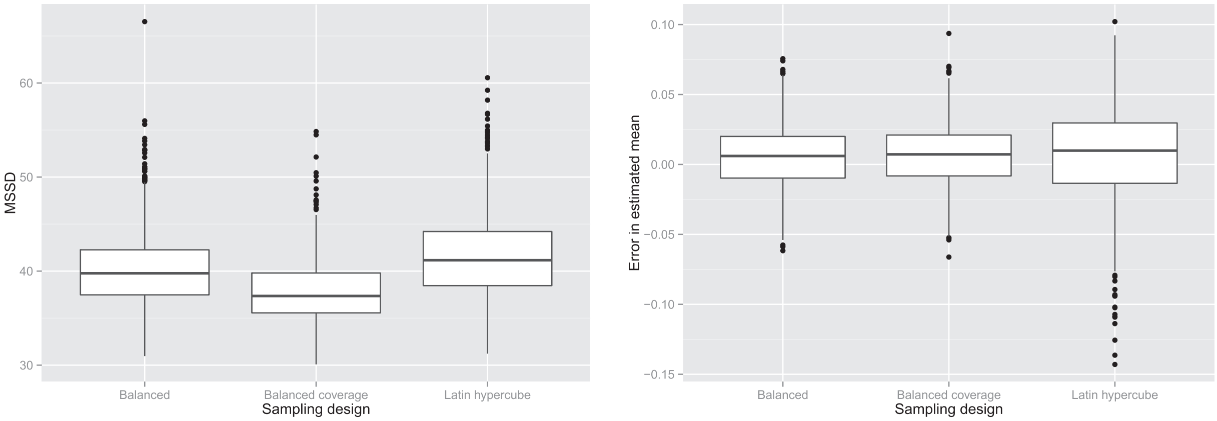

Fig. 8. Distribution of Mean Squared Shortest Distance and error in estimated mean for three balanced random sampling designs of size 100.

Table 1
Variance of estimated mean of balanced (b), balanced coverage (bc), and latin hypercube (lh) random sampling divided by the variance for simple random sampling, for two sample sizes.

	b	bc	lh
20	0.85	0.81	1.78
100	0.82	0.78	1.88

predictors. Sampling design lh was much less precise than SI. The variance of the ratio estimator of the mean with lh was about 1.8 times the variance of the estimated mean with SI.

All 1000 balanced and balanced coverage samples had the required sample size (20 or 100). With lh the mean sample size was 20.05 and 100.0. With an expected size of 20 the minimum was 17, and the maximum was 24; with an expected size of 100 these extremes were 95 and 106. For 72 out of the 1000 lh samples with an expected sample size of 20 the realized sample size was ≤17 or ≥23. For $n = 100$ there were 12 samples with a sample size ≤95 or ≥105.

Fig. 9 shows the marginal sample sizes for lh, averaged over the 1000 lh samples, for an expected total sample size of 20. Note that deviations from the targeted marginal sample size of 1 for $n = 20$ increased from *lst*, *dem*, *NIR* to *evi*. This order equals the order in which these covariates were suppressed in the landing phase of the cube algorithm. This order is determined by the columns in matrix **X**. Deviations were largest for the first three marginal strata of *NIR* and *evi*. These strata were most under-represented on average. For a sample size of 100 the pattern of the average marginal sample sizes was very similar (not shown).

4. Discussion

Balanced sampling as described in this paper generates samples with known inclusion probabilities. In balanced and balanced coverage random sampling these inclusion probabilities can be freely chosen. In latin hypercube random sampling these inclusion probabilities follow from the number of population units in the cross-classification strata. I think probability sampling with known inclusion probabilities has several advantages over non-probability sampling. First, when the inclusion probabilities are known, both design-based estimates as well as model-based predictions of population parameters such as the mean or total can be obtained. Second, probability sampling with equal inclusion probabilities protects against biased predictions in mapping by model-based inference. It is well-known that when a preferential sample is used in conventional kriging, it is certainly not guaranteed that the population mean of the prediction errors equals zero, despite the model-unbiased predictions at all points in the study area (Gelfand et al., 2012). So in kriging it is important to have a "representative" sample. A simple and straightforward way of selecting such a sample is probability sampling with equal inclusion probabilities.

This raises the question, should we use the inclusion probabilities in fitting models when these inclusion probabilities are not constant, but differ between population units, see for instance Smith (2001) and references therein for a discussion on this. The same question is relevant for random forest modeling. In this modeling approach the actually selected sample is resampled by bootstrapping. In latin hypercube random sampling units in large cross-classification strata have small inclusion probabilities, whereas units in small strata have large inclusion probabilities. In other words, small strata are under-represented, big strata are over-represented in the lh sample. Should we account for these differences in inclusion probability in selecting bootstrap samples, which is common in bootstrapping for estimating means and their estimation variances from complex surveys (Rao and Wu, 1988; Sitter, 1992)? I welcome research into whether weighted bootstrapping also leads to increased accuracy of random forest predictions.

In constrained latin hypercube sampling the objective function consists of three terms, one term equals two times *U*, a second term is a measure for the deviation from proportional sample sizes of strata based on categorical covariates, and a third term is a measure of the difference between the population and sample correlation matrix of the covariates. A weighted sum of the three terms is minimized. The weights must be chosen by the user. I am not aware of theory for optimizing these weights, which makes the choice somewhat arbitrary and the method heuristic. In latin hypercube random sampling this weighting is avoided.

In the case study the precision of the mean as estimated from the latin hypercube random sample was poor compared to simple random sampling. The model-assisted ratio-estimator performed much better than the Horvitz–Thompson estimator, but also for this estimator the variance was about twice the variance with simple random sampling. Further research is needed into how the quality of design-based or model-assisted estimates from latin hypercube random samples can be improved.

The balanced coverage random sample can be, I think, an alternative for the model-based design minimizing the variance of prediction errors obtained with kriging with an external drift, as proposed by Brus and Heuvelink (2007). With both designs sampling units are well-spread in

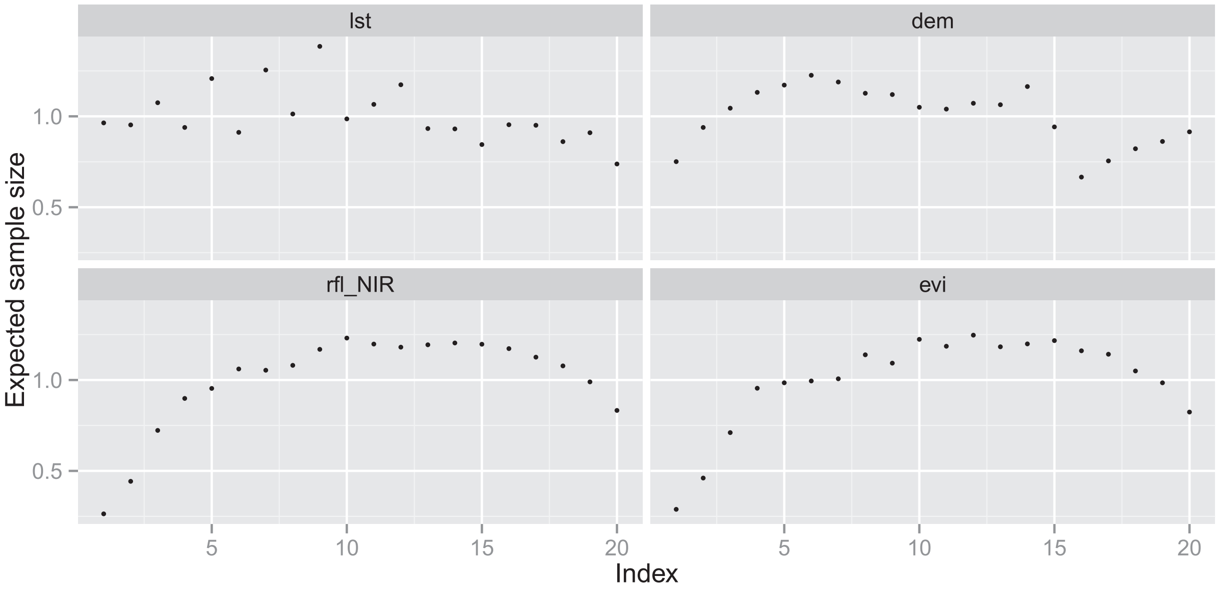

Fig. 9. Average sample sizes of the 4 × 20 marginal strata for a latin hypercube random sample of expected size 20.

geographic and feature space. The model-based design builds on the assumption of a linear relation between the covariates and the soil property of interest. In the extreme case of no autocorrelation of the model-residuals, locations will be selected with extreme (either very small or very large) values for the covariates. Spatial clustering is not avoided, and no locations will be selected with intermediate values. When a variogram is used with a partial sill >0, the resulting sample will have much better spatial coverage and, thanks to this, in practice also locations will be selected with intermediate values for the covariates. In contrast to this model-based design, in balanced coverage random sampling no assumption is made on the relation between the soil property of interest and the variables used in spreading the sample. It would be interesting to compare the performance of this balanced coverage random design with the model-based design in digital soil mapping.

5. Conclusions

- Balanced sampling is a flexible approach for selecting random samples. By balancing the sample on covariates that are correlated with the soil property of interest, the precision of design-based estimates of population parameters will be increased compared to simple random sampling. In the case study with an R-square value of the linear regression model for lnSOC of 14.6%, the variance of the estimated mean was about 0.8 times the variance with simple random sampling.
- Besides numeric also categorical variables can be used in balancing. Balancing on a categorical variable boils down to stratified random sampling. Latin hypercube random sampling is a special case of this.
- In the case study the use of the geographical coordinates and covariates as spreading variables in balanced coverage random sampling led to a modest improvement of the coverage of the marginal strata (smaller values for *U*) and of geographic space (smaller values for *MSSD*) as compared to balanced sampling without spreading.
- Coverage of the marginal strata by the latin hypercube random samples was satisfactory. However, the quality of the mean as estimated by the ratio estimator was poor as compared to the other two balanced designs and to simple random sampling.
- Contrary to constrained latin hypercube sampling (Minasny and McBratney, 2006) with latin hypercube random sampling the inclusion probabilities are known, so that we are flexible in estimating population means or totals either by design-based, model-assisted (Särndal et al., 1992) or model-based inference. The inverse of the inclusion probabilities can also be used as weights in growing random forests from latin hypercube random samples.

References

Brus, D.J., 2000. Using regression models in design-based estimation of spatial means of soil properties. Eur. J. Soil Sci. 51, 159–172.

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–59.

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., 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.

Chauvet, G., Tillé, Y., 2006. A fast algorithm of balanced sampling. Comput. Stat. 53–62.

Deville, J.-C., Tillé, Y., 2004. Efficient balanced sampling: the cube method. Biometrika 91, 893–912.

Falorsi, P.D., Righi, P., 2008. A balanced sampling approach for multi-way stratification designs for small area estimation. Surv. Methodol. 34, 223–234.

Gelfand, A.E., Sahu, S.K., Holland, D.M., 2012. On the effect of preferential sampling in spatial prediction. Environmetrics 23, 565–578.

Grafström, Anton, Tillé, Yves, 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24 (2), 120–131.

Grafström, A., Lundström, N.L.P., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520.

Hung, Y., Amemiya, Y., Wu, C.F.J., 2010. Probability-based latin hypercube designs for slid-rectangular regions. Biometrika 97, 961–968.

Lesch, S.M., 2005. Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. Comput. Electron. Agric. 46 (1–3 SPEC. ISS), 153–179.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resour. Res. 31, 387–398.

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

Rao, J.N.K., Wu, C.F.J., 1988. Resampling inference with complex survey data. J. Am. Stat. Assoc. 83, 231–241.

Robertson, B.L., Brown, J.A., McDonald, T., Jaksons, P., 2013. BAS: balanced acceptance sampling of natural resources. Biometrics 69, 776–784.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.

Sitter, R.R., 1992. A resampling procedure for complex survey data. J. Am. Stat. Assoc. 87, 755–765.

Smith, T.M.F., 2001. Biometrika centenary: sample surveys. Biometrika 88, 167–194.

Stevens, D.L., Olson, A.R., 2004. Spatially balanced sampling of natural resources. J. Am. Stat. Assoc. 99, 262–278.

Tillé, Y., 2006. Sampling Algorithms. Springer.

Tillé, Y., 2011. Ten years of balanced sampling with the cube method: an appraisal. Surv. Methodol. 37 (2), 215–226.

Tillé, Y., Matei, A., 2012. Survey Sampling (R package version 2.5).

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

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.