Sampling for digital soil mapping: A tutorial supported by R scripts

Brus, D. J.

Sampling for digital soil mapping: a tutorial supported by R scripts

D.J. Brus[a,b]

[a] *Biometris, Wageningen University and Research, PO Box 16, 6700 AA Wageningen, Netherlands*
[b] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application (School of Geography, Nanjing Normal University), Nanjing, 210023, China*

Abstract

In the past decade substantial progress has been made in model-based optimization of sampling designs for mapping. This paper is an update of the overview of sampling designs for mapping presented by de Gruijter et al. (2006). For model-based estimation of values at unobserved points (mapping), probability sampling is not required, which opens up the possibility of optimized non-probability sampling. Non-probability sampling designs for mapping are regular grid sampling, spatial coverage sampling, k-means sampling, conditioned Latin hypercube sampling, response surface sampling, Kennard-Stone sampling and model-based sampling. In model-based sampling a preliminary model of the spatial variation of the soil variable of interest is used for optimizing the sample size and or the spatial coordinates of the sampling locations. Kriging requires knowledge of the variogram. Sampling designs for variogram estimation are nested sampling, independent random sampling of pairs of points, and model-based designs in which either the uncertainty about the variogram parameters, or the uncertainty about the kriging variance is minimized. Various minimization criteria have been proposed for designing a single sample that is suitable both for estimating the variogram and for mapping. For map validation additional probability sampling is recommended, so that unbiased estimates of map

quality indices and their standard errors can be obtained. For all sampling designs R scripts are available in the supplement. Further research is recommended on sampling designs for mapping with machine learning techniques, designs that are robust against deviations of modeling assumptions, designs tailored at mapping multiple soil variables of interest and soil classes or fuzzy memberships, and probability sampling designs that are efficient both for design-based estimation of populations means and for model-based mapping.

Keywords:

spatial coverage sampling, spatial simulated annealing, k-means sampling, model-based sampling, latin hypercube sampling, kriging, variogram

---

1. Introduction

The design of a soil survey scheme is a crucial, first step in digital soil mapping (Domburg et al., 1994; de Gruijter et al., 2006). An important element in this design process is the choice of the sampling design. This paper describes and illustrates sampling designs for mapping of soil attributes. A rich plethora of sampling designs for mapping is available, from straightforward simple designs to advanced, complicated designs. De Gruijter et al. (2006) present an overview of these sampling methods. Since 2006 numerous publications have been published on sampling for mapping, especially on model-based sampling. The aim of this paper is to present an update of the overview of de Gruijter et al. (2006), to illustrate them with real-world case studies, and to describe how the sampling designs can be implemented using the popular statistical language R (R Core Team, 2016). R scripts are available as a supplement to this article at `https://github.com/DickBrus/TutorialSampling4DSM`.

The sampling methods are illustrated with four case studies:

- Cotton Research Field, Khorezm province, Uzbekistan
- Hunter Valley, New South Wales, Australia
- The woredas Alefa, Chilga and Dembia, Ethiopia
- Xuancheng, Anhui province, China

The first case study is a survey of the salinity of the soils at the Cotton Research Field in Khorezm, Uzbekistan. Electromagnetic induction (EMI) was measured with the EM38-MK2 instrument, with receivers at 1 m and 50 cm from the transmitter, positioned in the vertical dipole orientation. The effective depth of the measurements equals about 1.5 m and 0.75 m, respectively. Details can be found in Akramkhanov et al. (2013). This case study is used to illustrate the selection of sampling locations for calibrating a multiple linear regression model (section 4.2), and for mapping using kriging with an external drift (section 5.2)

For the Hunter Valley study area we have raster maps of five quantitative covariates: elevation, slope, aspect, compound topographic index (cti), and normalized difference vegetation index (ndvi). This case study is used to illustrate, amongst others, k-means sampling (section 3.3) and conditioned Latin hypercube sampling (section 4.1)

The data of the three woredas (administrative regions) in Ethiopia are concentrations of soil organic matter (SOM) in the A horizon. By far the most sampling locations are located along roads (convenience sample). Raster maps of near-infrared (NIR), visible infrared, land surface temperature, enhanced vegetation index and elevation are available for this study area. This data set is used to illustrate spatial infill sampling (section 3.2), model-based optimization of the spacing of a square grid (section 5.1) and model-based infill sampling (section 5.2).

In Xuancheng SOM concentration in the A horizon was measured at 121 sites. Besides the soil data, we have raster maps of elevation, precipitation and slope. These data and covariates are used to map SOM in the A horizon by kriging with an external drift (KED) and random forests (RF). A stratified simple random sample of 62 points is used as an illustration of how to estimate quality indices of the maps and how to test hypotheses about these quality indices (section 7).

2. Probability versus non-probability sampling

At the highest level one may distinguish random from non-random sampling methods. In random sampling a subset of population units is randomly selected from the population, using a random number generator. Examples of non-random sampling are convenience sampling e.g. along roads, arbitrary sampling i.e. sampling without a specific purpose in mind, and targeted sampling. In the literature the term random sampling is often used for arbitrary sampling, i.e. sampling without a specific purpose in mind. To avoid confusion the term probability sampling was introduced. Probability sampling is random sampling fulfilling two requirements. Firstly, all units in the population have a positive probability of being selected. No parts of the population may be excluded. Secondly, the selection probability of each possible sample is known. With arbitrary sampling these two requirements are often not met.

The choice between probability or non-probability sampling is closely connected with the choice between a design-based or model-based approach for statistical inference (estimation, hypothesis testing) (de Gruijter and ter Braak, 1990; Papritz and Webster, 1995; Brus and de Gruijter, 1997). In the design-based approach units are selected by probability sampling. Estimates are based on the selection probabilities of the sampling units as determined by the sampling design (design-based inference). No model is used in estimation. On the contrary, in a model-based approach a

stochastic model is used in estimation, for instance a linear regression or an ordinary kriging model. As the model already contains a random error term, probability sampling is not required in this approach, which opens up the possibility of optimized non-probability sampling. As an illustration, consider the following model: $z_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with z_i the variable of interest of unit i, x_i a covariate of that unit, β_0 and β_1 regression coefficients and ϵ_i the error (residual) at unit i, normally distributed with mean zero and a constant standard deviation σ. The errors are independent, so that $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. Figure 1 shows a simple random sample without replacement (SRS) and the sample optimized for the calibration of the simple linear regression model. Both samples are plotted on a map of the covariate (predictor). The standard errors of both regression coefficients (computed for a residual standard deviation σ of 2) are considerably smaller for the optimized sample (Table 1). The joint uncertainty about the two regression coefficients, quantified by the determinant of the variance-covariance matrix of the estimated regression coefficients, equals 0.0020 for SRS and 0.00010 for the optimized sample. So, we conclude that for mapping with a simple linear regression model, simple random sampling is not a good option.

A model-based approach for sampling and statistical inference does not necessarily imply model-based sampling. The adjective model-based refers to the model-based inference, not to the selection of the locations. In a model-based approach sampling locations can be, but need not be selected by model-based sampling. If they are, then both in selecting the locations and in mapping a statistical model is used. In most cases the two models differ: the sample data are used to update the postulated preliminary model used for sampling design. This updated model is then used for mapping.

3. Geometric sampling designs

3.1. Regular grid sampling

A straightforward, popular sampling method for mapping is sampling on a regular grid, for instance a square or triangular grid. As opposed to regular grid sampling in a design-based approach (systematic random sampling), in a model-based approach there is no need to place the grid randomly on the area, but can be placed in such way that the coverage of the study area by the grid is optimal.

When sampling on a regular grid we must decide on the grid-spacing, i.e. the distance between neighboring points. This boils down to a decision on the sample size, i.e. the number of grid points. There are two options to decide on this spacing, either by starting from the available budget or from a requirement on the quality of the map. The latter will be explained hereafter, in section 5.1, as this requires a model of the spatial variation, and as a consequence this is already an example of model-based sampling. Starting from the available budget and an estimate of the costs per point, we first compute the affordable sample size. Then, for a square grid the grid spacing can be computed by $d = \sqrt{A/n}$ with A the size of the area and n the affordable sample size. With units of area in m^2, the grid spacing is in m.

Square grids can be selected with function `spsample` of R package `sp` (Pebesma and Bivand, 2005).

3.2. Spatial coverage sampling

With regular grid sampling of irregularly shaped areas the geographical spreading of the sampling locations throughout the study area can be suboptimal. In some parts of the study area the distance to the nearest sampling point can be relatively large. In this case we would like to relax the constraint of sampling on a regular grid. We would like to shift grid points a bit into the undersampled areas, so that the

spatial pattern becomes irregular. This leads to spatial coverage sampling in which a geometric criterion defined in terms of the distances between the nodes of a fine discretisation grid and the sampling points is minimized (Royle and Nychka, 1998). Brus et al. (2007) proposed to minimize the Mean Squared Shortest Distance (MSSD) by k-means. The spatial coordinates of the centroids of the cells of a discretisation grid are used as variables in k-means clustering of the grid cells. The centroids of the clusters are used as sampling points.

If one already has measurements at locations with known spatial coordinates (legacy point data), and it is safe to assume that the measurements are still valid, it can be efficient to use these data in mapping. In this case we do not want to select new locations in the neighbourhood of the existing locations, but instead we want to fill-in the undersampled areas.

Figure 2 shows a spatial coverage and spatial infill sample of 100 points for the Ethiopia case study area. Legacy data are collected mainly along roads; this is a nice example of convenience sampling. The spatial coverage sample does not take these legacy data into account; this would be appropriate if we do not want to use the legacy data, for instance because the quality of the data is poor. If we do want to use the legacy data, a spatial infill sample can be designed. The new sampling locations are more in the interior parts of three woredas.

Spatial coverage and spatial infill samples can be selected with R package `spcosa` (Walvoort et al., 2010a,b), see `SpatialCoverageSample.R` and `SpatialInfillSample.R` in the supplement.

3.3. k-means sampling

In regular grid and spatial coverage sampling the selection of the sampling locations is entirely based on the spatial coordinates of the locations. Covariates possibly

related to the soil property of interest, are not accounted for in selecting sampling locations. This can be suboptimal when the soil property of interest is related to covariates of which maps are available, think for instance of remote sensing imagery. These maps can then be used in mapping the soil property of interest using, for instance using a multiple linear regression model. This subsection describes methods for selecting sampling locations on the basis of the covariate values of the grid cells.

3.3.1. Hard k-means

In hard k-means sampling the covariates are used to cluster the grid cells by the k-means clustering algorithm. Similar to spatial coverage sampling the MSSD is minimized, but now the distance is not measured in geographical space but in a p-dimensional space spanned by the p covariates (think of it as a multi-dimensional scatter plot with the covariates along the axes). In hard k-means each unit can only belong to exactly one cluster. Figure 3 shows an example for the Hunter Valley study area. The five quantitative covariates elevation, slope, aspect, cti and ndvi were used as covariates. The sample of size was set to 20, and so 20 clusters were constructed using hard k-means. Note that the number of clusters is based on the required sample size (number of clusters equals number of sampling locations), not on the number of subregions with a high density of points in the multivariate distribution. The covariates are scaled so that their standard deviations become 1. Grid cells with the shortest scaled Euclidean distance in covariate-space to the centroids of the clusters are selected as the sampling points. Figure 4 shows the selected sample in a scatter diagram of elevation versus cti.

Hard k-means clustering of the units (cells) can be done with function `kmeans` of package `stats`, see R script `KMSample.R` in the supplement.

3.3.2. Fuzzy k-means

Contrary to hard k-means, fuzzy k-means (also referred to as soft k-means) allows units to belong to one or more clusters. A vector containing k numbers is assigned to every unit, with all numbers in the interval $[0, 1]$; the numbers sum to 1. The numbers indicate the degree to which a unit belongs to each cluster. They are referred to as membership grades. With fuzzy k-means, the centroid of a cluster is the weighted mean of the covariates over all units, using the memberships of that cluster as weights. As before, grid cells with the shortest Euclidean distance in covariate-space to the centroids of these fuzzy clusters are selected as the sampling points. These are the locations with the largest membership in the fuzzy subsets $1 \cdots k$.

K-means clustering is a well-known technique for selecting a subsample from a larger sample with NIR and vis-NIR spectroscopy. On the subsample the variable of interest is measured (Naes, 1987). For recent applications of fuzzy k-means in soil spectroscopy, see Debaene et al. (2014) and Ramirez-Lopez et al. (2014).

Fuzzy k-means clustering can be done with function `FKM` of package `fclust` and function `runFuzme` of R package `fuzme`. R package `fuzme` can also be used for clustering using Mahalanobis distances. Clustering using Mahalanobis distances can also be achieved with function `fanny` of R package `cluster`. My experience is that computing time with these R packages is prohibitive when we have a large number of cells. In that case I recommend the software `FuzME`, which can be downloaded from internet at https://sydney.edu.au/agriculture/pal/software/fuzme.shtml. For postprocessing of the memberships to select sampling points, see FKMSampling.docx and R script `FKMSample_FuzME.R` in the supplement.

Fuzzy k-means with extragrades. As noted by de Gruijter et al. (2010), with hard and fuzzy k-means sampling the selected sampling points will tend to be concentrated in those parts of the multivariate distribution where the density of points is largest. The multivariate distribution is well represented by the sample, however no points are selected in the extremes of the distribution where the density of points is low. These points with extreme values, either near the minimum or near the maximum, for all or most covariates can have a considerable effect on the quality of the calibrated model. To overcome this problem, de Gruijter et al. (2010) proposed fuzzy k-means with extragrades. In this clustering method besides the k subsets of points represented by a centroid, an extra fuzzy subset is created with multivariate extremes or outliers. This fuzzy subset is not represented by a centroid; what the points share is that they are all distant from the k centroids. Finally, the k locations with the largest membership in the respective regular subsets are selected and completed by one or more locations with the largest memberships in the extra subset (de Gruijter et al., 2010). I am not aware of applications yet of this sampling design.

Fuzzy k-means with extragrades can be done with function `fkme` of R package `fuzme`. Again, computing time can become prohibitive, so that clustering with `FuZME` becomes attractive.

4. Adapted experimental designs

This section describes two experimental designs that have been adapted for spatial surveys. An adaptation was necessary because in contrast to experiments, in observational studies one is not free to choose combinations of levels of different factors. When two covariates are strongly correlated it may happen that there are no locations with a relatively large value for one covariate and a relatively small value for the other covariate.

In a full factorial design all combinations of factor levels are observed. For instance, suppose we have only two covariates, e.g. application rates for N and P in agricultural experiment, and four levels for each covariate. It is evident that the best option is to have multiple plots for all 4×4 combinations. This is referred to as a full factorial design. With k factors and l levels per factor the total number of observations is l^k. With numerous factors and/or numerous levels per factor this becomes unfeasible in practice. Alternative designs have been developed that need less observations but still provide detailed information about how the variable of interest responds to changes in the factor levels. Examples are Latin hypercube samples and response surface designs. The survey sampling analogues of these experimental designs are now described.

In a final subsection the Kennard-Stone design is described (Kennard and Stone, 1969). Although this design was proposed for experiments, this design can be used without adaptations as a sampling design in observational research.

More experimental designs have been applied in soil survey, for instance D-optimal designs, see Totaro et al. (2013) for an interesting application of this design.

4.1. Conditioned Latin hypercube sampling

Latin hypercube sampling (LHS) is used in designing (computer) experiments with numerous covariates and/or factors of which we want to study the effect on the output (McKay et al., 1979). With numerous covariates and/or levels per covariate, a full factorial design becomes unfeasible. A much cheaper alternative then is an experiment with, for all covariates, exactly one observation per level. So in the agricultural experiment this would entail four observations, distributed in a square in such way that we have in all rows and in all columns one observation. This is referred to as a Latin square. The generalisation of a Latin square to a higher number

of dimensions is a Latin hypercube.

Minasny and McBratney (2006) adapted LHS for observational studies; this adaptation is referred to as conditioned LHS (cLHS). For each covariate a series of intervals (marginal strata) is defined. The breaks of the marginal strata are chosen such that the numbers of pixels in these marginal strata are equal. This can be done by using the quantiles corresponding with evenly spaced cumulative probabilities as stratum breaks. For instance, for five marginal strata we use the quantiles corresponding with the cumulative probabilities 0.2, 0.4, 0.6 and 0.8.

Minasny and McBratney (2006) developed a search algorithm, based on heuristic rules and an annealing schedule, to select a cLHS (see for an explanation of annealing, section 5.2 hereafter). The objective function that is minimized is the weighted sum of three components, one of which is the sum over all marginal strata of the absolute difference between the marginal stratum sample size and targeted sample size (equal to 1). A second criterion is the sum over all entries of the matrix with absolute values of the difference between the correlation of the covariates in the population and in the sample. A third criterion is involved only when we have, besides quantitative covariates, categorical variables. This third component is the sum over all classes of the absolute difference between the sample proportion of a given class and the population proportion of that class.

With cLHS the marginal distributions of the covariates in the sample are close to these distributions in the population. This can be advantageous for mapping methods that do not rely on linear relations, for instance in machine learning techniques like classification and regression trees (CART), and random forests.

Figure 3 shows a cLHS sample of 20 points from the Hunter valley study area, using the same five covariates as before in k-means sampling. In Figure 4 the cLHS sample is plotted in a scatter diagram of elevation against cti. Besides the marginal

strata are shown. Ideally, each column and each row contains one sampling point.

Conditioned LHS is a very popular sampling design in digital soil mapping. Roudier et al. (2012) and Mulder et al. (2013) adapted cLHS to make it more suitable for areas in which some parts are difficult to access, think of remote and mountainous areas. Ramirez-Lopez et al. (2014) compared cLHS with fuzzy k-means and Kennard-Stone sampling for calibration of models for predicting clay content and Ca concentration at the field and regional scale, using soil spectroscopy as input. Schmidt et al. (2014) compared an extension of cLHS with fuzzy k-means and response surface sampling for calibration of model for predicting basic soil properties at the field scale, using electromagnetic induction (EM38 and EM31) and gamma spectroscopy (U, K, Th) data.

cLHS samples can be selected with R package `clhs` (Roudier, 2011) and function `optimCLHS` of R package `spsann` (Samuel-Rosa, 2016). For an application of the latter package, see `cLHS_spsann.R` in the supplement. Both R packages cannot be used to design a cLHS sample in the presence of legacy data. When we have legacy data we do not want to sample marginal strata that are already covered by these legacy sample data. Conditioned Latin hypercube infill sampling can be done with function `getCriterion.cLHS` of `Functions4SSA.R`, which is called by `cLHS.R` (see supplement). In both functions `optimCLHS` of R package `spsann` and `getCriterion.cLHS` of `Functions4SSA.R` the first and second component of the minimization criterion (O1 and O2) are not computed as sums but as means. For O2 this mean is computed over the off-diagonal elements of the matrix.

4.2. Response surface sampling

With response surface designs we aim at finding an optimum of the response within specified ranges of the factors. There are many types of response surface

designs, see Myers et al. (2002). A commonly used response surface design is the central composite design; the data of this design are used to fit a curved, quadratic surface (multiple linear regression model with quadratic terms).

Lesch et al. (1995) adapted the response surface methodology so that it can be applied in observational studies. Several problems needed to be tackled. First, when multiple covariates are used, the covariates must be decorrelated. Second, sampling locations may show strong spatial clustering, so that the assumption in linear regression modelling of spatially uncorrelated model residuals is violated. To tackle these two problems Lesch et al. (1995) proposed the following procedure (see also Lesch (2005)):

- Transform covariate matrix into a scaled, centered, de-correlated matrix by principal components analysis (PCA)
- Choose response surface design type. This leads to a set of combinations of factor levels, referred to as design-points
- Select candidate sampling locations based on the distance from the design-points in PC-space. Select multiple locations per design-point
- Select combination of candidate sampling locations with the highest value for a criterion that quantifies how uniform the sample is spread across the study area

Lesch (2005) proposed three maximization criteria that can be used in the final step: 1. the average separation distance between sampling locations; 2. the geometric mean separation distance, and 3. the minimum separation distance. The response surface sampling approach is an example of a model-based sampling design. From that viewpoint I should have described this sampling design in the next section.

With response surface sampling one assumes that some type of low order (linear or quadratic) regression model can be used to accurately approximate the relationship between the soil variable of interest and the covariates. The sampling locations are then selected to implicitly optimize the estimation of this model, subject to satisfying one or more explicit spatial optimization criteria (Lesch et al., 1995).

Note that in linear regression modeling one assumes that the data are independent. Optimization of the sampling design under this model will not prevent the locations for spatial clustering, see section 8. However, in reality the assumption of independent data might be violated when the sampling locations are spatially clustered. For that reason the response surface sampling design selects samples with good spatial coverage, so that the design becomes robust against violation of the independence assumption.

This design has been applied for mapping soil salinity (ECe), using electromagnetic induction (EMI) measurements and surface array conductivity measurements as predictors in multiple linear regression models. For applications, see Corwin and Lesch (2005), Lesch (2005), Fitzgerald et al. (2006), Corwin et al. (2010) and Fitzgerald (2010).

This sampling design is illustrated with the Cotton Research Field in Uzbekistan. We used the software ESAP (Lesch et al., 2000) to select a response surface sample of 12 points. ECa was measured with the EM device in vertical dipole mode with transmitters at 1 m and 50 cm from the receiver, on transects covering the Cotton Research Field (Figure 5). The natural logs of the two EM measurements are first interpolated to a fine grid by ordinary kriging. These interpolated EM data are then used to design the response surface sample. The two covariates are strongly correlated, $r = 0.73$. Figure 5 shows the selected sample plotted on the interpolated EM measurements. Figure 6 shows the selected response surface sample, plotted in

the space spanned by the two principal components, and in the scatter diagram of the two original covariates. The sample sizes that can be chosen in ESAP are 6, 12 or 20 points.

4.3. Kennard-Stone sampling

One of the motivations for this experimental design was that in experiments often only part of the space spanned by the factors can be covered by the design points. To circumvent this problem the Kennard-Stone design (KS) starts from a finite $N \times p$ matrix of points that discretise the factor space, with N the number of candidate points, and p the number of factors. A geometric criterion is used to select a subset of n candidate points that are used as design points. The response is observed for the combinations of factors at these design points.

The selection of the design points goes as follows. First two candidate points are selected with a maximum separation distance in factor space. The third point that is selected from the $N-2$ candidate points has maximum distance from the first two design points, *et cetera*. Kennard and Stone (1969) recommends to harmonize the dimensions of the factors by scaling them. They also suggest to take correlation of the factors into account by transforming the factors into orthogonal variables, and measuring distances in this transformed factor space.

This design is commonly used to select a subsample out of a large sample with spectroscopy data (spectral library) to calibrate a model relating a soil property of interest to the spectra, see for instance Viscarra Rossel and Brus (2018) and Riedel et al. (2018).

KS samples can be selected with function `ken.sto` of R package `soil.spec` (Sila et al., 2014).

5. Model-based sampling

5.1. Optimization of grid spacing

The alternative to deriving the grid spacing from the available budget is to derive the spacing from a requirement on the precision of the map. Suppose that the maximum variance of the prediction errors may not exceed a given threshold. The question then is what is the tolerable grid spacing so that the maximum prediction error variance does not exceed this threshold. Ignoring the relatively large variances near the border of the study area, we expect the prediction error variance to be largest at the centres of the grid cells with the measurements at their corners; these points have the largest distance to the points of the sampling grid. The larger the grid spacing, the larger the prediction error variances at these centres. The question is how large the spacing can be, so that the maximum prediction error variance is just below the threshold.

For finding this maximum grid spacing one must have prior knowledge of the spatial variation. First, I consider the situation in which it is reasonable to assume that the mean of the study variable in the area is constant, and that we have a prior variogram, for instance estimated from existing data from the study area or from data of similar areas.

There is no simple equation that relates the grid spacing to the variance of the prediction error (kriging variance). What can be done, is to calculate the kriging variance for a range of grid spacings, plot the kriging variances at the cell centres against the grid spacing, and use this plot inversely to determine, given a constraint on the maximum kriging variance, the maximum grid spacing (Burgess and McBratney, 1981; McBratney et al., 1981; McBratney and Webster, 1981).

For a requirement on the mean kriging variance instead of the maximum kriging

variance, I propose the following procedure:

1. Specify variogram type and parameter values
2. Select a simple random sample of points
3. Select a square grid with a given spacing
4. Compute the ordinary kriging variance at the simple random sample of evaluation points
5. Compute the sample average of the kriging variance
6. Repeat this for other grid spacings

The simple random sample of step 2 is used to *estimate* the population mean of the kriging variance (MKV). The sample should be large enough, say > 1000 points, so that the estimate has high precision. In step 3 the square grid can be selected using either a fixed or a random starting point. In the latter case, steps 3 - 5 must be repeated several times, leading to multiple values for the estimated MKV for each grid spacing. Note that the procedure is very general, and can also be used to determine the tolerable grid spacing for, for instance, the P95 of the kriging variance.

The same procedure can also be used to decide on the tolerable grid spacing for kriging with an external drift (KED). In this case the kriging variance is not only determined by the spatial coordinates of the grid nodes and evaluation points, but also by the covariate values at these points.

Figure 7 shows graphs of the mean variance for OK and KED versus the grid spacing for the Ethiopia case study. The expected sample sizes for the grid spacings range from 432 (5 km spacing) to 76 points (12 km spacing). In KED I used elevation, NIR, visible infrared, and land surface temperature as covariates. A large part of the variation is explained by the four covariates, and as a result for a given required MKV the tolerable grid spacing with KED is considerably larger than for OK. For

KED up to a spacing of about nine km, corresponding with an expected sample size of 134 points, the variance of the error in the interpolated residuals dominates the kriging variance. With wider spacings the contribution of the uncertainty in the estimated regression coefficients becomes more substantial, explaining the somewhat accelerated increase of the MKV beyond a spacing of nine km.

In practice we do not know the variogram. In the best case we have prior data that can be used to estimate the variogram. However, even in this case we are uncertain about the variogram type and the variogram parameters. Recently, Lark et al. (2017) worked out a Bayesian approach to account for this uncertainty. A sample from the multivariate posterior distribution of the variogram parameters is obtained by Markov Chain Monte Carlo (MCMC). Each unit of the sample is used to compute the kriging variances at the centre of square grid cells where the kriging variance is maximum. On his turn, each value of the kriging variance can be used to compute a tolerable grid spacing. The same procedure can be used using the mean kriging variance as a quality criterion. Figure 8 shows the histogram of the tolerable grid spacing for a mean ordinary kriging variance of 0.8 for the Ethiopia case study. The posterior distribution of the parameters of a spherical model with nugget was sampled by MCMC and differential evolution (ter Braak and Vrugt, 2008). Prior distributions for the sill variance, proportion of variance that is spatially structured, and range were all uniform with lower bounds equal to zero and upper bounds equal to 5 $(\mathrm{mg/kg})^2$, 1 and 100 km. The grid spacing with the largest number of MCMC samples equals 8 km, which corresponds with the tolerable grid spacing derived from Figure 7. The subfigure on the right in Figure 8 shows the proportion of MCMC samples with a MKV smaller or equal to the target MKV of 0.8, as a function of the grid spacing. If we require a probability of 80% that the MKV does not exceed the target MKV of 0.8, the tolerable grid spacing is about 6.25 km. With a grid spacing

of 8 km as determined from Figure 7, the probability that the MKV exceeds 0.8 is only about 55%.

Once we have decided on the required spacing, we may calculate from this the required sample size, or with a random start the expected sample size. We then may further optimize the design, by relaxing the constraint that the sampling locations must be on a square grid, and optimizing the coordinates of the locations. This can either be done by computing a spatial coverage sample, see section 2, or by model-based optimization of the sampling locations by spatial simulated annealing, see next section.

Function `ossfim` of package `gstat` (Pebesma, 2004) can be used for model-based optimization of the grid spacing, given a requirement on the maximum ordinary kriging variance (kriging variance at centre of cells). For optimizing the grid spacing given a requirement on the mean ordinary kriging variance or mean variance for kriging with an external drift, see `ModelBasedGridSpacingOK_MeanKV.R` and `ModelBasedGridSpacingKED_MeanKV.R` in the supplement. For the Bayesian approach of optimization of the grid spacing, see `Bayesian_GridSpacing.R`.

5.2. Optimization of coordinates of sampling locations

As argued in section 3.2, sampling on a regular grid can be suboptimal. I showed how the spatial coordinates of the sampling locations can be optimized by minimizing a geometric criterion, the MSSD, through k-means. This section describes optimization of the spatial coordinates of the sampling points through minimization of a criterion defined in terms of the prediction error variance, e.g. the mean kriging variance. Optimization by k-means as in spatial coverage sampling cannot be used for this. Inspired by the potentials of optimization through simulated annealing (Kirkpatrick et al., 1983; Aarts and Korst, 1987), van Groenigen and Stein (1998)

proposed to optimize the locations by spatial simulated annealing (SSA), see also van Groenigen et al. (1999, 2000). This is an iterative, random search procedure, in which a sequence of samples is generated. A new sample (proposed sample) is obtained by slightly modifying the current sample. One sampling location of the current sample is randomly selected, and this location is shifted to a random location within the neighbourhood of the selected location. The minimization criterion is computed for each sample. If the criterion of the proposed sample is smaller, it is accepted. If the criterion is larger, the proposed sample is accepted with a probability that is a function of the increase (the larger the increase, the smaller the acceptance probability) and of an annealing schedule parameter, referred to as the temperature, T. The larger T, the larger the probability that a proposed sample with a given increase of the criterion, is accepted. T is gradually decreased during the optimization, so that the acceptance probability of worse samples approaches zero towards the end of the optimization.

Minimization of the mean kriging variance (MKV) for ordinary kriging (OK) by SSA leads to a sample that is spread out throughout the area. Brus et al. (2007) found that the optimized samples were very similar to spatial coverage samples, and that the MKV were nearly equal. Figure 10 shows a model-based infill sample of 100 points for Ethiopia. The legacy data were used to estimate a variogram for SOM. The fitted spherical variogram had a nugget of 0.62, a partial sill of 0.56 and a range of 45 km. Comparison with the spatial infill sample of Fig. 2 shows that in a much wider zone on both sides of the roads no new sampling points are selected. This can be explained by the large range of the variogram.

Heuvelink et al. (2007) optimized the locations by SSA for kriging with an external drift (KED). Remember that in KED we assume that the mean of the variable of interest is a linear combination of one or more covariates of which we have a map

covering the area. Brus and Heuvelink (2007) showed that the optimized sample is a compromise between spreading in geographic space and feature space. More precisely, locations are selected by spreading them out throughout the study area, while accounting for the values of the covariates at the selected locations, in the sense that locations with covariate values near the minimum and maximum are preferred. This can be explained by noting that the variance of the KED prediction error can be decomposed in the variance of the interpolated residuals and the variance of the estimated mean. The contribution of the first variance component is minimized through geographical spreading, that of the second component by selecting locations with covariate values near the minimum and maximum. Figure 9 shows that the smaller the proportion of spatially structured variance, the more the sampling points shift towards the left and right side of the square where the covariate (Easting) has its minimum and maximum value, respectively.

Note that for optimizing the sampling locations for KED we must decide on the covariates that, we expect, explain part of the variation of the soil variable of interest. When one or more covariates are used in sample optimization, but not used in KED once the data are collected, the sample is suboptimal for the model used in prediction. Reversely, ignoring a covariate in sample optimization while using this covariate as a predictor, also leads to suboptimal samples.

Further, note that a sample with covariate values close to the minima and maxima only is not desirable if we do not want to rely on the assumption of a linear relation between the soil property of interest and the covariates. To identify a non-linear relation locations with intermediate covariate values are needed. Optimization using a variogram with clear spatial structure leads to geographical spreading of the sampling locations, so that most likely also locations with intermediate covariate values will be selected.

Figure 11 shows a sample of 50 points from the Cotton Research Field in Uzbekistan, optimized for KED of ECe using EMv1m as a covariate. The natural log of the EMv1m measurements (Fig. 6) are interpolated first to a square grid, and these interpolated values are used as a covariate in KED. The residual variogram for the natural log of ECe, the variable of interest, used in SSA is exponential with nugget 0.1, partial sill 0.075 and a distance parameter of 100 m (practical range 300 m).The good spreading in geographic space is immediately clear; a careful look shows that preferably locations with either very small or very large values of ln(EMv1m) are selected, disturbing locally the regular pattern. The pushing of the locations towards the margins of the distribution is evident when comparing the population and sample histogram of ln(EMv1m) (see Figure 1 in the supplement).

Function `optimMKV` of package `spsann` (Samuel-Rosa, 2016) can be used for model-based optimization of the coordinates of sampling locations, both for OK and KED, see R script `ModelBasedSample_KED_spsann.R` in the supplement. In the current version legacy data cannot be accounted for. R scripts `ModelBasedSample_SSA_OK.R` and `ModelBasedSample_SSA_KED.R` can be used for optimization of the locations of an infill sample in situations with legacy data. These R scripts call function `getCriterion.K` in `functions4SSA.R`.

6. Sampling for variogram estimation

For model-based sampling as described in sections 5.1 and 5.2 we need to specify the (residual) variogram. In cases we do not have the faintest idea, we might want to collect first data with the specific aim of estimating the variogram. This variogram is subsequently used to design a model-based sample for mapping. This section is about how to design this reconnaissance sample survey for estimating the variogram.

The first question is how many observations we need for this. Webster and Oliver (1992) gave as a rule of thumb that 150-225 points are needed to obtain a reliable variogram when estimated by the method-of-moments. Lark (2000) showed that with maximum likelihood (ML) estimation two-third to only half of the observations are needed to achieve equal precision of the estimated variogram parameters. Once we have decided on the sample size, we must select the locations. Two random sampling designs for variogram estimation are described in this section, nested sampling and independent sampling of pairs of points.

6.1. Nested sampling

Nested sampling can be used to estimate the semivariance at several chosen separation distances (Oliver and Webster, 1986; Webster et al., 2006).

We must first decide on these separation distances. Usually separation distances are chosen in a geometric progression, for instance 2, 8, 32, 128 and 512 m. The multiplier should be at least three. There are two implementations of nested sampling. In the first implementation, in the first stage several main stations are selected in a way that they cover the study area well, for instance by spatial coverage sampling. In the second stage each of the main stations is used as a starting point to select one point at a distance equal to the largest chosen separation distance (512 m in the example), in a random direction from the main station. This doubles the sample size. In the third stage at each of the points selected in the previous stages (main stations of stage 1 plus the points of stage 2) are used as starting points to select one point at a distance equal to the second largest separation distance, and so on. All points selected in the various stages are included in the nested sample.

The first stage of the second implementation is equal to that of the first implementation. In the second stage each of the main stations serves as a starting point

for randomly selecting a pair of points with a separation distance equal to the largest chosen separation distance, with the main station halfway. In the third stage each of the substations is used to select in the same way a pair of points separated by the second largest chosen distance, and so on. Only the points selected in the final stage are used as sampling points. Figure 12 shows a nested sample selected by this second approach. For illustration purposes, only one main station is selected (halfway the two stations with label 1). In total 16 points are selected in four stages. The stations that served as starting points in stage 1 to 3 for selecting pairs of points are also shown.

The sample of Figure 12 is an example of a balanced nested sample from the Hunter Valley case study area: in all stages all stations selected in the previous stage are used to select a pair of points. If in the first implementation of nested sampling all points selected in all previous stages are used to select a new point, then this also results in a balanced nested sample. The number of pairs of points separated by a given distance doubles with every stage. As a consequence, the estimated semivariances for the smallest separation distance are much more precise than for the largest distance. We are most uncertain about the estimated semivariances for the largest separation distances. If in the first stage only one pair of points is selected separated by the largest distance, then we have only one degree of freedom for estimating the variance component associated with this stage. It is more efficient to select more than one main station, say about ten, and to select less points in the final stages. For instance, with the second implementation we may decide to select a pair of points at only half the number of stations selected in the one-but-last stage. The nested sample then becomes unbalanced.

The model for nested sampling with four stages is a hierarchical ANOVA model

with random effects:

$$Z_{ijk} = \mu + A_i + B_{ij} + C_{ijk} + \epsilon_{ijkl} \tag{1}$$

with μ the mean, A_i the effect of the ith first stage station, B_{ij} the effect of the jth second stage station within the ith first stage station, and so on. A_i, B_{ij}, C_{ijk} and ϵ_{ijkl} are random quantities (random effects), all with zero mean, and variances σ_1^2, σ_2^2, σ_3^2 and σ_4^2 respectively.

For balanced designs the variance components can be estimated by the method of moments from a hierarchical ANOVA. The first step is to assign factors to the sampling points that indicate the grouping of the sampling points in the various stages. The number of factors needed is the number of stages minus 1. To illustrate this, in Figure 12 the first factor has two levels (in Eq. 1 $i = 1, 2$), the second factor has four levels (in Eq. 1 $j = 1, 2, 3, 4$) and the third factor has eight levels (in Eq. 1 $k = 1, 2, \cdots, 8$). For unbalanced nested designs the variance components can be estimated by restricted maximum likelihood (REML) (Webster et al., 2006). REML estimation is also recommended if in Eq. 1 iunstead of a constant mean μ the mean is a linear combination of one or more covariates (fixed effects). The semivariances at the chosen separation distances are obtained by cumulating the estimated variance components.

Random sampling of the points is not strictly needed because a model-based approach is followed here (the model of Eq. 1 is a superpopulation model, i.e we assume that our population is generated by this model). Papritz et al. (2011), for instance, selected the points (using the second implementation) non-randomly to improve the control of the nested subareas and the average separation distances.

Lark (2011) describes a method for optimization of a nested design, given the total number of points and the chosen separation distances.

The R script `NestedSampling_v1.R` in the supplement can be used to select a balanced nested sample, using the first implementation of nested sampling. The R script `NestedSampling_v2.R` can be used to select balanced and unbalanced nested samples, using the second implementation.

6.2. Independent sampling of pairs of points

With the nested design the estimated semivariances for the different separation distances are not independent. Independent estimated semivariances can be obtained by independent selection of pairs of points (IPP sampling) as proposed by Brus and de Gruijter (1994). For simple random sampling of point pairs this method is very straightforward. For each separation distance a point pair is selected by first selecting fully randomly one point from the study area. Then the second point is randomly selected randomly from the circle with the first point at its centre and a radius equal to the chosen separation distance. If this second point is outside the study area, both points are ignored. This is repeated until we have the required point pairs for this separation distance.

The R script `SI_PointPairs.R` can be used to select simple random samples of pairs of points for variogram estimation. In this R script bootstrap samples of the samples of point pairs are used to estimate the variances and covariances of the estimated semivariogram model parameters.

6.3. Model-based sampling for variogram estimation

There is rich literature on model-based optimization of the sampling locations for variogram estimation. Several design criteria (minimization criteria) have been proposed for optimizing the sample, such as the determinant of the variance covariance matrix of variogram parameters estimated by generalized least squares to the

experimental method-of-moments variogram (Müller and Zimmerman, 1999; Bogaert and Russo, 1999), the log determinant of the inverse Fisher information matrix in maximum likelihood (ML) estimation of the variogram (hereafter shortly denoted by logdet(F^{-1})) (Zhu and Stein, 2005), and the variance of the kriging variance at the centre of square grid due to uncertainty in the ML estimates of the variogram parameters (hereafter shortly denoted by $V(\sigma_K^2)$) (Lark, 2002). This variance is approximated by a first order Taylor series, requiring the partial derivates of the kriging variance to the variogram parameters. All these minimization criteria are a function of the variogram parameters θ, showing that the problem is circular. Using a preliminary 'estimate' of the variogram parameters, $\hat{\theta}$ leads to a locally optimal design at $\hat{\theta}$. For that reason Bogaert and Russo (1999) and Zhu and Stein (2005) proposed a Bayesian approach in which a multivariate prior distribution for the variogram parameters is postulated, and the expected value over this distribution of the criterion is minimized.

Figure 13 shows for the Hunter Valley case study area samples of 100 points, the locations of which are optimized by SSA, using logdet(F^{-1}) (left subfigure) or $V(\sigma_K^2)$ (right subfigure) as a minimization criterion. The postulated variogram is exponential with a range of 500 m and a nugget-to-sill ratio of 0.2. For both criteria the points show strong spatial clustering: nearly all points have one or more points at a very short distance (< 2 m).

R script `ModelBasedSample_SSA_EK.R` (which calls `Functions4SSA.R`) in the supplement can be used to design a model-based sample for variogram estimation. Either logdet(F^{-1}), or $V(\sigma_K^2)$ can be selected as a minimization criterion.

6.4. One sample both for estimating model parameters and prediction

In practice, often a reconnaissance survey for variogram estimation is not feasible, and a single sample must be designed that is suitable both for estimating the model parameters and prediction with the estimated model parameters. Another reason is that in a reconnaissance survey we seldom can afford a sample size large enough to obtain reliable estimates of the model parameters. Papritz et al. (2011) found that for a sample size of 192 points the estimated variance components with balanced and unbalanced nested designs were highly uncertain. For this reason it is attractive to use also the sampling points designed for spatial prediction (mapping) for estimating the variogram. From this it follows that designing two samples, one for estimating the variogram and one for spatial prediction, is suboptimal. Designing one sample that can be used both for estimation of the model parameters and for prediction potentially is more efficient.

Finally, with nested sampling and IPP sampling we aim at estimating the variogram of the 'residuals' of a constant mean (see Eq. 1). In other words, with these designs we aim at estimating the parameters of model used in ordinary kriging. In situations where we have covariates that can partly explain the spatial variation of the soil variable of interest, kriging with an external drift is more appropriate. In these situations the reconnaissance survey should be tailored at estimating both the regression coefficients associated with the covariates and the parameters of the residual variogram.

Model-based methods for designing a single sample for estimating the model parameters and for prediction with the estimated model parameters are proposed, amongst others, by Zimmerman (2006), Zhu and Stein (2006), Zhu and Zhang (2006) and Marchant and Lark (2007). The methods use a different minimization criterion. Zimmerman (2006) proposed to minimize the kriging variance (at the centre of a

square grid) that is augmented by an amount that accounts for the additional uncertainty in the kriging predictions due to uncertainty in the (residual) variogram parameters (hereafter denoted by $\sigma^2_{K^+}$). The uncertainty in the ML estimates of the variogram parameters is estimated by the inverse of the Fisher information matrix. Marchant and Lark (2007) proposed the same criterion, but following Zhu and Stein (2005), accounted for uncertainty in the postulated preliminary variogram by adopting a Bayesian approach. Zhu and Stein (2006) proposed as a minimization criterion the Estimation Adjusted Criterion (EAC), which is the spatial average of a weighted sum of the variance of the prediction error (including a term that accounts for uncertainty about the variogram parameters as in Zimmerman (2006)) and the variance of the kriging variance (quantified in the same way as by Lark (2002)).

Computing time for optimization of the coordinates of a large sample, say > 50 points, can become prohibitively large. To reduce computing time Zhu and Stein (2006) proposed a two-step approach. In the first step, for a fixed proportion $p \in (0, 1)$ the locations of $(1-p)n$ points are optimized for prediction with given parameters, for instance by minimizing MKV. This 'prediction sample' is supplemented with pn points, so that the two combined samples of size n minimize $\text{logdet}(F^{-1})$ or $V(\sigma^2_K)$). This is repeated for different values of p. In the second step EAC is computed for the combined samples of size n, and the proportion and associated sample with minimum EAC is selected.

A simplification of this two-step approach is to select in the first step a spatial coverage sample (obtained by minimizing MSSD), and to supplement this by a fixed number of points whose coordinates are optimized by SSA, using EAC computed from both samples (spatial coverage + supplemental sample) as a minimization criterion. In SSA the spatial coverage sample is fixed, i.e. the locations are not further optimized. Lark and Marchant (2018) recommended as a rule of thumb to add about

10% of the spatial coverage sample as short distance points. Figure 14 shows for the Hunter Valley case study area spatial coverage samples of 90 points, supplemented by 10 points optimized by SSA, using $\sigma^2_{K^+}$ (left subfigure) or EAC (right subfigure) as a minimization criterion.

Figure 14 shows that all, or nearly all supplemental points are very close to a point of the spatial coverage sample. Based on this, a very straightforward, simple sampling design for estimating the model parameters and for prediction is a spatial coverage sample supplemented with randomly selected points in between the points of the spatial coverage sample at some chosen, fixed distances. Figure 15 shows an example. A subsample of 10 points is selected from the 90 points of the spatial coverage sample, using simple random sampling without replacement. These points are used as a starting point to select a close distance point in a random direction. R script `SpatialCoveragePlusSample.R` in the supplement can be used to select such samples.

R script `ModelBasedSample_SSA_EK.R` in the supplement can be used to design a model-based sample both for estimation of the variogram and for kriging. The core of the sample is a spatial coverage sample, to which a fixed number of sampling points is added. The locations of the supplemental sample are optimized given the locations of the spatial coverage sample. Both above mentioned minimization criteria ($\sigma^2_{K^+}$ and EAC) are implemented in function `getCriterion.EK` of `Functions4SSA.R` which is called by `ModelBasedSample_SSA_EK.R`.

7. Sampling for validation

An important step in a mapping project is the validation of the model and evaluation of the quality of the map. As argued by Brus et al. (2011) this can best be

done by collecting additional data, not used for mapping, through probability sampling. This is superior to validation through data splitting or cross-validation, as the samples used for mapping, and subsequently for data splitting or cross validation, generally are not probability samples. Probability sampling enhances model-free, design-based estimation of map quality indices, such as overall and map unit purity of categorical maps and the population mean error (ME) and population mean squared error (MSE), as well of our uncertainties about these estimates, expressed, for instance, as a standard error or a confidence interval.

I illustrate map validation with the case study Xuancheng. Using 121 observations of soil organic matter concentration (g/kg) in the A-horizon, two maps are made, one with a random forest model (RF), and one with KED (see Figure 2 in the supplement). For the RF model seven covariates are used: planar curvature, profile curvature, slope, temperature, precipitation, topographic wetness index and elevation. In KED only the two most important covariates in the RF model are used: precipitation and elevation. For validating the two maps a stratified random sample was selected of 62 units, using eight map units of the geological map as strata (Fig. 16). The population ME of both maps was estimated by

$$\widehat{ME} = \sum_{h=1}^{L} w_h \widehat{ME}_h \tag{2}$$

with $w_h = N_h/N$ the relative size of stratum h (N_h is number of pixels in stratum h, N is total number of pixels in study area), and $\widehat{ME}_h$ the estimated mean error of stratum h:

$$\widehat{ME}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}, \tag{3}$$

with e_{hi} the error of validation unit i in stratum h: $e_{hi} = z_{hi} - \hat{z}_{hi}$. Note that $\widehat{ME}_h$ is simply the unweighted sample average of the errors in stratum h.

The variance of the estimated ME was estimated by

$$\widehat{V}(\widehat{ME}) = \sum_{h=1}^{L} w_h^2 \widehat{V}(\widehat{ME}_h) \tag{4}$$

with $\widehat{V}(\widehat{ME}_h)$ the estimated variance of the estimated ME in stratum h:

$$\widehat{V}(\widehat{ME}_h) = \frac{s_h^2(e)}{n_h} \tag{5}$$

with $s_h^2(e)$ the sample variance of the errors in stratum h. By taking the square root we obtain an estimate of the standard error of the estimated ME. The population MSE and its standard error can be estimated by the same formulas, replacing the errors in Eq. 3 by squared errors.

A problem in estimating the standard error is that there is one stratum with only one observation. Following Cochran (1977) we collapsed this stratum with a similar geological map unit stratum. Note that after collapsing the stratum weights w_h must be adapted, so that they sum to one again.

The estimated population mean errors were used to test the null-hypothesis $ME = 0$, against the two-sided alternative hypothesis $ME \neq 0$. In words the null-hypothesis states 'the predictions are unbiased', or 'there is no systematic error in the predictions'. This hypothesis can be tested with a one-sample t test. The number of degrees of freedom can be approximated by $n - L$, with L the number of strata (Lohr, 1999). For both maps the null-hypothesis is not rejected (p-value $>> \alpha = 0.05$), so there is no evidence at all for biased predictions, neither for RF, nor for KED (Table 2).

I also tested the null-hypothesis $MSE(KED) = MSE(RF)$. As alternative hypothesis we chose $MSE(KED) > MSE(RF)$, because in KED only two covariates are used as predictors, and besides in KED we assume a linear relation between SOM and the covariates, which can be too restrictive (in RF no such assumption is made).

This hypothesis is tested by a paired t-test, i.e. for each validation unit the difference of the two squared errors is computed: $d_i = e_i^2(KED) - e_i^2(RF)$. The null-hypothesis can now be reformulated as $MD = 0$, with MD the population mean of the pairwise differences in squared erors; the alternative hypothesis is $MD > 0$. In this procedure we automatically account for correlation of the two squared errors. To our surprise the estimated MSE with KED is smaller than with RF. The t-value is -0.632, with a p-value of 0.735. If we test the null-hypothesis against the two-sided alternative $MSE(KED) \neq MSE(RF)$ the p-value equals 0.530, so that we conclude that we have no evidence that the population MSE of the RF map is smaller than that of the KED map.

When parts of the mapped area are difficult to access, think of remote areas, rough terrain conditions, cost-efficiency of the validation can be increased by accounting for these access costs (Yang et al., 2018). In stratified random sampling we may take the differences in access costs in allocating the total sample size to the strata. Besides, these access costs can be used to construct the strata (Yang et al., 2018).

Stratified random samples can be selected with function `strata` of R package `sampling` (Tillé and Matei, 2015). Many other probability sampling designs are implemented in this R package. These packages select the units from a data.frame, which implies that the population is considered finite, whereas in reality we have an infinite population of points. In our case the units are often the nodes of a fine grid discretising the area, or the cells of a raster map. After the random selecting of the nodes (raster cells), a random point location is selected within the selected raster cells. A simple random sample of units can be selected by the function `sample.int` of the `base` package. Optimal stratifications can be computed with R package `stratification` (Baillargeon and Rivest, 2014).

`StratifiedRandomSampling.R` in the supplement is an R script for selecting a

stratified simple random sample. `Validation.R` is used to estimate the ME and MSE of both maps of Xuancheng, and `StatisticalTesting.R` is used for testing the hypotheses about the ME and MSE.

8. Choosing a sampling design and further research

8.1. No single best sampling design

There is no single best sampling design for digital soil mapping. The best design depends on the method used for mapping the soil. This is illustrated with Figure 1. We have seen before that the optimized sample for mapping with a simple linear regression model contains the units with the smallest or the largest values of the covariate x. In this case the optimized sample shows strong spatial clustering. Spatial clustering is not avoided because in a simple linear regression model we assume that the data are independent. In the optimized sample for mapping by KED (for KED we need many more points, but this is just for illustration purposes) spatial clustering is avoided, the selected units are spread throughout the area. At the same time units near the minimum (unit with coordinates (13.5, 12.5)) and maximum (unit with coordinates (13.5, 6.5)) of x are selected, see also section 5.2. So if we believe that the soil can better be mapped by KED instead of simple linear regression, because we expect the data to be spatially autocorrelated, the optimized sample largely differs from the optimized sample for mapping using a simple linear regression model.

If we foresee a quadratic relation, $z_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, the optimized sample will also include locations with covariate values near the mean of x. And if we expect an even more complicated, non-linear relation, stratified sampling using quantiles of covariate x as stratum breaks, so that the distribution of x in the sample and in the population are similar (as is done in cLHS sampling for multiple covariates) can be advantageous (Figure 1).

831 *8.2. Rules for choosing*

832 Table 3 is an attempt to link mapping methods and sampling designs. As one
833 can see there is no 1:1 relation; for most mapping methods there are multiple options
834 for the sampling design.

835 Two situations are distinguished, one in which we have one or more maps with
836 covariates, one in which we have none. In the latter case the soil variable of interest
837 is necessarily mapped by some spatial interpolation technique, like ordinary kriging
838 (OK). For spatial interpolation sampling points must be evenly spread throughout
839 the area, which can be achieved by sampling on a regular grid, or even better by spa-
840 tial coverage sampling. If one has prior knowledge of the variogram, this variogram
841 can be used to optimize the grid spacing, given a requirement on the maximum or
842 mean kriging variance or any percentile of the frequency distribution of the kriging
843 variance. The tolerable grid spacing leads to a minimum sample size. This sample
844 size can subsequently be used to further optimize the locations of the sampling units,
845 through minimization of the MSSD by k-means (spatial coverage sample), or min-
846 imization of the mean kriging variance by spatial simulated annealing. For OK we
847 need a variogram, and therefore I recommend to supplement the sample with short
848 distance points as explained in subsection 6.4.

849 When we have one or more maps of covariates there are various options for
850 mapping. At a high level we may distinguish mapping methods that rely, after
851 transformation of the variable of interest and or the covariates, on the assumption
852 of a linear relation of the soil variable of interest and the covariates, from methods
853 that do not rely on such assumption. The former methods involve, amongst others,
854 prediction with a simple or multiple linear regression model (LR) and KED. KED
855 mapping requires more observations (higher sampling density) than LR mapping.
856 For LR mapping, response surface sampling (RSS) can be a good option. If we have

more than two covariates, for RSS with the software ESAP the first two principle components can be used only, which can be a limitation.

For KED, in principle the same sampling options as for OK come into scope: regular grids, spatial coverage and model-based sampling. For model-based sampling we must decide on the covariates that are used in the optimization. Besides, we must specify the residual variogram. Both choices may have an adverse effect on the quality of the sample. If one or more covariates are used in designing the sample, but not used in prediction because they do not improve predictions, the model-based sample is suboptimal. Misspecification of the distance parameter ((effective) range), and especially of the nugget-to-sill ratio of the residual variogram also affects the quality of the optimized sample. Again, supplementing the sample with short distance points for residual variogram estimation is recommended.

For mapping using machine learning with one or more covariates sampling options are (fuzzy) k-means sampling, cLHS and KS sampling.

8.3. Further research

More studies into the efficiency of alternative sampling designs for a given mapping method are needed to improve and extend Table 3. Such studies are especially needed for mapping with machine learning techniques like random forest, cubist, boosted regression, neural networks, support vector machines *et cetera*.

In many cases we may not have decided yet on the mapping method at the stage of designing the sample. It is more realistic that we postpone this decision to after the sample data are collected, so that we can use the data to select an appropriate mapping method. In this situation it is important to choose a sampling design that is robust against deviations of modeling assumptions. For instance, if we neither want to rely on the assumption of a linear relation, nor on the assumption of independent

residuals, good options can be (fuzzy) k-means and cLHS sampling in which the sampling points are also spread in geographic space. A simple and straightforward way of achieving this is to add the spatial coordinates to the set of covariates, see Gao et al. (2016) for an example.

Often interest is not only in a single soil variable, but in multiple soil variables. Vășát et al. (2010) used a linear model of coregionalisation to optimize the sample size and coordinates of sampling locations for mapping with ordinary cokriging. They applied the method in a situation where prior data are available to calibrate the model, but when no or few prior data are available, the postulated model used in the optimization can be rather hypothetical. I welcome more research in this area. An alternative to designing a sampling scheme for mapping multiple quantitative soil properties, is to design a sampling scheme for mapping soil classes or fuzzy memberships. Predicted soil classes or fuzzy memberships can then be used to predict the soil properties. Studies into efficient sampling designs for mapping soil classes or fuzzy memberships are needed.

Although probability sampling is not required when the soil is mapped with a statistical model of the spatial variation, probability sampling still can be attractive for various reasons. When we have a dual aim, both estimating the population mean and mapping, it can be attractive to select a probability sample so that the population mean can be estimated model-free, by design-based inference. In this context the work of Grafström and Tillé (2013) and de Gruijter et al. (2016) is of interest. Grafström and Tillé (2013) adapted a sampling algorithm for balanced sampling, which is an efficient sampling design for estimating a population mean that exploits auxiliary variables, so that the sampling units are well spread throughout the study area, see (Brus, 2015) for a detailed description of the algorithm. The geographical spreading may increase the precision of the estimated population mean

(less redundant information), and besides we may profit from this spreading when the balanced sample is used for mapping, for instance by KED. De Gruijter et al. (2016) proposed a method in which a map of carbon content with associated uncertainty is used to optimize stratified random sampling for soil carbon auditing at the farm-scale. Once the data are collected, these data can be used to update the map and the stratification. The updated stratification is then used to select new sampling locations. In this way a series of samples is obtained that is used both for design-based estimation of the population total and for mapping the soil C content. In both sampling designs the primary aim seems to be design-based estimation of the population mean or total. Studies into probability sampling designs optimized for a criterion that is a function of the qualities of both the design-based estimate and of the map are recommended.

Supplement

R scripts, data sets and supplementary figures are available at `https://github.com/DickBrus/TutorialSampling4DSM`.

Acknowledgements

I thank Akmal Akramhanov (International Center for Agricultural Research in the Dry Areas (ICARDA), Tasjkent, Uzbekistan), A-Xing Zhu (Nanjing Normal University, China), Budiman Minasny (University of Sydney, Australia) and Hailu (Agricultural Transformation Agency (ATA), Addis Abeba, Ethiopia) for providing the data used in this paper to illustrate the sampling methods.

References

Aarts, E. and Korst, J. (1987). *Simulated Annealing and Boltzmann Machines*. Wiley.

931 Akramkhanov, A., Brus, D. J., and Walvoort, D. J. J. (2013). Geostatistical moni-
932 toring of soil salinity in uzbekistan by repeated emi surveys. *Geoderma*.

933 Baillargeon, S. and Rivest, L.-P. (2014). *stratification: Univariate Stratification of*
934 *Survey Populations*. R package version 2.2-5.

935 Bogaert, P. and Russo, D. (1999). Optimal spatial sampling design for the estimation
936 of the variogram based on a least squares approach. *Water Resources Research*,
937 35:1275–1289.

938 Brus, D. J. (2015). Balanced sampling: A versatile sampling approach for statistical
939 soil surveys. *Geoderma*, 253-254:111–121.

940 Brus, D. J. and de Gruijter, J. J. (1994). Estimation of nonergodic variograms
941 and their sampling variance by design-based sampling strategies. *Mathematical*
942 *Geology*, 26:437–454.

943 Brus, D. J. and de Gruijter, J. J. (1997). Random sampling or geostatistical mod-
944 elling? Choosing between design-based and model-based sampling strategies for
945 soil (with Discussion). *Geoderma*, 80:1–59.

946 Brus, D. J., de Gruijter, J. J., and van Groenigen, J. W. (2007). Designing spa-
947 tial coverage samples using the k-means clustering algorithm. In Lagacherie, P.,
948 McBratney, A., and Voltz, M., editors, *Digital Soil Mapping. An introductory per-*
949 *spective*, pages 183–192. Elsevier.

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

952 Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation
953 of digital soil maps. *European Journal of Soil Science*, 62(3):394–407.

Burgess, T. and McBratney, R. W. A. (1981). Optimal interpolation and isarithmic mapping of soil properties .4. sampling strategy. *JOURNAL OF SOIL SCIENCE*, 32(4):643–659.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley, New York.

Corwin, D., Lesch, S. M., Segal, E., Skaggs, T. H., and Bradford, S. A. (2010). Comparison of sampling strategies for characterizing spatial variability with apparent soil electrical conductivity directed soil sampling. 15:147–162.

Corwin, D. L. and Lesch, S. M. (2005). Characterizing soil spatial variability with apparent soil electrical conductivity: Part ii. case study. *Computers and Electronics in Agriculture*, 46(1-3 SPEC. ISS.):135–152. Cited By (since 1996): 7.

de Gruijter, J., McBratney, A., Minasny, B., Wheeler, I., Malone, B., and Stockmann, U. (2016). Farm-scale soil carbon auditing. *Geoderma*, 265:120–130.

de Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., and Knotters, M. (2006). *Sampling for Natural Resource Monitoring*. Springer, Berlin.

de Gruijter, J. J., McBratney, A. B., and Taylor, J. (2010). *Sampling for High Resolution Soil Mapping*, chapter 1, pages 3–14. Springer Netherlands, Sydney, Australia.

de Gruijter, J. J. and ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22:407–415.

Debaene, G., Niedzwiecki, J., Pecio, A., and Zurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma*, 214-215:114–125.

977 Domburg, P., de Gruijter, J. J., and Brus, D. J. (1994). A structured approach to
978 designing soil survey schemes with prediction of sampling error from variograms.
979 *Geoderma*, 62:151–164.

980 Fitzgerald, G. (2010). *Response Surface Sampling of Remotely Sensed Imagery for*
981 *Precision Agriculture*, chapter 10, pages 121–129. Springer Netherlands, Sydney,
982 Australia.

983 Fitzgerald, G. J., Lesch, S. M., Barnes, E. M., and Luckett, W. E. (2006). Directed
984 sampling using remote sensing with a response surface sampling design for site-
985 specific agriculture. *Computers and Electronics in Agriculture*, 53(2):98–112.

986 Gao, B., Pan, Y., Chen, Z., Wu, F., Ren, X., and Hu, M. (2016). A spatial condi-
987 tioned latin hypercube sampling method for mapping using ancillary data. *Trans-*
988 *actions in GIS*, 20(5):735–754.

989 Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading
990 and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

991 Heuvelink, G., Brus, D. J., and de Gruijter, J. J. (2007). Optimization of sample
992 configurations for digital mapping of soil properties. In Lagacherie, P., McBratney,
993 A., and Voltz, M., editors, *Digital Soil Mapping. An introductory perspective*, pages
994 1020–1030. Elsevier.

995 Kennard, R. and Stone, L. (1969). Computer aided design of experiments. *Techno-*
996 *metrics*, 11(1):137–148.

997 Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated an-
998 nealing. *Science*, 220(4598):671–680.

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, 51:717–728.

Lark, R. M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105:49–80.

Lark, R. M. (2011). Spatially nested sampling schemes for spatial variance components: Scope for their optimization. *Computers & Geosciences*, 37(10):1633–1641.

Lark, R. M., Hamilton, E. M., Kaninga, B., Maseka, K. K., Mutondo, M., Sakala, G. M., and Watts, M. J. (2017). Planning spatial sampling of the soil from an uncertain reconnaissance variogram. *SOIL*.

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

Lesch, S., Rhoades, J., and Corwin, D. (2000). *ESAP-95 Version 2.01R User Manual and Tutorial Guide*.

Lesch, S. M. (2005). Sensor-directed response surface sampling designs for characterizing spatial variation in soil properties. *Computers and Electronics in Agriculture*, 46(1-3 SPEC. ISS.):153–179.

Lesch, S. M., Strauss, D. J., and 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 Resources Research*, 31:387–398.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, USA.

Marchant, B. P. and Lark, R. M. (2007). Optimized sample schemes for geostatistical surveys. *Mathematical Geology*, 39(1):113–134.

McBratney, A., Webster, R., and Burgess, T. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables .1. theory and method. *Computers & Geosciences*, 7(4):331–334.

McBratney, A. B. and Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: II Program and examples. *Computers & Geosciences*, 7:335–365.

McKay, M., Beckman, R., and Conover, W. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

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

Mulder, V., de Bruin, S., and Schaepman, M. (2013). Representing major soil variability at regional scale by constrained latin hypercube sampling of remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 21:301 – 310.

Müller, W. G. and Zimmerman, D. L. (1999). Optimal designs for variogram estimation. *Environmetrics*, 10:23–27.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2002). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Third edition.* Wiley, New York.

Naes, T. (1987). The design of calibration in near infra-red reflectance analysis by clustering. *Journal of Chemometrics*, 1:121–134.

Oliver, M. and Webster, R. (1986). Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. 18:227–242.

Papritz, A., Duemig, A., Zimmermann, C., Gerke, H. H., Felderer, B., Koegel-Knabner, I., Schaaf, W., and Schulin, R. (2011). Uncertainty of variance component estimates in nested sampling: a case study on the field-scale spatial variability of a restored soil. *European Journal of Soil Science*, 62(3):479–495.

Papritz, A. and Webster, R. (1995). Estimating temporal change in soil monitoring: I. statistical theory. *European Journal of Soil Science*, 46:1–12.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2).

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Dematte, J., and Scholten, T. (2014). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, 226:140–150.

Riedel, F., Denk, M., Mller, I., Barth, N., and Gler, C. (2018). Prediction of soil parameters using the spectral range between 350 and 15,000 nm: A case study based on the permanent soil monitoring program in saxony, germany. *Geoderma*, 315:188–198. cited By 1.

Roudier, P. (2011). *clhs: a R package for conditioned Latin hypercube sampling.*

Roudier, P., Beaudette, D., and Hewitt, A. (2012). *A conditioned Latin hypercube sampling algorithm incorporating operational constraints*, pages 227–231.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, 24:479–488.

Samuel-Rosa, A. (2016). *spsann: Optimization of Sample Configurations using Spatial Simulated Annealing*. R package version 2.0-0.

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

Sila, A., Hengl, T., and Terhoeven-Urselmans, T. (2014). *soil.spec: Soil Spectroscopy Tools and Reference Models*. R package version 2.1.4.

ter Braak, C. J. F. and Vrugt, J. A. (2008). Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18:435–446.

Tillé, Y. and Matei, A. (2015). *Survey sampling*. R package version 2.7.

Totaro, S., Coratza, P., Durante, C., Foca, G., Li Vigni, M., Marchetti, A., Marchetti, M., and Cocchi, M. (2013). Soil sampling planning in traceability studies by means of experimental design approaches. *Chemometrics and Intelligent Laboratory Systems*, 124:14–20. cited By 6.

van Groenigen, J. W., Pieters, G., and Stein, A. (2000). Optimizing spatial sampling for multivariate contamination in urban areas. *Environmetrics*, 11:227–244.

van Groenigen, J. W., Siderius, W., and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, 87:239–259.

van Groenigen, J. W. and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, 27:1078–1086.

Vășát, R., Heuvelink, G., and Boruvka, L. (2010). Sampling design optimization for multivariate soil mapping. 155:47–153.

Viscarra Rossel, R. A. and Brus, D. J. (2018). The cost-efficiency and reliability of two methods for soil organic c accounting. *Land Degradation and Development*, 29(3):506–520.

Walvoort, D., Brus, D., and de Gruijter, J. (2010a). *Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata*. R package version 0.2-3.

Walvoort, D. J. J., Brus, D. J., and de Gruijter, J. J. (2010b). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*, 36:1261–1267.

Webster, R. and Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43:177–192.

Webster, R., Welham, S. J., Potts, J. M., and Oliver, M. A. (2006). Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. 32:1320–1333.

Yang, L., Brus, D. J., A-X Zhu, Li, X., and Shi, J. (2018). Accounting for access costs in validation of soil maps: a comparison of design-based sampling strategies. *Geoderma*, 315:160–169.

1113 Zhu, Z. and Stein, M. L. (2005). Spatial sampling design for parameter estimation
1114 of the covariance function. *Journal of Statistical Planning and Inference*, 134:583–
1115 603.

1116 Zhu, Z. and Stein, M. L. (2006). Spatial sampling design for prediction with esti-
1117 mated parameters. *Journal of Agricultural Biological and Environmental Statis-*
1118 *tics*, 11:24–44.

1119 Zhu, Z. and Zhang, H. (2006). Spatial sampling under the infill asymptotic frame-
1120 work. *Environmetrics*, 17:323–337.

1121 Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance
1122 parameter estimation, and empirical prediction. *Environmetrics*, 17:635–652.

Table 1: Standard errors of the estimated intercept (β_0) and slope (β_1) for a simple random sample (SRS) and the sample optimized for simple linear regression, see Figure 1

	β_0	β_1
SRS	1.51	0.086
Optimized sample	1.08	0.051

Table 2: Estimated population mean error and population mean squared error, with standard errors in paranthesis; t: outcome of test statistic of hypothesis $ME = 0$, against two-sided alternative hypothesis, with p-value in paranthesis

	RF	KED
$\widehat{ME}$	0.546 (1.306)	0.814 (1.203)
$\widehat{MSE}$	95.9 (26.3)	89.4 (25.5)
t	0.418 (0.678)	0.676 (0.502)

Table 3: Overview of mapping methods and sampling designs; OK: ordinary kriging; LR: linear regression; KED: kriging with an external drift; ML: machine learning techniques; cLHS: conditioned Latin hypercube sample

Covariate maps available?	Mapping method	Sampling design	Remark
No	OK	Regular grid	Option: optimized grid spacing
		Spatial coverage/infill sample	
		Model-based sample	Min. crit: mean or max OK-var
Yes	LR	Response surface sample	
	KED	Regular grid	Option: optimized grid spacing
		Spatial coverage/infill sample	
		Model-based sample	Min. crit: mean or max KED-var
	ML	k-means sample	
		cond. Latin hypercube sample	
		Kennard-Stone sample	

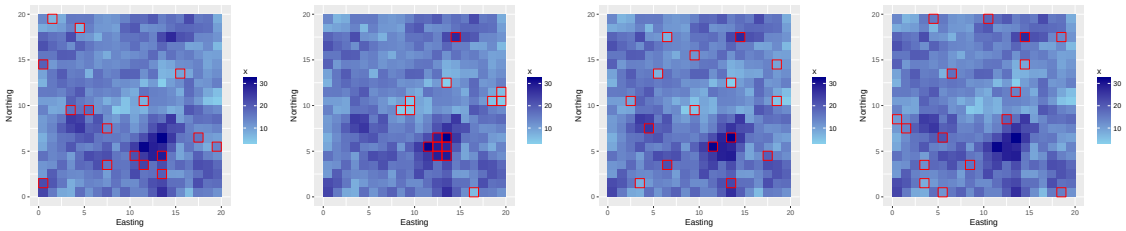

Figure 1: From left to right: simple random sample, optimized sample for mapping with simple linear regression model, optimized sample for kriging with an external drift, and stratified sample using sixteen equal-sized covariate strata (quantiles of covariate used as stratum boundaries). All samples are plotted in a map of the covariate.

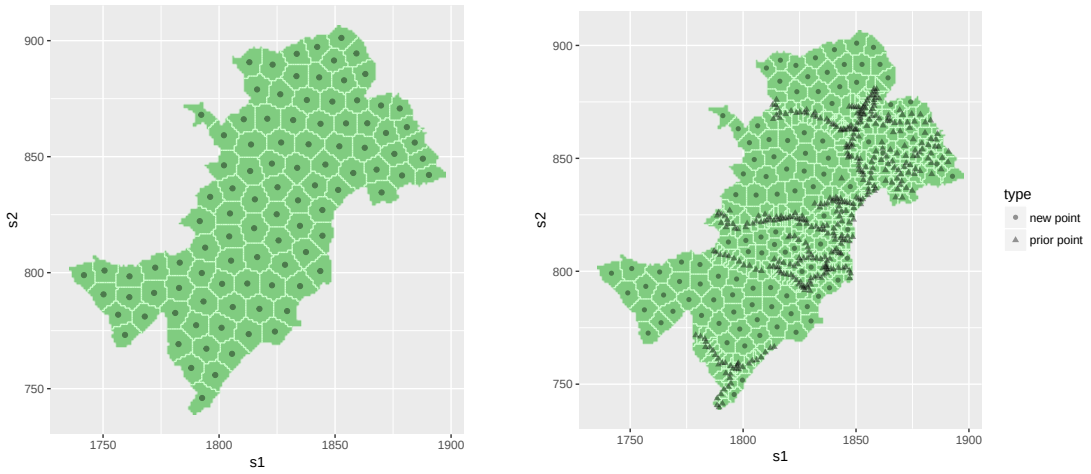

Figure 2: Spatial coverage and spatial infill sample in three woredas of Ethiopia, optimized by minimizing MSSD by k-means.

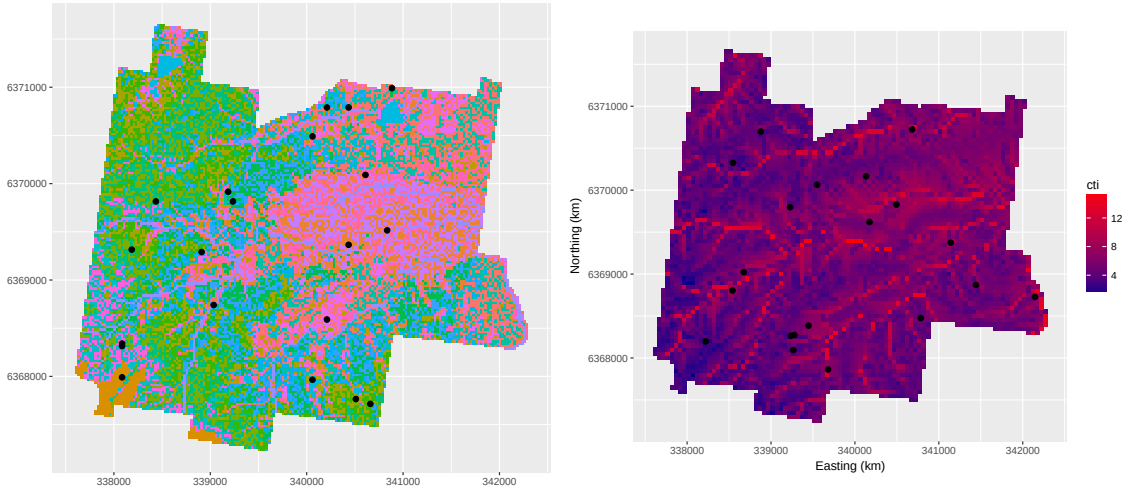

Figure 3: Hard k-means (left) and cLHS sample (right) of 20 points in Hunter Valley, using elevation, slope, aspect, cti, and ndvi as covariates.

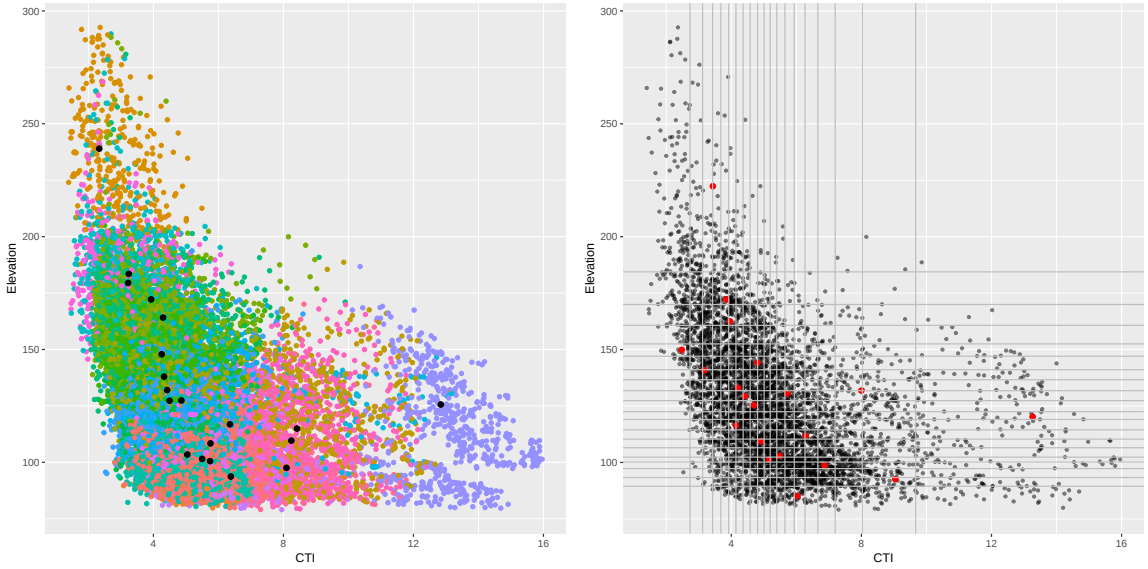

Figure 4: Hard k-means (left) and cLHS sample (right) plotted in scatter diagram of elevation against compound topographic index. Vertical and horizontal lines in scatter diagram of cLHS are at breaks of marginal strata.

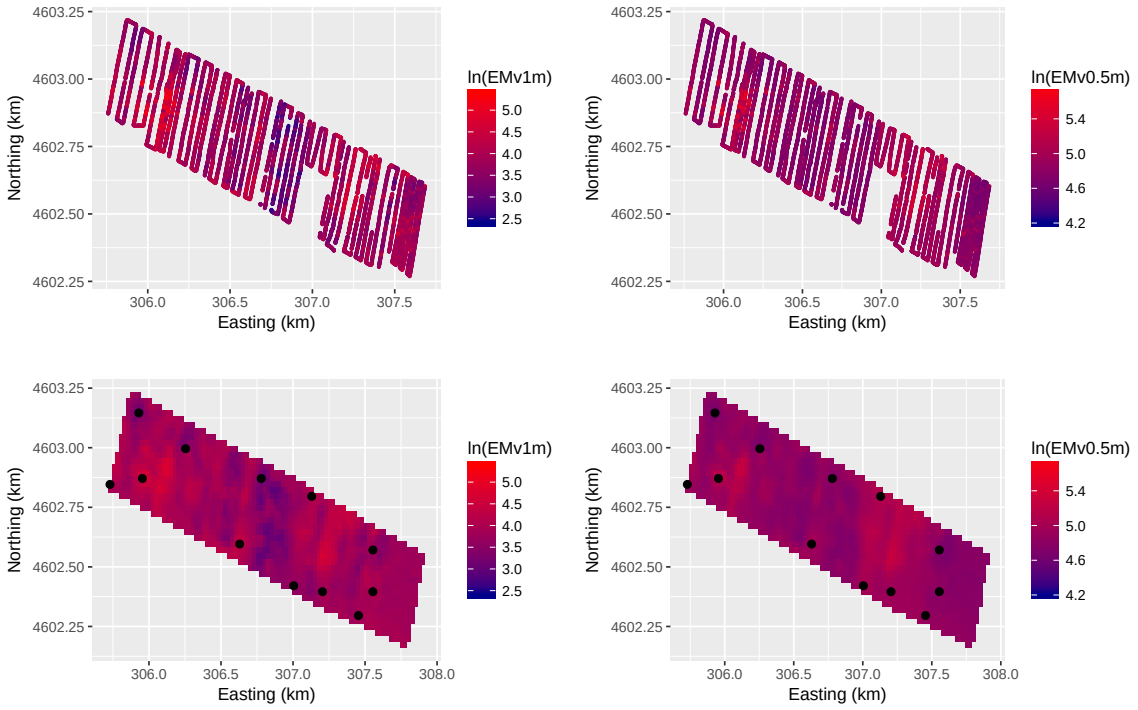

Figure 5: Natural log transformed measurements of EMv-1m and EMv-0.5m in Cotton Research Field, Uzbekistan (top), and response surface sample plotted on ordinary kriging predictions of ln(EMv-1m) and ln(EM-0.5m) (bottom)

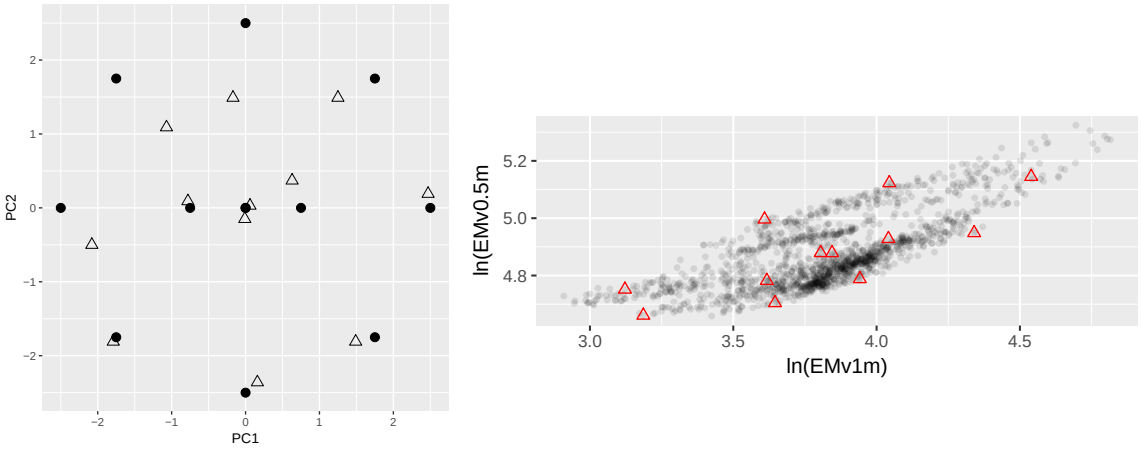

Figure 6: Design points (dots) and principal component scores (triangles) of selected response surface sample (left), and response surface sample plotted in the scatter diagram of the two covariates (right)

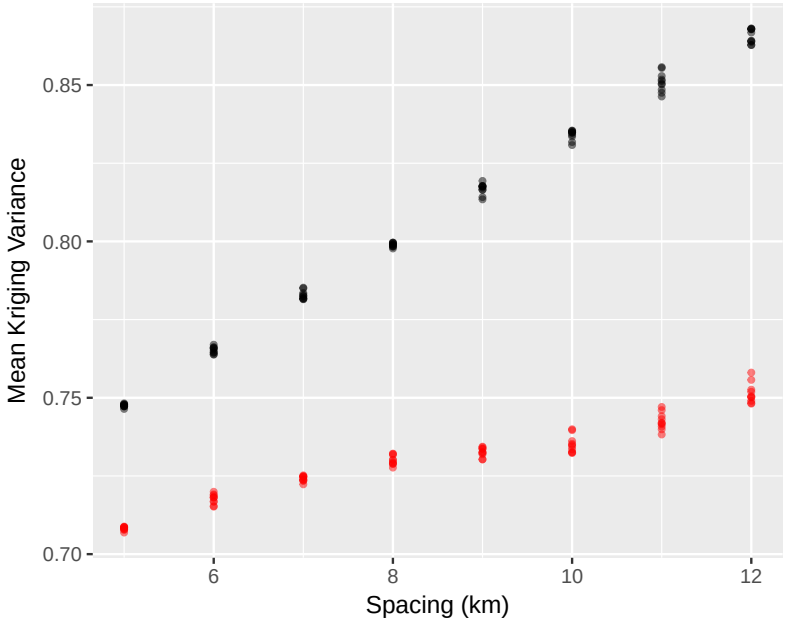

Figure 7: Mean ordinary kriging variance (black dots) and mean variance of kriging with an external drift (red dots) for square grids of variable spacing, selected from the three woredas in Ethiopia. For each spacing ten grids are selected with a random start.

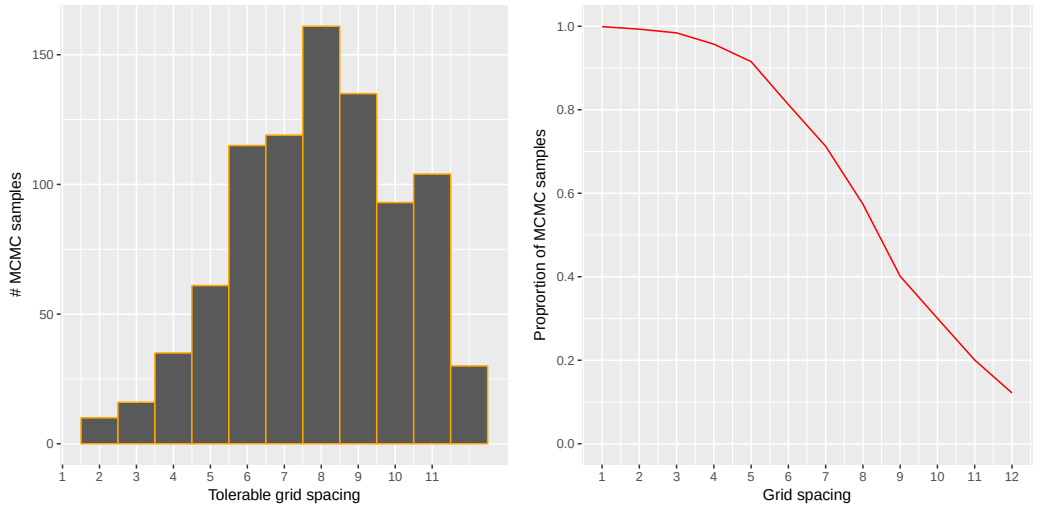

Figure 8: Histogram of tolerable grid spacing for a target MKV of 0.8 (left) and proportion of MCMC samples with a MKV smaller than or equal to a target MKV of 0.8

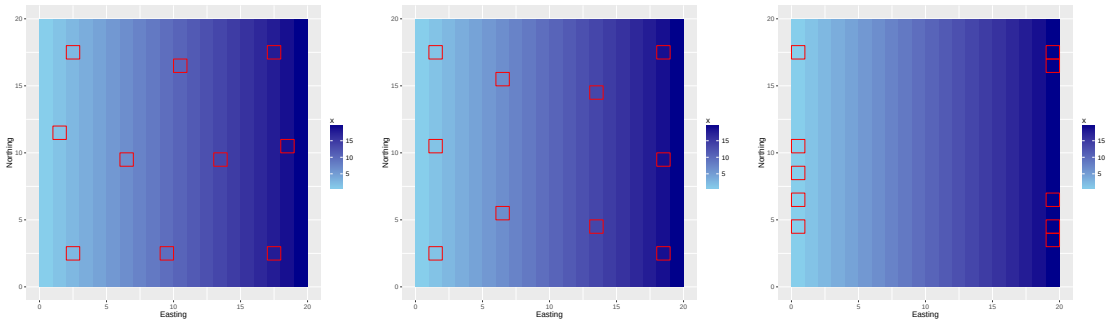

Figure 9: Samples optimized by SSA for KED, using Easting as a covariate, for three exponential residual variograms with a distance parameter of five distance units. The nugget and partial sill parameters of the residual variogram are 0 and 1 (left), 0.5 and 0.5 (middle), and 1 and 0 (right), respectively.

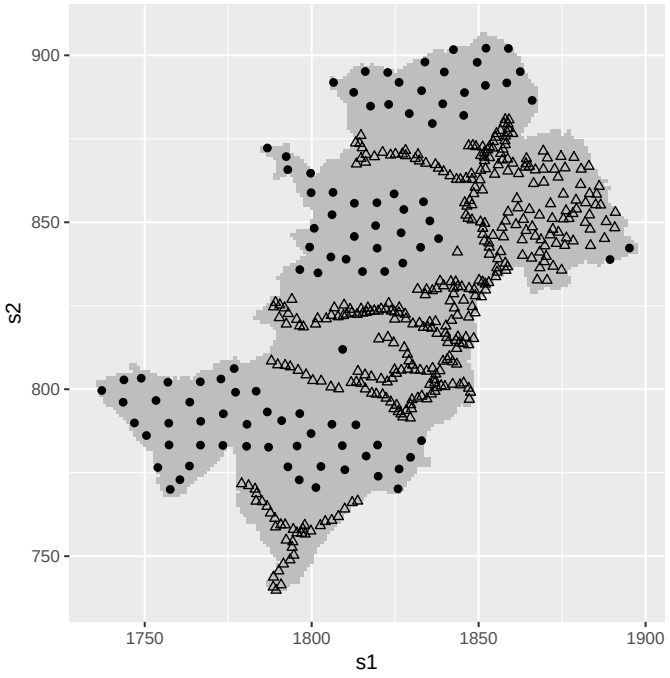

Figure 10: Infill sample optimized by SSA, for ordinary kriging of SOM in three woredas of Ethiopia. The variogram for SOM estimated from the legacy sample data (triangles) was spherical, with nugget 0.62, partial sill 0.56, and range 45 km.

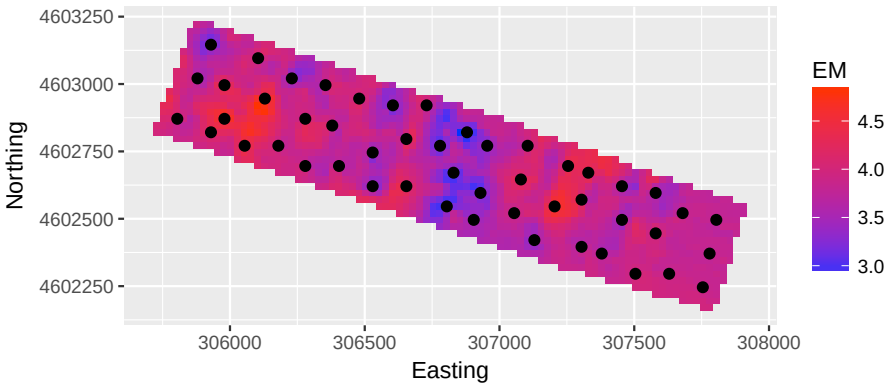

Figure 11: Sample optimized by SSA, for KED of ECe in the Cotton Research Field (Ethiopia), using interpolated values of natural log of EMv1m as covariate.

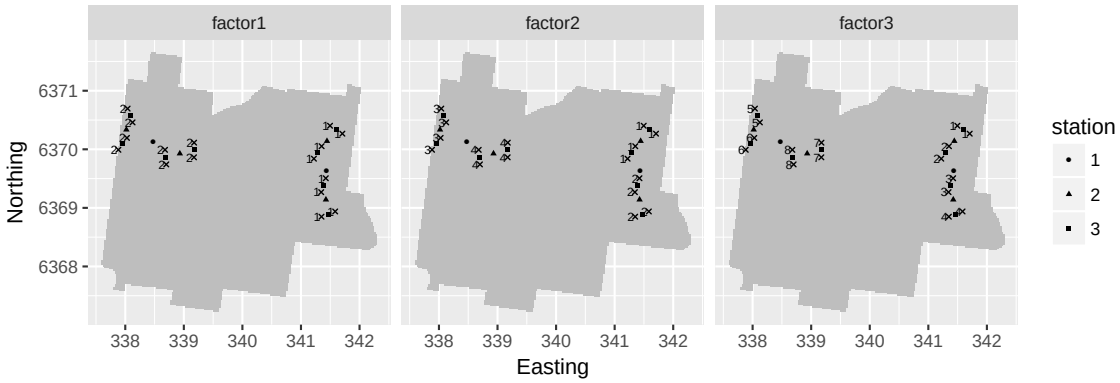

Figure 12: Balanced nested sample from Hunter Valley. In the first stage two stations are selected with a separation distance of 3000 m (first order stations). In the second stage each first order station is used to select a pair of second order stations with the second largest separation distance. In the third stage each second order station is used to select a pair of third order stations. Finally, in the fourth stage each third order station is used to select a pair of points included in the nested sample (symbol x). This results in a balanced nested sample of 16 points. The three panels show the factor levels needed for the hierarchical ANOVA.

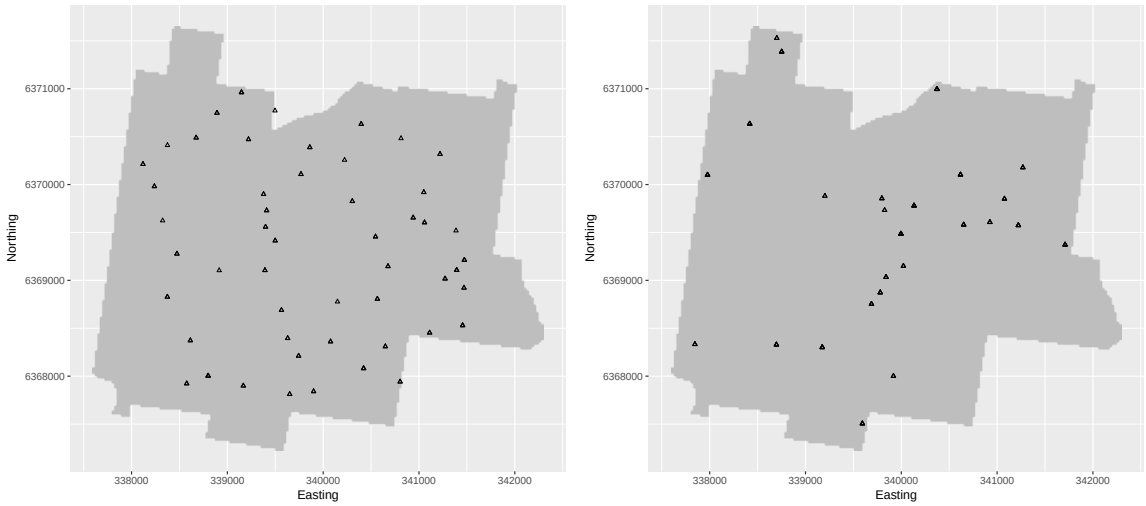

Figure 13: Samples of 100 points optimized for variogram estimation. Minimization criterion: average of logdet(F^{-1}) (left) and $V(\sigma_K^2)$ (right). Postulated variogram: exponential with distance parameter of 500 m and nugget to sill ratio of 0.2.

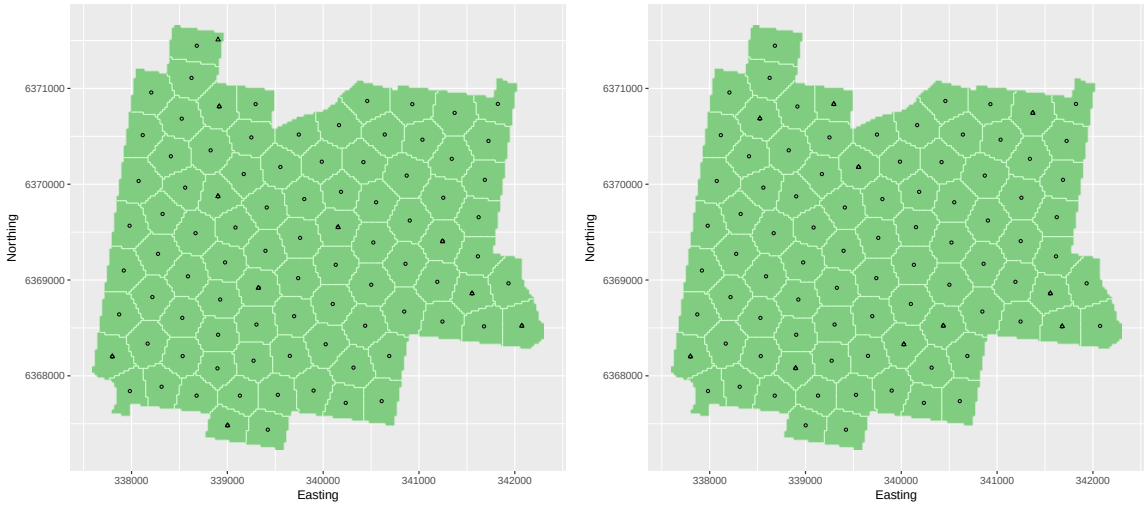

Figure 14: Spatial coverage sample of 90 points (circles), supplemented by 10 points (triangles) optimized by SSA. Minimization criterion in SSA: average of augmented kriging variance (left) and EAC (right). Postulated variogram: exponential with distance parameter of 500 m and nugget to sill ratio of 0.2.

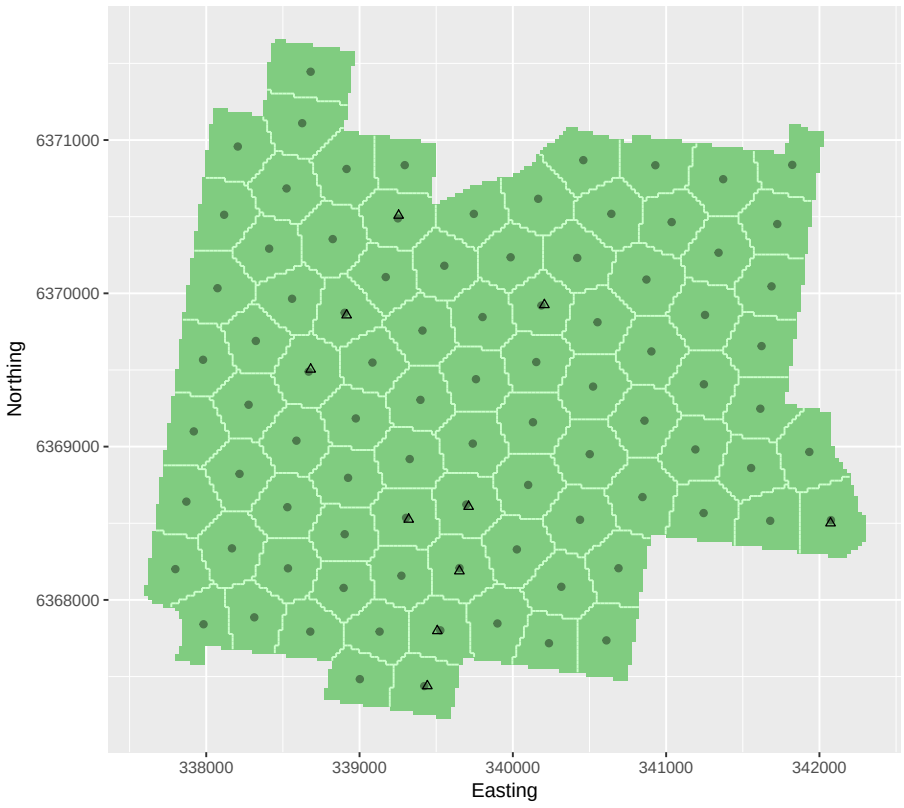

Figure 15: Spatial coverage sample of 100 points supplemented with 10 points at a short distance of randomly selected point of spatial coverage sample.

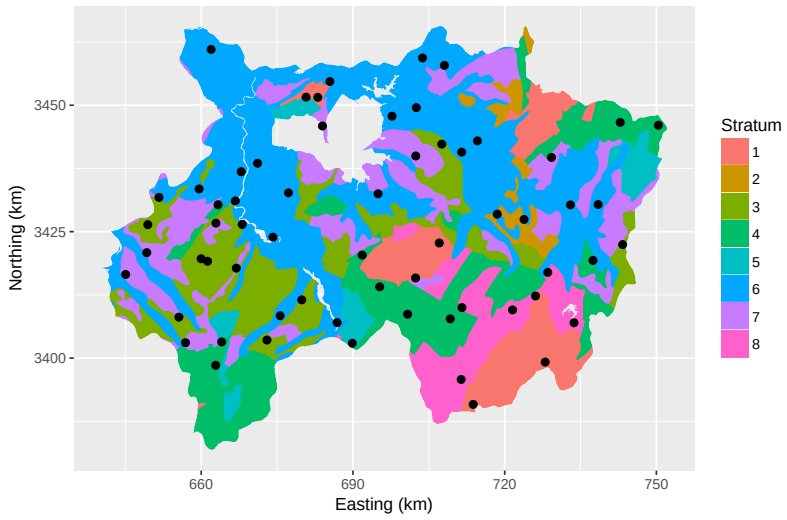

Figure 16: Stratified simple random sample of 62 points for validation of two maps of soil organic matter concentration in A horizon in Xuancheng. Strata are the eight units of a geological map.