Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping

Felix Stumpf[1]*, Karsten Schmidt[1], Thorsten Behrens[1], Sarah Schönbrodt-Stitt[1], Giovanni Buzzo[2], Christian Dumperth[3], Alexandre Wadoux[4], Wei Xiang[5], and **Thomas Scholten[1]**

[1] University of Tübingen, Department of Geosciences, Chair of Soil Science and Geomorphology, Germany
[2] University of Trier, Department of Environmental Remote Sensing and Geoinformatics, Chair of Spatial and Environmental Planning, Germany
[3] University of Erlangen-Nuremberg, Department of Geology and Mineralogy, Chair of Applied Geology, Germany
[4] Wageningen UR, Department of Soil Geography and Landscape, Wageningen, Netherlands
[5] China University of Geosciences Wuhan, Department of Geotechnical Engineering and Engineering Technology, China

Abstract

Most calibration sampling designs for Digital Soil Mapping (DSM) demarcate spatially distinct sample sites. In practical applications major challenges are often limited field accessibility and the question on how to integrate legacy soil samples to cope with usually scarce resources for field sampling and laboratory analysis. The study focuses on the development and application of an efficiency improved DSM sampling design that (1) applies an optimized sample set size, (2) compensates for limited field accessibility, and (3) enables the integration of legacy soil samples. The proposed sampling design represents a modification of conditioned Latin Hypercube Sampling (cLHS), which originally returns distinct sample sites to optimally cover a soil related covariate space and to preserve the correlation of the covariates in the sample set. The sample set size was determined by comparing multiple sample set sizes of original cLHS sets according to their representation of the covariate space. Limited field accessibility and the integration of legacy samples were incorporated by providing alternative sample sites to replace the original cLHS sites. We applied the modified cLHS design ($\text{cLHS}_{\text{adapt}}$) in a small catchment (4.2 km^2) in Central China to model topsoil sand fractions using Random Forest regression (RF). For evaluating the proposed approach, we compared $\text{cLHS}_{\text{adapt}}$ with the original cLHS design ($\text{cLHS}_{\text{orig}}$). With an optimized sample set size n = 30, the results show a similar representation of the cLHS covariate space between $\text{cLHS}_{\text{adapt}}$ and $\text{cLHS}_{\text{orig}}$, while the correlation between the covariates is preserved (r = 0.40 *vs.* r = 0.39). Furthermore, we doubled the sample set size of $\text{cLHS}_{\text{adapt}}$ by adding available legacy samples ($\text{cLHS}_{\text{adapt+}}$) and compared the prediction accuracies. Based on an external validation set cLHS_{val} (n = 20), the coefficient of determination (R^2) of the $\text{cLHS}_{\text{adapt}}$ predictions range between 0.59 and 0.71 for topsoil sand fractions. The R^2-values of the RF predictions based on $\text{cLHS}_{\text{adapt+}}$, using additional legacy samples, are marginally increased on average by 5%.

Key words: digital soil mapping / field accessibility / legacy soil samples / sample set size / conditioned Latin Hypercube Sampling / random forest / Three Gorges Reservoir Area

Accepted May 03, 2016

1 Introduction

Digital Soil Mapping (DSM) couples soil information obtained at distinct spatial locations with statistically related, co-located, and area-covering predictor covariates. The coupling is accomplished by regression and classification approaches resulting in continuous or discrete maps of soil properties (*McBratney* et al., 2003; *Scull* et al., 2003; *McMillan*, 2008). DSM presents an established framework for soil mapping and has been successfully applied in numerous studies, addressing various soil properties, landscapes, and scales (*Florinsky* et al., 2002; *Behrens* et al., 2005; *Mora-Vallejo* et al., 2008; *Lacoste* et al., 2011; *Wang* et al., 2012; *Behrens* et al., 2014; *de Carvalho Junior* et al., 2014; *Mansuy* et al., 2014; *Taghizadeh-Mehrjardi* et al., 2014; *Thomas* et al., 2015).

DSM predictor covariates can be derived inexpensively from existing data sets such as digital elevation models (DEM) and remote sensing data (*McKenzie* and *Ryan*, 1999; *Gessler* et al., 2000; *Behrens* et al., 2010). By contrast, the field sampling of soil data remains a limiting factor. This is attributed to the requirements of statistical sampling designs, which need to be suited to local environmental conditions and pursue the incorporation of real field costs and budgetary constraints at the same time (*Lagacherie*, 2008; *Kidd* et al., 2015).

* Correspondence: Dr. F. Stumpf; e-mail: felix.stumpf@uni-tuebingen.de

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Primarily, the sampling design should reflect the variation of the target soil property in the study area (*Heuvelink* et al., 2007; *Brungard* and *Boettinger*, 2010). Suggested strategies infer sampling in the geographical space (*Brus* et al., 2006), in the soil related covariate space (*Minasny* and *McBratney*, 2006), or in a combination of both (*Dobermann* and *Simbahan*, 2007). Secondly, the sampling design should support field operability in terms of constrained accessibility, *e.g*., due to difficult terrain and restricted areas (*Kidd* et al., 2015). Few studies addressed this issue by excluding inaccessible areas in the process of sample site selection (*e.g*., *Roudier* et al., 2012; *Mulder* et al., 2013; *Clifford* et al., 2014) or by applying models from accessible areas to inaccessible areas based on similar environmental conditions (*Cambule* et al., 2013). Third, the sampling design should incorporate available legacy soil information to accommodate the demand on reducing high labor and monetary costs for sampling and laboratory analysis (*Lagacherie*, 2008). Existing soil maps served as covariates (*Mayr* and *Palmer*, 2007) and in disaggregated form as a source to calibrate prediction models (*Naumann* and *Thompson*, 2014). Legacy soil profiles provide local information on soil properties and served as input for statistical model procedures (*Carré* and *Girard*, 2002; *Hengl* et al., 2004). Yet, a spatial mismatch of statistically predefined sample sites, a lack in harmonization with the target soil property, and different spatial resolutions, formats and objectives remain problems when incorporating legacy data into sampling designs (*Carré* et al., 2007; *Krol*, 2008; *Sulaeman* et al., 2013). A further possibility to increase the efficiency in soil data acquisition comprises an optimized sample set size. Few studies addressed this issue by comparing model results based on different calibration set sizes (*Brungard* and *Boettinger*, 2010; *Ramirez-Lopez* et al., 2014; *Schmidt* et al., 2014).

As a consequence of the described restriction and limitations, most sampling designs solely focus on reflecting the variation of the target soil property. They often do not consider operability and efficiency improvements in terms of accessibility, the integrative use of legacy samples and optimization of the sample set size (*Lagacherie*, 2008; *Cambule* et al., 2013). Thus, advances in surveying soil data for DSM depend on comprehensively addressing the statistical, operational, and efficiency potentials of sampling designs. Such comprehensive DSM sampling designs have been addressed by few studies (*e.g*., *Roudier* et al., 2012; *Mulder* et al., 2013; *Clifford* et al., 2014) that are based on conditioned Latin Hypercube Sampling (cLHS). The cLHS method presents a stratified random sampling design that provides an optimal stratification of a covariate space with a reduced number of spatially distinct sample sites (*Minasny* and *McBratney*, 2006). *Roudier* et al. (2012) and *Mulder* et al. (2013) implemented data on travelling costs, which were derived from terrain and land use parameters, into the cLHS algorithm to account for limited field accessibility. Similar to these approaches, thus, aiming to preserve an optimal stratification of the samples in the covariate space while compensating limited field accessibility, *Clifford* et al. (2014) proposed 'flexible LHS'. The method additionally produces an ordered list of alternative sample sites by analyzing the relation of each location in a defined neighborhood to the initial sample site, in case the latter turns out to be inaccessible during field sampling.

The above approaches outline improvements concerning the applicability of hypercube sampling designs, mainly by addressing limited field accessibility. However, further improvements should combine these advances with an optimized sample set size and the integrative use of legacy soil samples. Thus, the objective of our study is to develop a comprehensive and efficiency improved DSM sampling design that (1) applies an optimal sample set size, (2) compensates for limited field accessibility, and (3) enables the integration of legacy soil samples.

The study is embedded in the joint Sino-German project 'YANGTZE GEO—Land use change, soil erosion, mass movements and matter fluxes along the Yangtze river, Three Gorges Reservoir Area' (*Schönbrodt-Stitt* et al., 2013; *Strehmel* et al., 2015). Therefore, the study is situated in a small catchment in the Three Gorges Reservoir Area and targets to provide a methodological frame to produce sensitive data of topsoil sand fractions for process-based erosion modeling.

2 Material and methods

2.1 Methodological overview

The proposed sampling design outlines a modification of the standard cLHS procedure (*Minasny* and *McBratney*, 2006; *Schmidt* et al., 2014). The idea of cLHS is that the covariate space of the entire study area is represented by a distinct sample set. Therefore, cLHS divides the range of each covariate into a number of equally probable strata, which corresponds to the sample set size. The sample set is derived by iteratively sampling randomly from the entire covariate space and finally selecting sample sites that in combination represent the entire covariate space by one distinct site within each stratum. This optimization procedure is accomplished by simulated annealing and ensures that each covariate is uniformly sampled in the final sample set (*Metropolis* et al., 1953; *Flannery* et al., 1992). Due to the purely statistical nature of the method, sample sites might be selected that do not exist in the real world. Hence, the site selection is conditioned by rejecting covariate combinations that do not exist in the real world (*Minasny* and *McBratney*, 2006).

We modified the standard cLHS procedure to provide alternative sample sites for the spatially distinct cLHS sites and to integrate legacy soil samples. Primarily, we predefined accessible cLHS sites and locations in a specific stratum with an available legacy sample as fixed sample sites. Secondly, we identified all accessible locations in the remaining strata to serve as a potential alternative for the respective inaccessible cLHS site. Subsequently, we generated test sample sets, from all possible combinations of fixed sample sites and potential alternative sites across the strata. Finally, we identified the test sample set that is most similar to the original cLHS set according to the cLHS criteria, formulated by *Minasny* and *McBratney* (2006). Therefore, the sample set should (1) represent the covariate space and (2) preserve the correlation between the covariates.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

In cLHS, the number of strata and potential alternative sample sites within each stratum is determined by the complexity of the covariate space and the sample set size. Thus, we reduced the number of covariates and optimized the sample set size to ensure that at least one accessible alternative sample site is present in each stratum of the modified cLHS. This was accomplished by (1) a correlation analysis between covariates and legacy soil samples to identify the most relevant covariates and (2) by analyzing the representation of the covariate space by various cLHS sets with different sample set sizes to identify an optimal sample set size.

Finally, we applied the modified cLHS design in the field, conducted a laboratory analysis to obtain the target soil properties, and used the data for calibrating a DSM model using Random Forest regression (RF). For the RF models we applied all available covariates since RF is robust to multi-collinearity.

2.2 Study area and geodatabase

Our study area is a drainage basin of 4.2 km^2, referred to as Upper Badong catchment (31°1′24′′N, 110°20′35′′E). It is located at the middle reaches of the Yangtze River, approximately 74 km upstream the Three Gorges Dam (TGD) in the western Hubei province in Central China (*Strehmel* et al., 2015). With 72% the vast majority of the Upper Badong catchment is exposed to the north. The altitude ranges from 469 m to 1,483 m asl with an average altitude of 1,053 m asl. Slope angles range from 0 to 53° with an average slope angle of 26°. The land use in the study area is dominated by woodland (81%) with scattered plots of cropland (15%) and small farm buildings (4%). The cropland area with mainly soybean, corn, and cabbage as agricultural products is located in the northern, lower part of the Upper Badong catchment. Woodland is predominant in the steep sloping southern catchment area.

Land use information is based on a RapidEye satellite image from September 28th 2012, providing five spectral bands in a spatial resolution of 5 m × 5 m (*RapidEye*, 2012). We derived a set of six land use classes according to ‘Cropland’, ‘Grassland’, ‘Broadleaf’, ‘Conifer’, ‘Shrub’, ‘Woodland’, and ‘Built-up’, adapted from *Liu* et al. (2005). For processing, we applied Support Vector Machine (SVM) classification, producing an overall class agreement of 74% (*Huang* et al., 2004; *Rabe* et al., 2010). Moreover, we generated a digital elevation model (DEM) based on semi-automated digitizing contour lines at 10 m intervals of a topographical map with a spatial resolution of 1:25,000. Subsequently, we interpolated the polylines for every raster cell of an underlying grid with a cell size of 25 m × 25 m. The root mean squared error (RMSE) of the DEM amounts to 8.3 m (*SAGA GIS*, 2011).

Based on the DEM, we derived a pool of continuous terrain parameters. For processing, we applied the default settings of the SAGA GIS-toolboxes ‘Terrain Analysis–Morphometry/Hydrology’ (*SAGA GIS*, 2011). The terrain parameters (Table 1) served as a geodatabase of covariates (1) to select a subset that forms the covariate space for the modified cLHS design, and (2) to be used as predictor covariates for modelling the target soil properties.

In October 2012, 55 topsoil samples were collected in a preliminary reconnaissance survey at 24 sample sites in the land use class ‘Cropland’, 12 in ‘Woodland’, nine in ‘Conifer’, six in ‘Broadleaf’, and four in ‘Shrub’. The samples were obtained from an approximated depth of 25 cm by the process of composite sampling. At each sample site of 40 cm × 40 cm, five subsamples from the center point and four from the surrounding corner points were mixed, reduced to 500 g, placed into labeled plastic bags, and subsequently analyzed in the laboratory. The proportional fractions of coarse sand (CS: 2–0.63 mm), medium sand (MS: 0.63–0.2 mm), and fine sand (FS: 0.2–0.063 mm) were derived according to *DIN ISO 11277:2002-08* (2002). Therefore, the samples were oven-dried (105°C), grinded, and dispersed into primary particles using $Na_4P_2O_7$ in the suspension, and finally sieved (*Müller* et al., 2009). In case the sum of the sand fractions deviated by more than 5% from the ideal of 100%, the analysis was repeated. In the present study, these samples are considered as legacy samples.

2.3 Sampling design

2.3.1 Covariate space and sample set size

We selected a subset of covariates to determine the covariate space for the cLHS design. The selection was based on a reasonable averaged correlation ($r_{\text{average}} > 0.4$) between the covariates (Table 1) and the target soil properties of the legacy samples. Moreover, the selection followed the criterion of a low correlation ($r_{\text{collinearity}} < 0.4$) within the subset of covariates to avoid collinearity (*Gessler* et al., 2000; *Hengl* et al., 2003; *Mulder* et al., 2013). We selected (1) the covariate with the highest r_{average}, and (2) all other covariates with $r_{\text{average}} >$ 0.4 and simultaneously $r_{\text{collinearity}} < 0.4$.

Subsequently, we determined the optimal sample set size n by identifying the best tradeoff between the representation of the cLHS covariate space and sampling effort. Using the statistical variance (*var*), we compared the cLHS covariate space x_i ($i = 1, \ldots, k$) referring to the entire study area (global variance) and 10 cLHS sets (sample set variance) with a size n_j ($j = 10, 20, \ldots, 100$):

$$n(j) = var(x_i) - \mathrm{var}(x_{ij}). \qquad (1).$$

The kneepoints of the curves with the corresponding sample set sizes are assumed to be the best tradeoffs for the specific covariates in the cLHS covariate space (*Schmidt* et al., 2014; *Ramirez-Lopez* et al., 2014). We compared the kneepoints of each cLHS covariate to finally select the maximum corresponding sample set size for the cLHS design.

2.3.2 Modified conditioned Latin Hypercube Sampling

We applied the cLHS covariate space and sample set size as described in the previous section to set up a cLHS design according to *Schmidt* et al. (2014). This cLHS version includes an extension which facilitates to set the extreme values in the covariate space as fixed and thereby ensuring a full representation by the sample set.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1: Environmental covariates with summary statistics.

Covariate	Unit	Minimum	Maximum	Average	Standard deviation
Altitude	m a.s.l.	469	1483	1054	255
Northness	–	1.42E-02	1.75E-02	1.62E-02	1.10E-03
Eastness	–	0	1.30E-02	4.98E-03	3.74E-03
Wetness Index (SWI)	–	0	14.8	5.9	1.8
Slope angle	degree	0	53.2	26.4	6.9
Slope length	m	0	2854	184	293
Catchment area	m^2 (log)	6.43	15.17	8.66	1.40
Plane curvature	m^{-1}	–1.03E-02	1.09E-02	–4.28E-05	2.82E-03
Profile curvature	m^{-1}	–1.09E-02	1.04E-02	–1.90E-04	2.30E-03
Combined curvature	m^{-1}	–8.80E-01	8.10E-01	–4.90E-03	1.54E-01
Flow accumulation	pixels (log)	2.8	6.1	3.9	0.56
Overland flow distance	m	0	377	91.9	75.1
Vertical flow distance	m	0	135	29.8	26.1
Horizontal flow distance	m (log)	0	2.6	1.5	0.88
Altitude above channel (AAC)	m	0	307	92	62
Terrain ruggedness	–	0.18	17.2	8.4	2.4
Mass balance index	–	–0.79	2.04	0.13	0.52
Convergence index	–	0	28.8	8.7	3.8
Position index	m	–26.9	35.7	0.25	7.2
Protection index	–	0	0.14	0.07	0.02

Within the resulting cLHS strata of the original cLHS sample sites (cLHS$_{orig}$), we defined areas with a slope angle larger than 35° and the land use classes 'Broadleaf', 'Conifer', 'Shrub', 'Built-up', as well as 'Water bodies' as inaccessible. The land use class 'Woodland' was also excluded from sampling, unless the location was in distance of less than 150 m from a path and, therefore, accessible in reasonable temporal expense.

A legacy sample, which occupied a stratum with an inaccessible cLHS$_{orig}$ sample site, substituted the latter and was defined as fixed sample site for the final sampling design. Accessible cLHS$_{orig}$ sample sites in the remaining strata were also fixed. For strata with neither an accessible cLHS$_{orig}$ sample site nor a matching legacy sample, an alternative sample site was selected from the accessible area within the respective stratum.

The alternative sample sites were selected by setting up test sample sets based on the fixed sample sites with all possible combinations of accessible alternative sites across the strata. We compared the test sample sets to the cLHS$_{orig}$ sample set according to (1) the preservation of the correlation *r* between the covariates and (2) the representation of the cLHS covariate space. For the latter, we used the frequency distribution of samples across the quartiles (Q) as simple measures. Ideally, 25% of the samples would fall in the first and third quartile (Q1 and Q3) and 50% of the samples would fall in the second quartile (Q2). The test sample set with the smallest deviation from the ideal distribution, while preserving the correlation between the covariates, was selected as final sample set cLHS$_{adapt}$. The additional samples were obtained in April 2013 according to the same procedure as for the legacy samples.

2.4 Spatial predictions

We applied RF to set up a DSM model. RF is an ensemble classifier based on multiple randomized decision trees, which use a set of binary rules to compute a target variable (*Breiman*, 2001). In context of DSM, the binary rules are based on multiple environmental covariates and the target soil property. The final prediction is computed by averaging the results over all individual trees for each location of the map. The single trees of a RF model should be as diverse as possible (*Grimm* et al., 2008). This is accomplished by (1) setting up each tree based on a bootstrap sample of the respective soil sample set, and by (2) identifying the best split predictor covariate at each tree node from a random subset (*Peters* et al., 2007).

We applied RF using the R-package 'randomForest' by *Liaw* and *Wiener* (2002). The number of trees (*k*) and the size of the random subset at each node (m_{try}) were defined with k = 1,500 and $m_{try} = 2\sqrt{p}$, with *p* representing the total number of predictor covariates (*Breiman*, 2001). For model calibration we used cLHS$_{adapt}$ and all available covariates (Table 1).

2.5 Evaluation

Evaluation of the DSM approach is based on a three-fold comparative analysis, which addresses (1) the quality of the proposed sampling design cLHS$_{adapt}$, (2) the predictive performance of the model results, and (3) the spatial uncertainty of the model approach.

Primarily, we used cLHS$_{orig}$ as baseline to compare sample sites of cLHS$_{adapt}$ with a further extended calibration set cLHS$_{adapt+}$, which comprises cLHS$_{adapt}$ and unused legacy samples. The comparison is based on the cLHS criteria of representing the cLHS covariate space and preserving the covariate correlation in the sample set.

Secondly, we compared the predictive performance of the DSM approaches referring to all target soil properties and both calibration sets cLHS$_{adapt}$ and cLHS$_{adapt+}$, which are equal in terms of model specifications and the RF covariate space. The comparison is based on external and bootstrap validation, using the coefficient of determination (R^2) and root mean squared error (RMSE) as accuracy estimators. A set of 20 external validation samples were previously derived by random data-splitting from the pool of unused legacy samples.

Furthermore, the target soil properties represent compositional data, while their sum ideally amounts to 100%. Thus, we used the coherence of the predicted sums for each pixel as spatial uncertainty measure to compare uncertainty patterns of both DSM approaches referring to cLHS$_{adapt}$ and cLHS$_{adapt+}$.

3 Results

3.1 Sampling design

3.1.1 Covariate space and sample set size

The covariates 'Altitude Above Channel' (AAC) and 'Wetness Index' (SWI) were selected to form the cLHS covariate space for the modified cLHS design. AAC shows the highest correlation to the target soil properties with $r_{average}$ = 0.65 and was therefore selected. The covariates 'Altitude' ($r_{average}$ = 0.65), 'Plane curvature' ($r_{average}$ = 0.63) and SWI ($r_{average}$ = 0.41) are reasonably correlated to the target variables with $r_{average}$ > 0.4. However, 'Altitude' and 'Plane curvature' were rejected, since both covariates indicate collinearity to ACC with $r_{collinearity}$ = 0.62 and $r_{collinearity}$ = 0.55. SWI shows a lower collinearity with $r_{collinearity}$ = 0.38 and was retained (Table 2).

The sample set size n was determined by analyzing the representation of the global cLHS covariate space by various sample set sizes using the statistical variance as indicator. The results show a decreasing difference between the global variance and the sample set variances with increasing sample set size (Fig. 1). This trend reveals an increased representation of the cLHS covariate space by sample sets with larger sample set sizes. The kneepoints of both curves of the covariate space (AAC, SWI) correspond to the sample set size n = 30, indicating the best tradeoff in terms of sampling effort. Thus, n = 30 was defined as final sample set size, resulting in 7.1 samples per km^2.

3.1.2 Modified conditioned Latin Hypercube Sampling

Corresponding to the final sample set size n = 30, the proposed cLHS design (cLHS$_{adapt}$) includes 30 strata. Each stratum contains a varying number of spatially scattered and discrete pixels, ranging from 2 to 82. These pixels represent potential alternative sample sites. The variation in the number of alternative sample sites and their spatial scattering is determined by the statistical distributions of the cLHS covariates 'AAC' and 'SWI'. After exclusion of inaccessible areas from the strata, the number of potential alternative sample sites ranges between 1 and 4 (Table 3).

Each stratum needs to be sampled at one specific sample site, while the combination of all sampled strata results in an optimized representation of the cLHS covariate space. Seven strata are occupied by fixed sample sites based on cLHS$_{orig}$ or the legacy samples: two sites refer to the cLHS$_{orig}$ design and match accessible pixels of a stratum. In the remaining five strata, legacy samples are available and used for the respective stratum. This results in 23 unsampled strata whether due to limited accessibility at the cLHS$_{orig}$ site or absent legacy samples (Table 3).

Since we integrated five legacy samples into the cLHS$_{adapt}$ design and designated 20 samples for validation, 30 legacy samples remained unused. We combined the latter with cLHS$_{adapt}$ (n = 30), resulting in the further calibration set

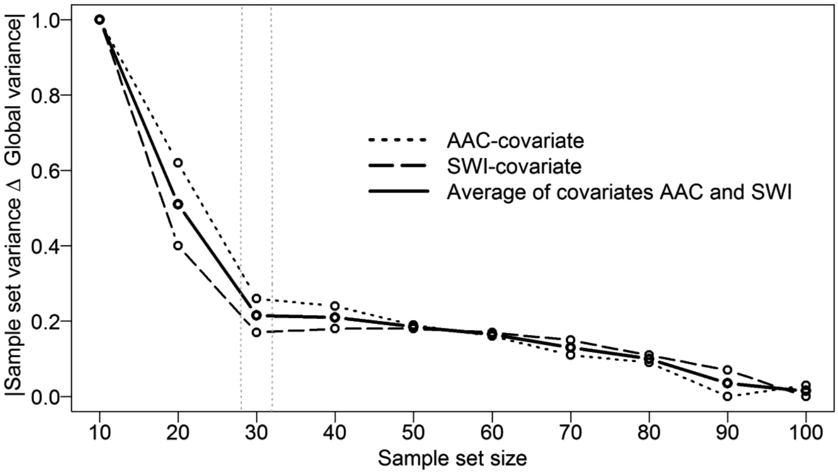

Figure 1: Representation of the global cLHS covariate space by various sample set sizes. The kneepoints of the curves (dashed line demarcation in grey) indicate the best tradeoff sample set size between sampling effort and representation of the specific covariate in the covariate space.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2: The covariates ranked according to the averaged correlations across the target soil properties ($r_{average}$), and the correlation between the covariates and the top ranked covariate (AAC) to indicate collinearity ($r_{collinearity}$).

Covariate	$r_{average}$	Rank	$r_{collinearity}$
AAC	**0.65**	**1**	**1**
Altitude	0.65	2	0.62
Plane curvature	0.63	3	0.55
SWI	**0.41**	**4**	**0.38**
Slope angle, maximum	0.38	5	0.44
Terrain ruggedness	0.31	6	0.01
Slope angle	0.29	7	0.03
Slope length	0.28	8	0.37
Convergence index	0.23	9	0.33
Land use	0.23	10	0.22
Mass balance index	0.2	11	0.22
Combined curvature	0.19	12	0.29
Northness	0.19	13	0.15
Eastness	0.18	14	0.10
Horizontal flow distance	0.18	15	0.07
Protection index	0.17	16	0.08
Profile curvature	0.16	14	0.27
Vertical flow distance	0.12	18	0.12
Flow accumulation	0.09	19	0.35
Overland flow distance	0.08	20	0.10
Position index	0.08	21	0.29
Catchment area	0.07	22	0.41

$cLHS_{adapt+}$ ($n = 60$), which we used to evaluate the proposed sampling and DSM approach.

The 23 lacking sample sites were determined by analyzing all possible combinations of accessible alternative sample sites across the so far unsampled strata. For each combination the 7 fixed sites were added, resulting in 10,592 test sample sets. The latter were tested for (1) the preservation of the covariate correlation and (2) the representation of the covariate space. Using $cLHS_{orig}$ as baseline, the most similar test sample set was designated as the final sample set $cLHS_{adapt}$.

Referring to the preserved correlation of the covariates, $cLHS_{adapt}$ shows a difference to $cLHS_{orig}$ in $r < 0.01$. Contrary, the extended sample set $cLHS_{adapt+}$ differs by $r > 0.05$. Referring to the representation of the cLHS covariate space, $cLHS_{adapt}$ deviates by 22% from the $cLHS_{orig}$, summed up across all quartiles of the entire covariate space (dev). The deviation of $cLHS_{adapt+}$ to the baseline is increased and amounts to 60% (Table 4). Thus, $cLHS_{adapt+}$ represents the covariate space with $n = 60$ far less than $cLHS_{adapt}$ with $n = 30$.

Table 3: Number of potentially (accessible) alternative sample sites and total number of sample sites per stratum. The strata sampled by original cLHS sites and strata occupied by legacy samples are indicated (dots).

Strata	Number of accessible alternative sample sites	Total number of sample sites	$cLHS_{orig}$ sample	Legacy sample
1	3	22	–	●
2	2	5	–	–
3	2	8	–	–
4	3	18	–	–
5	1	7	–	–
6	1	19	–	–
7	1	84	–	●
8	1	9	–	●
9	1	18	–	–
10	1	9	–	–
11	1	20	–	–
12	1	7	–	–
13	1	6	–	–
14	1	8	–	–
15	2	6	–	–
16	2	12	–	–
17	2	19	–	–
18	2	3	–	–
19	4	9	●	–
20	3	26	–	–
21	3	10	–	●
22	1	4	–	–
23	2	2	●	–
24	3	16	–	–
25	1	7	–	–
26	2	13	–	–
27	2	5	–	–
28	4	25	–	–
29	2	9	–	–
30	2	64	–	●

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 4: Comparison of the proposed sample set cLHS$_{adapt}$ (n = 30) and the extended sample set cLHS$_{adapt+}$ (n = 60) to the baseline sample set cLHS$_{orig}$ (n = 30). Deviations (dev) are indicated by (1) the correlation in the covariate space of the sample set (r), and (2) the proportional frequency of samples across the quartiles in the covariate space of the entire study area (Q1–3$_{AAC}$, Q1–3$_{SWI}$).

Sample set	*r*	Q1$_{AAC}$ / %	Q2$_{AAC}$ / %	Q3$_{AAC}$ / %	Q1$_{SWI}$ / %	Q2$_{SWI}$ / %	Q3$_{SWI}$ / %	dev / %
cLHS$_{orig}$	0.4003	23	50	27	23	50	27	–
cLHS$_{adapt}$	0.3942	27	46	27	27	53	20	22
cLHS$_{adapt+}$	0.3365	45	38	17	17	58	25	60

3.3 Spatial predictions

3.3.1 Calibration and validation sets

The results of the lab analysis (Fig. 2) of the two calibration sets cLHS$_{adapt}$ (n = 30) and cLHS$_{adapt+}$ (n = 60) show an average topsoil sand content of 6% in cLHS$_{adapt}$ and 9.5% in cLHS$_{adapt+}$. Comparing both calibration sets and all target sand fractions (CS, MS, FS), cLHS$_{adapt+}$ reveals higher averages and variabilities. This is pronounced for CS with an average of 2.4% and an inter quartile range (IQR) of 1.3% in cLHS$_{adapt}$ *versus* an average of 4.6% and an IQR of 8.9% in cLHS$_{adapt+}$. By contrast, the distributions of the target variables MS and FS show increased similarities between the calibration sets. For MS, the average amounts to 2% with an IQR of 0.9% in cLHS$_{adapt}$, while cLHS$_{adapt+}$ shows an average of 3% with an IQR of 4%. For FS, cLHS$_{adapt}$ shows an average of 1.6% with an IQR of 0.8% *versus* an average of 2% with an IQR of 2.2% in cLHS$_{adapt+}$. The distributions of the validation set (cLHS$_{val}$) show increased averages compared to both calibration sets and across all target variables (CS: 5%; MS: 3.5%; FS: 2.8%). The variabilities in cLHS$_{val}$ show higher similarities to cLHS$_{adapt+}$ across all target variables with IQR of 5.8% for CS, 3.7% for MS, and 3.6% for FS (Fig. 2).

3.3.2 Accuracy and uncertainty

The RF prediction models couple the calibration sets cLHS$_{adapt}$ (n = 30) and cLHS$_{adapt+}$ (n = 60) with all available covariates (Table 1) to spatially estimate all target sand fractions (CS, MS, FS). The accuracy evaluation by external and bootstrap validation performs similar with deviations in R^2 ranging from 1% to 4% across all target variables and validation methods. The absolute R^2-values across all target variables, both calibration sets, and validation methods range between 0.57 and 0.80 (Table 5).

The FS-models (cLHS$_{adapt}$: R^2 = 0.59; cLHS$_{adapt+}$: R^2 = 0.61) are outperformed by the CS-models (cLHS$_{adapt}$: R^2 = 0.63; cLHS$_{adapt+}$: R^2 = 0.67), while the MS-models (cLHS$_{adapt}$: R^2 = 0.71; cLHS$_{adapt+}$: R^2 = 0.80) perform best according to external validation. Across all target variables, the models calibrated by cLHS$_{adapt+}$ (n = 60) outperform the models using cLHS$_{adapt}$ (n = 30). The deviations amount to 2% for FS, 4% for CS and 9% for MS (Table 5).

The sum of all predicted sand fractions ideally amounts to 100%. We mapped the positive and negative deviations from this ideal for the predictions of both calibration sets, cLHS$_{adapt}$ and cLHS$_{adapt+}$ (Fig. 3). Both model approaches show increased deviations in the north and along the depression lines of the entire study area. For the cLHS$_{adapt}$-models, the deviations are pronounced with hotspots of overestimations in the northern depression lines, and underestimations in the central part. The summary statistics of the absolute deviations show a maximum of 9.4% and an average of 1.8% for the cLHS$_{adapt}$-models. The maximum of the cLHS$_{adapt+}$-models amounts to 7.6% with an average of 1.3%. Thus, the calibration set cLHS$_{adapt+}$ marginally outperforms cLHS$_{adapt}$, confirming the results of the accuracy estimations by external and bootstrap validation (Table 5).

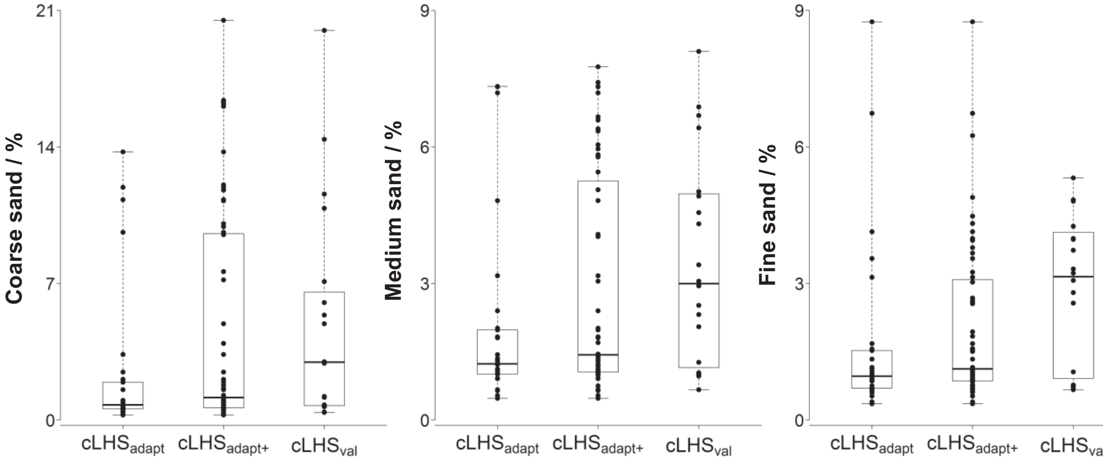

Figure 2: Distributions of the target soil properties (coarse sand, medium sand, fine sand) for both calibration sets (cLHS$_{adapt}$, cLHS$_{adapt+}$) and the validation set (cLHS$_{val}$).

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 5: Random Forest model accuracies (R^2, RMSE) based on external and bootstrap validation. The accuracy estimations refer to all calibration sets (cLHS$_{adapt}$, cLHS$_{adapt+}$) and target soil properties (CS: coarse sand, MS: medium sand, FS: fine sand).

External validation	**cLHS$_{adapt}$**		**cLHS$_{adapt+}$**	
	R^2	RMSE	R^2	RMSE
CS	0.63	4.03	0.67	3.75
MS	0.71	1.07	0.80	0.94
FS	0.59	0.34	0.61	0.38
Bootstrap validation	**cLHS$_{adapt}$**		**cLHS$_{adapt+}$**	
	R^2	RMSE	R^2	RMSE
CS	0.64	2.16	0.71	3.48
MS	0.69	0.91	0.78	1.01
FS	0.57	0.38	0.64	0.36

4 Discussion

4.1 Conditioned Latin Hypercube Sampling

The cLHS method identifies sample sites which are stratified in the cLHS covariate space in a two-step approach. First, strata are arranged in a hypercube, which spans the covariate space. Second, one sample site per stratum is selected according to an optimization procedure (*Metropolis* et al., 1953) to cover the covariate space by a combination of unique sample sites. The latter sample site selection process is conditioned by only selecting sites in the covariate space that also exist in the real world (*Minasny* and *McBratney*, 2006).

Roudier et al. (2012), *Mulder* et al. (2013), and *Clifford* et al. (2014) further constrained the sample site selection by penalizing locations with limited accessibility. This implies a potential bias in the final sample set, since inaccessible areas might occupy segments of the covariate space that are excluded from sampling. Instead of penalizing locations for the site selection process, we arranged test sample sets of all possible combinations of accessible sites to quantify their deviations from the original cLHS design. This enables to select the test sample set most similar to the original cLHS design.

Yet, the existence of alternative accessible sites in the strata cannot be guaranteed. *Clifford* et al. (2014) eludes the latter problem by analyzing the relation of each location in a defined neighborhood to the specific initial sample site, providing an ordered list of alternative sites for each initial site. According to a simulation study, *Clifford* et al. (2014) showed that the coverage of the covariate space remains preserved by replacing up to 50% of the initial sites by alternatives. However, it is the combination of one unique site per stratum that covers the cLHS covariate space. Thus, the generation of alternatives for each target site individually, follows the biased assumption that all other sites have been successfully sampled before. Instead of generating alternatives for each initial site individually, we selected alternatives by considering the combination of all accessible sites using the test sample sets. This avoids assumptions about previously sampled sites, while preserving the cLHS concept of covering the covariate space by the combination of one unique site within each stratum. Therefore, the bias, introduced by the additional constraint of including accessibility, is reduced for the presented approach. However, ensuring that at least one alternative site per stratum is available requires a decreased complexity of the covariate space to potentially increase the number of accessible sites in each stratum. Thus, we optimized the sample set size and reduced the number of covariates to the two most relevant (AAC; SWI; *Gessler* et al., 2000). Other studies, applying cLHS for soil sampling, used more cLHS covariates (four to ten) to build the cLHS covariate space (*Hengl* et al., 2003; *Roudier* et al., 2012; *Mulder* et al., 2013; *Schmidt* et al., 2014; *Taghizadeh-Mehrjardi* et al., 2014; *Kidd* et al., 2015). For these studies, a less

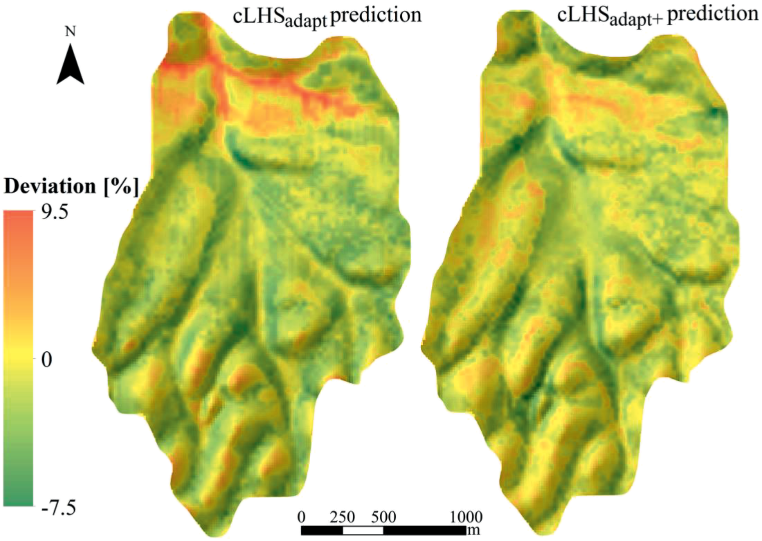

Figure 3: Mapped coherence of the predicted sum for the compositional target soil properties (coarse sand, medium sand, fine sand) and the calibration sets cLHS$_{adapt}$ and cLHS$_{adapt+}$. The color ramp indicates the range of positive and negative deviations from the ideal of 100%.

complex covariate space is irrelevant since either additional cLHS criteria are not included or the bias, due to assumptions about previously sampled sites, is not considered. The ratio of samples per km^2 varies between 0.005 and 0.465 in study areas with a size ranging from 720 km^2 to 12,800 km^2 for cLHS applications (*Mulder* et al., 2013; *Clifford* et al., 2014; *Taghizadeh-Mehrjardi* et al., 2014; *Kidd* et al., 2015). *Brungard* and *Boettinger* (2010) suggested a minimum ratio of 0.7 to 1 samples per km^2 using cLHS in a study area of a size of 300 km^2. We optimized the sample set size to n = 30 in a study area of 4.2 km^2, thus, 7.1 samples per km^2. Moreover, we showed that the deviation of the covariate space coverage from the original cLHS set is marginally (Table 4), while 93% of the original samples were replaced by alternatives (Table 3).

Clifford et al. (2014) used 8,669 legacy samples as predefined basis to locate 300 additional sample sites. Since the subset of legacy samples was not obtained purposively to cover the covariate space, redundancies within the subset potentially led to a bias in the combined sample set. In this context, *Carré* et al. (2007) proposed to analyze the distribution of legacy samples across the strata to evaluate the adequacy of legacy samples for representing the covariate space. Following this principle, we only used those five legacy samples which are located within a stratum, thus, simultaneously avoiding redundancies and reducing the sampling effort. Thus, the increased number of potential alternative sites per stratum also accommodates with the goal to integrate legacy samples, which are adequate to represent the covariate space. In the present study, only one legacy sample is available in the respective stratum, thus, no evaluation of preference is necessary. Apart from this situation, we suggest to consider multiple legacy samples within a stratum simply as potential accessible sites and follow the methodological procedure as described before.

Any cLHS or stratified sampling design must be seen as a specific case study, resulting in a partially limited generalizability. This is attributed to the assumptions that the covariate space, determined by local framework conditions, sufficiently describes the target soil property in space and time (*Clifford* et al., 2014). Besides these limitations, our approach is transferable to any other study area, considering that an increasing number of samples and resolution increases the computational load. Thus, our approach is suitable for small to medium study areas with medium to high limitations in accessibility.

4.2 Spatial prediction

Soil texture estimation using DSM typically shows accuracies less than R^2 = 0.5, while studies with R^2 > 0.7 are rare (*Malone* et al., 2009; *Lacoste* et al., 2011; *Wang* et al., 2012; *de Carvalho Junior* et al., 2014; *Mansuy* et al., 2014). Our results (Table 5) showed prediction accuracies with R^2 > 0.5 for all sand fractions and for both calibration sets (cLHS$_{\text{adapt}}$; cLHS$_{\text{adapt+}}$). Comparing the accuracies for external and bootstrap validation, the cLHS$_{\text{adapt+}}$ approach with doubled sample set size (n = 60) showed a marginal increase of 0.5% compared to the proposed cLHS$_{\text{adapt}}$ approach (n = 30). This trend is also reproduced considering the spatial uncertainty based on the coherence of the total sand fraction, which ideally amounts to 100%. For both approaches of the present study, the models overestimated the target variables in the northern depression lines, while the central part of the study area showed underestimations (Fig. 3). This variability in the uncertainty pattern is attributed to various errors in the entire modelling attempt (*Nelson* et al., 2011). The error sources refer to the measurement and processing of the covariates and soil property samples, to the model specifications and oversimplifications, to an imprecise localization of spatial data, and to calibration sets that represent the global distribution erratically in the space (*Bishop* et al., 2006; *Grimm* and *Behrens*, 2009).

However, the similarity between the two DSM approaches confirms the robustness of the proposed cLHS approach. Moreover, since our sample set size n = 30 is in statistical terms relatively small compared to other RF studies using n between 165 and 4,920 (*Grimm* et al., 2008; *Lacoste* et al., 2011; *Mansuy* et al., 2014), the accuracy of the proposed DSM approach is noticeable.

5 Conclusion

We present a DSM sampling design, which is based on the principles of conditioned Latin Hypercube Sampling (cLHS). The final sample set adequately reproduces the variation of selected terrain covariates, which serve as proxies for the target soil properties. The design compensates for limited field accessibility, integrates the use of legacy samples and uses an optimized sample set size. Consequently, our approach provides better operability in difficult terrain and improves efficiency in terms of temporal and monetary constraints compared to other cLHS approaches. We applied the approach in a small catchment in the Three Gorges Reservoir area at the Yangtze River in Hubei, Central China. Using a Random Forest regression model we estimated the topsoil sand fractions with convincing accuracies.

Acknowledgments

We thankfully acknowledge funding from the *German Ministry of Education and Research* (*BMBF, grant No. 03 G 0827A*) for the German-Sino research collaboration *YANGTZE GEO*. We thank *Bi Renneng* from the University of Erlangen-Nuremberg and *Alexander Strehmel* from the Christian-Albrechts-University of Kiel for data sharing. We thank all students from the China University of Geosciences (CUG) in Wuhan that were involved in field campaigns and laboratory analyzes. We are also very grateful for the general support of the whole YANGTZE GEO research group.

References

Behrens, T., *Schmidt, K.*, *Ramirez-Lopez, L.*, *Gallant, J.*, *Zhu, A.*, *Scholten, T.* (2014): Hyper-scale digital soil mapping and soil formation analysis. *Geoderma* 213, 578–588.

Behrens, T., *Schmidt, K.*, *Zhu, A.*, *Scholten, T.* (2010): The ConMap approach for terrain-based digital soil mapping. *Eur. J. Soil Sci.* 61, 133–143.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

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

Bishop, T. F. A., *Minasny, B.*, *McBratney, A. B.* (2006): Uncertainty analysis for soil-terrain models. *Int. J. Geog. Inf. Sci.* 20, 117–134.

Breiman, L. (2001): Random Forests. *Mach. Learn.* 45, 5–32.

Brungard, C. W., *Boettinger, J. L.* (2010): Conditioned Latin Hypercube Sampling: Optimal sample size for Digital Soil Mapping of Arid Rangelands in Utah, USA, in Boettinger, J. L., Howell, D. W., Moore, A. C., Hartemink, A. E., Kienast-Brown, S. (eds.): Digital Soil Mapping. Bridging Research, Environmental Application, and Operation. Progress in Soil Science. Vol. 2. Springer, Dordrecht, The Netherlands, pp. 67–75.

Brus, D., *De Gruijter, J.*, *Van Groeningen, J.* (2006): Designing Spatial Coverage Samples Using the *k*-means Clustering Algorithm, in Lagacherie, P., McBratney, A. B., Voltz, M. (eds.): Digital Soil Mapping, An Introductory Perspective. Developments in Soil Science. Vol. 31. Elsevier, Amsterdam, The Netherlands, pp. 183–192.

Cambule, A., *Rossiter, D.*, *Stoorvogel, J.* (2013): A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341–353.

Carré, F., *Girard, M.* (2002): Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. *Geoderma* 110, 241–263.

Carré, F., *McBratney, A. B.*, *Minasny, B.* (2007): Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma* 141, 1–14.

Clifford, D., *Payne, J. E.*, *Pringle, M. J.*, *Searle, R.*, *Butler, N.* (2014): Pragmatic soil survey design using flexible Latin hypercube sampling. *Comput. Geosci.* 67, 62–68.

de Carvalho Junior, W., *Lagacherie, P.*, *da Silva Chagas, C.*, *Filho, B.*, *Bhering, S. B.* (2014): A regional-scale assessment of digital soil attributes in a tropical hillslope environment. *Geoderma* 232-234, 479–486.

Dobermann, A., *Simbahan, G. C.* (2007): Methodology for using Secondary Information in Sampling Optimization for Making Fine-Resolution Maps of Soil Organic Carbon, in Lagacherie, P., McBratney, A. B., Voltz, M. (eds.): Digital Soil Mapping, An Introductory Perspective. Developments in Soil Science. Vol. 31. Elsevier, Amsterdam, The Netherlands, pp. 167–182.

DIN ISO 11277:2002-08 (2002): Bodenbeschaffenheit–Bestimmung der Partikelgrößenverteilung in Mineralböden–Verfahren mittels Siebung und Sedimentation, ISO 11277: 1998/Cor. 1: 2002. Beuth Verlag, Berlin, Germany.

Flannery, B. P., *Press, W. H.*, *Teukolsky, S. A.*, *Vetterling, W. T.* (1992): Numerical Recipes in FORTRAN: The Art of Scientific Computing. Cambridge University Press, New York, NY, USA.

Florinsky, I. V., *Eilers, R. .G.*, *Manning, G. R.*, *Fuller, L. G.* (2002): Prediction of soil properties by digital terrain modelling. *Environ. Modell Softw.* 17, 295–311.

Gessler, P. E., *Chadwick, O. A.*, *Chamran, F.*, *Althouse, L.*, *Holmes, K.* (2000): Modeling soil–landscape and ecosystem properties using terrain attributes. *Soil Sci. Soc. Am. J.* 64, 2046–2056.

Grimm, R., *Behrens, T.*, *Märker, M.*, *Elsenbeer, H.* (2008): Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. *Geoderma* 146, 102–113.

Grimm, R., *Behrens, T.* (2009): Uncertainty analysis of sample locations within digital soil mapping approaches. *Geoderma* 155, 154–163.

Hengl, T., *Rossiter, D. G.*, *Stein, A.* (2003): Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Soil Res.* 41, 1403–1422.

Hengl, T., *Heuvelink, G. B. M.*, *Stein, A.* (2004): A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 120, 75–93.

Heuvelink, G. B., *Brus, D. J.*, *de Gruijter, J. J.* (2007): Optimization of sample configurations for digital soil mapping of soil properties with universal Kriging, in Lagacherie, P., McBratney, A. B., Voltz, M. (eds.): Digital Soil Mapping, An Introductory Perspective. Developments in Soil Science. Vol. 31. Elsevier, Amsterdam, The Netherlands, pp. 137–152.

Huang, C., *Davis, L. S.*, *Townshend, J. R. G.* (2004): An assessment of support vector machines for land cover classification. *Int. J. Remote Sens.* 23, 725–749.

Kidd, D., *Malone, B.*, *McBratney, A.*, *Minasny, B.*, *Webb, M.* (2015): Operational sampling challenges to Digital Soil Mapping in Tasmania, Australia. *Geoderma Reg.* 4, 1–10.

Krol, B. (2008): Towards a Data Quality Management Framework for Digital Soil Mapping with Limited Data, in Hartemink, A. E., McBratney, A. B., Mendonça-Santos, M. L. (eds.): Digital Soil Mapping with Limited Data. Springer, Dordrecht, The Netherlands, pp. 136–149.

Lacoste, M., *Lemercier, B.*, *Walter, C.* (2011): Regional mapping of soil parent material by machine learning based on point data. *Geomorphology* 133, 90–99.

Lagacherie, P. (2008): Digital Soil Mapping: A State of the Art, in Hartemink, A. E., McBratney, A. B., Mendonça-Santos, M. L. (eds.): Digital Soil Mapping with Limited Data. Springer, Dordrecht, The Netherlands, pp. 3–14.

Liaw, A., *Wiener, M.* (2002): Classification and regression by randomForest. *R News* 2/3, 18–22.

Liu, J., *Liu, M.*, *Tian, H.*, *Zhuang, D.*, *Zhang, Z.*, *Zhang, W.*, *Tang, X.*, *Deng, X.* (2005): Spatial and temporal patterns of China's cropland during 1990–2000: An analysis based on Landsat TM data. *Remote Sens. Environ.* 98, 442–456.

Malone, B. P., *McBratney, A. B.*, *Minasny, B.*, *Laslett, G. M.* (2009): Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154, 138–152.

Mansuy, N., *Thiffault, E.*, *Paré, D.*, *Bernier, P.*, *Guindon, L.*, *Villemaire, P.*, *Poirier, V.*, *Beaudoin, A.* (2014): Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using k-nearest neighbor method. *Geoderma* 235–236, 59–73.

Mayr, T., *Palmer, R.* (2007): Digital Soil Mapping: An England and Wales Perspective, in Lagacherie, P., McBratney, A. B., Voltz, M. (eds.): Digital Soil Mapping, An Introductory Perspective. Developments in Soil Science. Vol. 31. Elsevier, Amsterdam, The Netherlands, pp. 365–376.

McBratney, A. B., *Mendonça-Santos, M. L.*, *Minasny, B.* (2003): On digital soil mapping. *Geoderma* 117, 3–52.

McKenzie, N. J., *Ryan, P. J.* (1999): Spatial prediction of soil properties using environmental correlation. *Geoderma* 89, 67–94.

McMillan, R. A. (2008): Experiences with Applied DSM: Protocol, Availability, Quality and Capacity Building, in Hartemink, A. E., McBratney, A. B., Mendonça-Santos, M. L. (eds.): Digital Soil Mapping with Limited Data. Springer, Dordrecht, The Netherlands, pp. 113–135.

Metropolis, N., *Rosenbluth, A. W.*, *Rosenbluth, M. N.*, *Teller, A. H.*, *Teller, E.* (1953): Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Minasny, B., *McBratney, A. B.* (2006): A conditioned Latin hypercube method for sampling in presence of ancillary data. *Comput. Geosci.* 32, 1378–1288.

Mora-Vallejo, A., *Claessens, L.*, *Stoorvogel, J.*, *Heuvelink, G. B. M.* (2008): Small scale digital soil mapping in Southeastern Kenya. *Catena* 76, 44–53.

Mulder, V. L., *De Bruin, S.*, *Schaepman, M. E.* (2013): Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. *Int. J. Appl. Earth Obs. Geoinf.* 21, 301–310.

Müller, H.-W., *Dohrmann, R.*, *Klosa, D.*, *Rehder, S.*, *Eckelmann, W.* (2009): Comparison of two procedures for particle-size analysis: Köhn pipette an X-ray granulometry. *J. Plant. Nutr. Soil Sci.* 172, 172–179.

Naumann, T. W., *Thompson, J. A.* (2014): Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma* 213, 385–399.

Nelson, M. A., *Bishop, T. F. A.*, *Triantafilis, J.*, *Odeh, I. O. A.* (2011): An error budget for different sources of error in digital soil mapping. *Eur. J. Soil Sci.* 62, 417–430.

Peters, J., *De Baets, B.*, *Verhoest, N.*, *Samson, R.*, *Degroeve, S.*, *De Becker, P.*, *Huybrechts, W.* (2007): Random forests as a tool for ecohydrological distribution modelling. *Ecol. Model.* 207, 304–318.

Rabe, A., *Van der Linden, S.*, *Hostert, P.* (2010): imageSVM, Version 2.1. Available at: www.hu-geomatics.de (last accessed: August 29, 2013).

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

Roudier, P., *Hewitt, A. E.*, *Beaudette, D. E.* (2012): A Conditioned Latin Hypercube Sampling Algorithm Incorporating Operational Constraints, in Minasny, B., Malone, B. P., McBratney, A. B. (eds.): Soil Assessment and Beyond. CRC Press, Sydney, Australia, pp. 227–231.

RapidEye (2012): Satellite Imagery Specifications–version 4.1. Available at: www.rapideye.com (last accessed: August 05, 2012).

SAGA GIS (2011): System for Automated Geoscientific Analyses (Version 2.0.6.). SAGA User Group Association, Hamburg, Germany. Available at: www.saga-gis.org (last accessed: September 14, 2013).

Schmidt, K., *Behrens, T.*, *Daumann, J.*, *Ramirez-Lopez, L.*, *Werban, U.*, *Dietrich, P.*, *Scholten, T.* (2014): A comparison of calibration sampling schemes at the field scale. *Geoderma* 232–234, 243–256.

Schönbrodt-Stitt, S., *Bosch, A.*, *Behrens, T.*, *Hartmann, H.*, *Shi, X.*, *Scholten, T.* (2013): Approximation and spatial regionalization of rainfall erosivity based on sparse data in a mountainous catchment of the Yangtze River in Central China. *Environ. Sci. Pollut. Res.* 20, 6917–6933.

Scull, P., *Franklin, J.*, *Chadwick, O. A.*, *McArthur, D.* (2003): Predictive soil mapping: a review. *Prog. Phys. Geog.* 27, 171–197.

Strehmel, A., *Schönbrodt-Stitt, S.*, *Buzzo, G.*, *Dumperth, C.*, *Stumpf, F.*, *Zimmermann, K.*, *Bieger, K.*, *Behrens, T.*, *Schmidt, K.*, *Bi, R.*, *Rohn, J.*, *Hill, J.*, *Udelhoven, T.*, *Xiang, W.*, *Shi, X.*, *Cai, Q.*, *Jiang, T.*, *Fohrer, N.*, *Scholten, T.* (2015): Assessment of geo-hazards in a rapidly changing landscape: the three Gorges Reservoir Region in China. *Environ. Earth Sci.* 74, 4939–4960.

Sulaeman, Y., *Minasny, B.*, *McBratney, A. B.*, *Sarwani, M.*, *Sutandi, A.* (2013): Harmonizing legacy soil data for digital soil mapping in Indonesia. *Geoderma* 192, 77–85.

Taghizadeh-Mehrjardi, R., *Minasny, B.*, *Sarmadian, F.*, *Malone, B.* (2014): Digital soil mapping of soil salinity in Ardakan region, central Iran. *Geoderma* 213, 15–28.

Thomas, M., *Clifford, D.*, *Bartley, R.*, *Philip, S.*, *Brough, D.*, *Gregory, L.*, *Willis, R.*, *Glover, M.* (2015): Putting regional digital soil mapping into practice in Tropical Northern Australia. *Geoderma* 241, 145–157.

Wang, D.-C., *Zhang, G.-L.*, *Pan, X.-Z.*, *Zhao, Y.-G.*, *Zhao, M.-S.*, *Wang, G.-F.* (2012): Mapping soil texture of a plain area using Fuzzy-*c*-Means Clustering method on land surface diurnal temperature difference. *Pedosphere* 22, 394–403.

© 2016 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim