Catena 153 (2017) 30–38

Contents lists available at ScienceDirect

Catena

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

Uncertainty-guided sampling to improve digital soil maps

Felix Stumpf [a,b,*], Karsten Schmidt [a], Philipp Goebes [a], Thorsten Behrens [a], Sarah Schönbrodt-Stitt [c], Alexandre Wadoux [d], Wei Xiang [e], Thomas Scholten [a]

[a] Department of Geoscience, Chair of Soil Science and Geomorphology, University of Tübingen, Rümelinstraße 19–23, 72070 Tübingen, Germany
[b] Agroscope, Institute for Sustainable Sciences ISS, Reckenholzstrasse 191, 8046 Zürich, Switzerland
[c] Department of Remote Sensing, Institute of Geography and Geology, University of Wuerzburg, Oswald-Kuelpe-Weg 86, 97074 Wuerzburg, Germany
[d] Wageningen UR, Department of Soil Geography and Landscape, Droevendaalsesteeg 3, 6708 Wageningen, Netherlands
[e] Department of Geotechnical Engineering and Engineering Geology, China University of Geosciences, Lumo Road 388, 430074 Wuhan, PR China

ARTICLE INFO

Article history:
Received 2 July 2015
Received in revised form 22 December 2016
Accepted 26 January 2017
Available online 31 January 2017

Keywords:
Soil landscape modeling
Spatial uncertainty
Soil sampling
Soil prediction improvement
Random Forest
Three Gorges Reservoir Area

ABSTRACT

Digital soil mapping (DSM) products represent estimates of spatially distributed soil properties. These estimations comprise an element of uncertainty that is not evenly distributed over the area covered by DSM. If we quantify the uncertainty spatially explicit, this information can be used to improve the quality of DSM by optimizing the sampling design. This study follows a DSM approach using a Random Forest regression model, legacy soil samples, and terrain covariates to estimate topsoil silt and clay contents in a small catchment of 4.2 km^2 in the Three Gorges Reservoir Area, Central China. We aim (i) to introduce a method to derive spatial uncertainty, and (ii) to improve the initial DSM approaches by additional sampling that is guided by the spatial uncertainty. The proposed uncertainty measure is based on multiple realizations of individual and randomized decision tree models. We used the spatial uncertainty of the initial DSM approaches to stratify the study area and thereby to identify potential sampling areas of high uncertainties. Further, we tested how precisely available legacy samples cover the variability of the covariates within each potential sampling area to define the final sampling area and to apply a purposive sampling design. For the final Random Forest model calibration, we combined the legacy sample set with the additional samples. This uncertainty-driven DSM refinement was evaluated by comparing it to a second approach. In this second approach, the additional samples were replaced by a random sample set of the same size, obtained from the entire study area. For the comparative analysis, external, bootstrap-, and cross-validation was applied. The DSM approach using the uncertainty-driven refinement performed best. The averaged spatial uncertainty was reduced by 31% for silt and by 27% for clay compared to the initial DSM approach. Using external validation, the accuracy increased by the same proportions, while showing an overall accuracy of $R^2 = 0.59$ for silt and $R^2 = 0.56$ for clay.

1. Introduction

Information on soil properties and their spatial distribution is essential for environmental protection and management. Conventional Soil Mapping (CSM) is based on a qualitative description of the soil-landscape relationship (Grunwald, 2005; Jenny, 1941; Ruhe, 1956). This conceptual model is used to select representative locations for soil profile observations, which are associated with discrete and homogeneous landscape entities (Kempen et al., 2012; McBratney et al., 2000). CSM products have been criticized for irreproducibility, discrete data representation, and impracticability due to a lack of specificity and quantified uncertainties (Behrens and Scholten, 2006; Behrens et al., 2010; Goovaerts and Journel, 1995; Minasny and McBratney, 2016; Scull et al., 2003).

Digital Soil Mapping (DSM) addresses these shortcomings by using quantitative models to link soil observations at points with spatially exhaustive environmental covariates (McBratney et al., 2003; McMillan, 2008; Scull et al., 2003). Due to its quantitative nature, DSM products are reproducible and allow for continuous data representation. Moreover, models can be updated for specific purposes, and uncertainty can be quantified (Heuvelink, 1996; Kempen et al., 2012).

The success of DSM largely depends on the availability and distribution of soil observations, and on the availability and quality of environmental covariates (Krol, 2008). While covariates can be derived inexpensively from remote sensing data (McBratney et al., 2003; Mendonça-Santos et al., 2008), the use of legacy soil data becomes increasingly interesting to avoid cost-intensive sampling campaigns (Krol, 2008; Lagacherie, 2008; Stumpf et al., 2016). Moreover, legacy data potentially serve as a knowledge base for understanding the local

[*] Corresponding author at: University of Tübingen, Department of Geoscience, Chair of Soil Science and Geomorphology, Rümelinstraße 19–23, 72070 Tübingen, Germany.
E-mail address: felix.stumpf@uni-tuebingen.de (F. Stumpf).

http://dx.doi.org/10.1016/j.catena.2017.01.033
0341-8162/© 2017 Elsevier B.V. All rights reserved.

soil-landscape relationship, and thus for planning new sampling campaigns (Bui and Moran, 2003; Lagacherie et al., 1995).

DSM products present an approximated description of spatially distributed soil properties that comprise an element of uncertainty (Minasny and Bishop, 2008; Webster and Oliver, 2006). According to Nelson et al. (2011), the uncertainty in DSM originates from four error sources, which are (i) covariate error, (ii) model error, (iii) positional error, and (iv) analytical error. The covariate error is sourced in the measurement or in an additional interpolation error if the data requires geostatistical pre-processing (cf. Bishop et al., 2006). The model error refers to an insufficient understanding, therefore to an oversimplification of real processes (Minasny and Bishop, 2008). Depending on the type of model, the error can be ascribed to model specifications, estimations of model parameters, or interpolations. McBratney et al. (2006) quantified the model error by applying bootstrapping that fits a model to different realizations of a data set to obtain an error variance. The positional error refers to imprecise localizations of spatial data, which is sourced in measurement errors of the GPS technology (cf. Grimm and Behrens, 2010). The analytical error refers to measurement errors of soil properties occurring during the laboratory analysis. According to Viscarra Rossel and McBratney (1998), the analytical errors are low compared to other DSM error sources and, in most cases, well documented. Additionally, the use of legacy data for DSM prevents monetary and temporal expenditures such as field sampling and laboratory analyses (Cambule et al., 2013). However, the increase in practicability is accompanied by additional errors, particularly if legacy data stem from multiple sources (Carré et al., 2007; Krol, 2008).

Due to the various error sources and in order to valorize DSM products from a user's perspective, a comprehensive uncertainty analysis is required (Mora-Vallejo et al., 2008; Stoorvogel et al., 2009; Sun et al., 2013; Webster and Oliver, 2006; Wellmann, 2013). According to Minasny and Bishop (2008), an uncertainty analysis should give response to three questions:

(i) How accurate is the prediction?
(ii) How is uncertainty spatially distributed?
(iii) Where should additional data be acquired to reduce uncertainties?

In DSM, uncertainty analysis is often limited to the prediction accuracy, which is conventionally based on the variance between observed and predicted values only at validation sites (Bishop et al., 2001; Brus et al., 2011; Everitt, 2002; Finke, 2007; Grunwald, 2009; Hengl and Husnjak, 2006; Krol, 2008). However, if kriging methods are applied to interpolate soil properties, a spatially explicit error based on the variogram function can be reported along with predictions (Bourennane et al., 2007; Carré and Girard, 2002; Diodato and Ceccarelli, 2006; Lark and Lapworth, 2012; Li and Heap, 2011; Knotters et al., 1995; Li and Heap, 2011; Qu et al., 2013; Sun et al., 2013; Tutmez and Hatipoglu, 2010). Recent studies derived spatial uncertainty from pixel-wise prediction intervals (PI) or confidence intervals (CI) using simulation experiments or spatial modeling (Heuvelink, 2014; Lacoste et al., 2016; Malone et al., 2011; Mulder et al., 2016; Viscarra Rossel et al., 2014).

The present study builds on two initial DSM campaigns using Random Forest regressions and legacy samples for estimating topsoil silt (2–63 μm) and clay (<2 μm) contents in a small catchment in the Three Gorges Reservoir Area (TGRA), Central China.

The goal of the study is to improve the initial DSM products by using a spatial uncertainty analysis to acquire additional soil data for model calibration. Thus, the objective of the study is a method (i) to derive spatially explicit uncertainty, and (ii) to identify areas relevant to acquire additional soil data.

Our study was evaluated by comparing three DSM protocols, which only differ by the calibration sample sets. We used (i) the legacy sample set of the initial DSM campaigns, (ii) the legacy sample set augmented by random sampling in the entire study area, and (iii) the legacy sample set augmented by the presented methods for additional soil data acquisition. For the comparative analysis, the DSM products are assessed by accuracy estimates of external, bootstrap-, and cross-validation, as well as by the presented spatially explicit uncertainty measure.

2. Material and methods

2.1. Study site and geodatabase

The study area is a drainage basin of 4.2 km^2. As a part of the TGRA, it is located at the middle reaches of the Yangtze River in Hubei province, Central China (31°1′24″N, 110°20′35″E). The elevation ranges from 469 m to 1483 m a.s.l. with an average elevation of 1053 m a.s.l. Slope inclinations range from 0° to 53° with an average of 26°. With 72%, the majority of slopes is exposed to the north (Schönbrodt-Stitt et al., 2013; Strehmel et al., 2015; Stumpf et al., 2016). According to FAO (2006), silty clay (SiC) is predominant. A large area of woodland (81%) alternates with scattered plots of cropland (15%) and sparsely distributed farm buildings (4%).

Conventionally, DSM input data is based on Jenny's state factor paradigm for soil development (Jenny, 1941) and its advancement to the SCORPAN approach for quantitative soil mapping (McBratney et al., 2003). According to SCORPAN, the soil property (S_a) or soil class (S_c) is a function of other soil properties (s), climate (c), organisms (o), relief (r), parent material (p), age (a), and the space or spatial position (n). However, in the study area, data availability was generally scarce and hazy atmospheric conditions throughout the year prevent the use of remotely sensed spectral data as covariates. Thus, we deviated from the SCORPAN paradigm and used a digital elevation model (cell size 25 m × 25 m) as primary data source (cf. Behrens et al., 2014; Behrens et al., 2010; Ließ et al., 2012; McBratney et al., 2003). We derived a pool of continuous terrain attributes using SAGA GIS (SAGA GIS, 2011) and selected a subset (Table 1) to describe regional and local repositioning processes of silt and clay contents (cf. Stumpf et al., 2016).

In two sampling campaigns in 2012 and 2013 from a previous study (Stumpf et al., 2016), 80 topsoil samples were obtained and denoted as legacy samples for this study. We selected 60 samples, using a stratified random sampling across the quartiles of the original data set to preserve the distribution and form the calibration set (LD). The remaining 20 samples were used for external validation (LD$_{val}$). We augmented LD by additional 30 samples according to the proposed uncertainty-guided sampling to form a second calibration set (LD$_{+\mathrm{lhs}}$) with a sample set size of $n = 90$. For a comparative evaluation, we generated a third calibration set (LD$_{+\mathrm{rand}}$) by augmenting LD with additional 30 samples, obtained according to a simple random sampling within the entire study area ($n = 90$).

All soil samples used in this study were obtained and analyzed identically. At each sampling site five topsoil (0–25 cm depth) sub samples were pooled; four from the corner points of a surrounding 40 cm × 40 cm square and one from the center point of the square. For particle size analysis, the samples were dried (40 °C) and sieved (<0.63 mm). Subsequently, silt and clay contents were separated using the Sedigraph III 5120 by micromeritics GmbH (cf. Yalamaç et al., 2014).

2.2. Modeling and spatial uncertainty

We used the Random Forest (RF) regression algorithm to link point observations of silt and clay contents with environmental covariates. RF presents a non-parametric ensemble learner, which is based on averaging the results of multiple randomized decision tree models for the final predictions (Breiman, 2001). The performance of RF primarily depends on a large number of decision trees, while this tree population should be as diverse as possible (Grimm and Behrens, 2010; Hansen and Salamon,

Table 1
Environmental covariates with summary statistics.

Covariate	Unit	Minimum	Maximum	Average	Standard deviation
Elevation	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 angle, maximum	Degree	0	0.72	0.42	0.13
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

1990; Peters et al., 2007; Prasad et al., 2006). RF increases the diversity within the tree population by using a randomized (by bootstrapping) subset of point observations for each tree. Furthermore, the best split is selected from a randomly sampled subset of covariates at each node of the individual trees. The best split is found by maximizing the difference in the mean square error (MSE) between the parent node and the left and right child nodes (Ließ et al., 2012; Viscarra Rossel et al., 2014). The subset of point observations, which is not used to build a specific tree, refers to out-of-bag (OOB) data. Using this OOB data for validating the respective trees, a RF error estimate is derived by averaging the MSE over all trees (OOB-error). The number of trees (k_{trees}) and the size of the covariate subset at each node (m_{try}) are user-defined model parameters. Both can be determined by comparing the OOB-errors of various RF realizations with different settings for k_{trees} and m_{try} (Grimm et al., 2008; Schmidt et al., 2010).

We derived a measure of spatially explicit uncertainty by using the mapped predictions of the randomized tree models within RF as a raster stack of equiprobable simulations (Goovaerts, 2001). The pixel-wise variability among the stacked simulations refers to their local uniformity, and thus to the local uncertainty of the RF predictions. Consequently, a high statistical variance at a specific pixel indicates increased uncertainty and vice versa (Sun et al., 2013). We defined the spatial uncertainty measure err_{var} for each pixel j as follows:

$$err_{var}(j) = \frac{1}{k-1}\ \sum_{i=1}^{k}(x_i-\overline{x})^2, \qquad (1)$$

where k is the number of simulations ($i = 1, 2, \ldots, k$) and x refers to their prediction results.

We set up two initial RF approaches ($RFLD_{silt}$ and $RFLD_{clay}$) to predict silt and clay contents by using the terrain attributes as covariates (Table 1) and the legacy sample set LD as point observations (with $k_{trees} = 1500$ and $m_{try} = 9$). Moreover, we used err_{var} to compute uncertainty maps for both approaches, denoted as $RFLD_{silt}{}^*$ and $RFLD_{clay}{}^*$. The R-packages ‘randomForest’ by Liaw and Wiener (2002) and “base” (R Core Team, 2014) were used for processing.

2.3. Uncertainty-guided sampling

The initial RF approaches $RFLD_{silt}$ and $RFLD_{clay}$ were calibrated by legacy samples (LD), which were not purposively obtained to cover the variability of the topsoil silt or clay contents in the study area. We refined the initial approaches by augmenting LD with respect to both target soil properties. The sampling design for the additional samples is based on identifying an area of high uncertainties using $RFLD_{silt}{}^*$ and $RFLD_{clay}{}^*$, and an adapted Latin Hypercube sampling design according to Stumpf et al. (2016).

We normalized $RFLD_{silt}{}^*$ and $RFLD_{clay}{}^*$ by scaling their ranges between 0 and 1. Then, we stacked the normalized uncertainty maps and retained the respective larger value for each pixel. This results in a new uncertainty map $RFLD_{silt/clay}{}^*$, which we used to determine a subarea of increased uncertainty, and thus to obtain additional samples.

Primarily, we defined four potential sampling areas referring to the entire study area (Q1) and to the quartile-breaks >25% (Q2), >50% (Q3), and >75% (Q4) of $RFLD_{silt/clay}{}^*$. Subsequently, we tested how precisely the legacy samples cover the variability of the target soil properties within each of those potential sampling areas. Since the covariates serve as proxies for the soil variability, we analyzed the divergence of the covariate distributions between the potential sampling area and available legacy sample sites. Comparing the divergence of all potential sampling areas and considering their demarcation by the quartile-breaks of $RFLD_{silt/clay}{}^*$, we assumed an increasing divergence from Q1 to Q4. Therefore, we selected this certain area for additional sampling, where the divergence starts to level up disproportionally. As measure for the divergence, we used the sum of the Kullback-Leibler divergences (KL; Kullback and Leibler, 1951) over all covariates, defined as follows:

$$KL\ (P_1 \| P_2) = \sum_{t=1}^{m}\left(\int p_1(x)\ log\frac{p_1(x)}{p_2(x)}dx\right)_t \qquad (2)$$

where P_1 is the covariate distribution in the potential sampling area, P_2 is the covariate distribution based on available legacy sample sites, p_1 and p_2 indicate the densities of P_1 and P_2, and m is the number of covariates ($t = 1, 2, \ldots, m$). For processing, we used the R-packages “flexmix” (Grün and Leisch, 2008) and “base” (R Core Team, 2014).

Within the final sampling area, we applied the adapted Latin Hypercube sampling design to obtain 30 additional topsoil samples (Stumpf et al., 2016). This design follows conditioned Latin Hypercube Sampling (cLHS), which optimally covers the variability of multiple covariates by the sample set with a limited sample set size (Brungard and Boettinger, 2010; Minasny and McBratney, 2006). The adaption additionally compensates for limited field accessibility, prevents redundancies in the covariate space referring to legacy samples, and enables for sampling several soil properties by one design. We used the R-packages “clhs” (Roudier, 2011) and “base” (R Core Team, 2014) for processing.

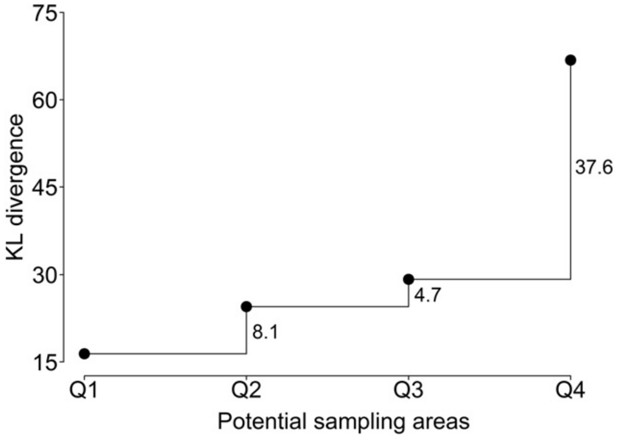

Fig. 1. Determination of the sampling area. The x-axis shows the potential sample areas Q1 to Q4. The y-axis shows the summarized Kullback-Leibler (KL) divergences between the covariate distributions in the area and for the legacy samples.

2.4. Prediction and evaluation

For estimating topsoil silt and clay contents, we set up three RF DSM approaches (RFLD, RFLD$_{+\text{rand}}$, RFLD$_{+\text{lhs}}$) by using all predictors (Table 1) combined with the calibration sample sets LD, LD$_{+\text{rand}}$, and LD$_{+\text{lhs}}$ for the respective approach (RFLD$_{\text{silt}}$, RFLD$_{\text{clay}}$, RFLD$_{+\text{rand\_silt}}$, RFLD$_{+\text{rand\_clay}}$, RFLD$_{+\text{lhs\_silt}}$, RFLD$_{+\text{lhs\_clay}}$). RF model settings were defined as the number of trees $k_{trees} = 1500$ and the size of the covariate subset at each node $m_{try} = 9$. For processing, we used the R-package 'randomForest' (Liaw and Wiener, 2002).

We evaluated the DSM approaches by using the presented pixel-wise uncertainty measure err_{var}, which was normalized across the full spatial domain and its average $\overline{ERR_{var}}$ as aggregated measure. Moreover, we used the accuracy measures coefficient of determination (R^2) and root mean squared error ($RMSE$) for observed sample sites. R^2 presents a measure that gives the proportion how well the variance of observed values is explained by predicted values. R^2 ranges between 0 and 1, while an increased R^2-value indicates increased certainty of making predictions from the model. $RMSE$ outlines a measure of the differences between predicted and observed values. These residuals, respectively prediction errors are aggregated into a single $RMSE$ error, giving a scale-dependent measure of the predictive capability of the model. For v instances in a data set ($l = 1, 2, v$) R^2, respectively $RMSE$ is defined as follows:

$$R^2 = 1 - \frac{\sum_{l=1}^{v}\left(y_l^{ob} - y_l^{pre}\right)^2}{\sum_{l=1}^{v}\left(y_l^{ob} - \overline{y^{ob}}\right)^2} \tag{3}$$

and

$$RMSE = \sqrt{\frac{\sum_{l=1}^{v}\left(y_l^{pre} - y_l^{ob}\right)^2}{v}} \tag{4}$$

where y^{pre} presents the predicted values and y^{ob} the observed values.

We applied external- (val), cross- (cv; 10-fold), and bootstrap- ($boot$; 10-fold) validation as accuracy estimation methods. For val, we used LD$_{\text{val}}$ as validation set, while cv and $boot$ are based on resampling the respective calibration set. Thus, cv and $boot$ provide a generalization error of a calibration model and do not require model-free validation data (Good, 2006). The cv estimator provides a nearly unbiased but highly variable accuracy estimate (Borra and Di Ciaccio, 2010; Kim, 2009), while $boot$ is biased upwards but less variable, and thus appropriate for smaller sample sets (Efron and Tibshirani, 1993; Molinaro et al., 2005). For processing, we used the R-package "caret" by Kuhn (2009).

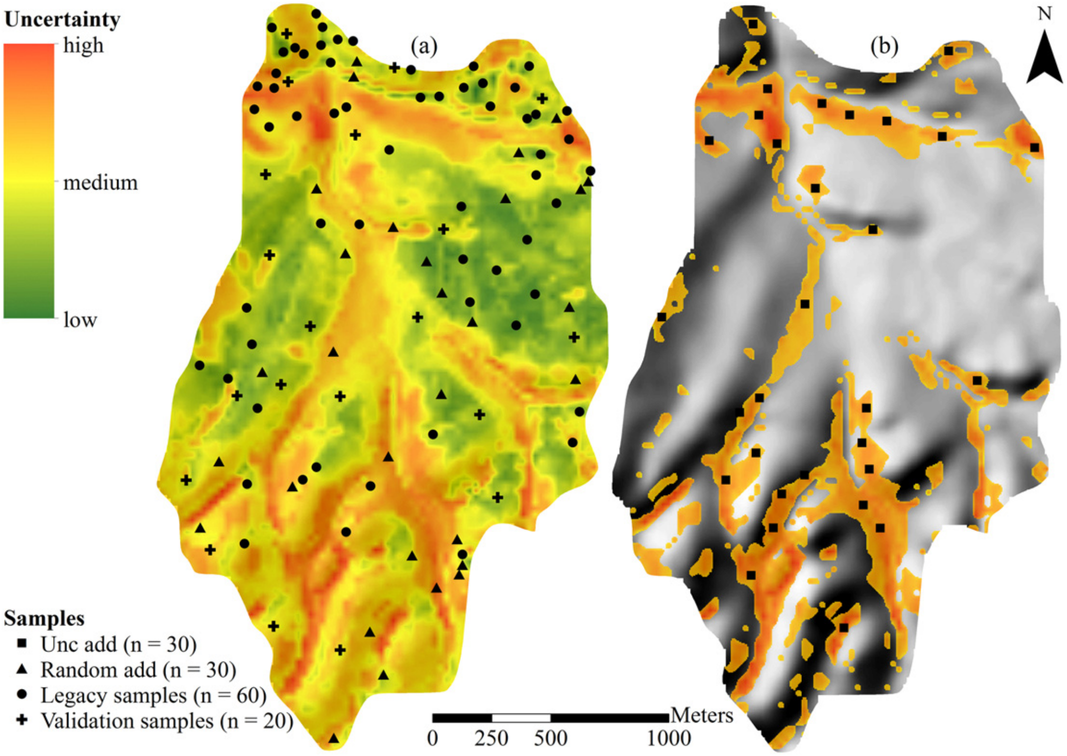

Fig. 2. (a) Combined spatial uncertainty (RFLD$_{\text{silt/clay}}$*) of initial silt and clay predictions with the legacy sample set (LD), the additive samples according to a simple random sampling (Random add) and the validation data (LD$_{\text{val}}$). (b) Area for uncertainty-guided sampling with the additive samples (Unc add).

3. Results

3.1. Spatial uncertainty-guided sampling

The combined uncertainty map of the initial DSM approaches $\mathrm{RFLD}_{\mathrm{silt/clay}}{}^{*}$ shows an uncertainty $\overline{ERR_{var}}$ of 0.63 with a standard deviation (SD) of 0.12 in the potential sampling area Q1. Since the potential sampling areas Q2, Q3, and Q4 were defined according to the quartiles-breaks of the combined uncertainty (>25%, >50%, >75%), we expected a successively increasing$\overline{ERR_{var}}$. We found $\overline{ERR_{var}} = 0.72$ in Q2 (SD = 0.1), $\overline{ERR_{var}} = 0.78$ in Q3 (SD = 0.07), and $\overline{ERR_{var}} = 0.86$ in Q4 (SD = 0.04).

Comparing each distribution of exhaustive covariates with the covariate distributions from available legacy samples in Q1 to Q4, a disproportionally increased divergence in Q4 was revealed. The summarized Kullback-Leibler (KL) divergences in Q1 amounts to 16.4, in Q2 to 24.5, in Q3 to 29.2, and in Q4 to 66.8. However, the increase in Q4 amounts to 130% compared to 19% in Q3 and 49% in Q2. Thus, the available legacy samples in Q4 outline a disproportionally decreased coverage of the exhaustive covariate space (Fig. 1). This results in Q4 as the final sampling area to obtain the additional samples of $\mathrm{LD}_{+\mathrm{lhs}}$ (Fig. 2b).

The combined uncertainty map $\mathrm{RFLD}_{\mathrm{silt/clay}}{}^{*}$ reveals increased values in the north and south of the study area where the topography shows increased heterogeneity. Uncertainty hotspots with $err_{var} > 0.8$ were predicted for areas along the topographic depression lines. In the central-west and -east of the study area, coherent areas of decreased uncertainties with $err_{var} < 0.5$ occur (Fig. 2a).

The $\mathrm{LD}_{+\mathrm{lhs}}$ samples were obtained according to an adapted cLHS design that optimally covers the variability of multiple covariates (Fig. 2b). The calibration set LD shows a cluster in the north of the study area. The sample sites of the additional samples for $\mathrm{LD}_{+\mathrm{rand}}$ as well as the sites for the validation set (*val*) were evenly distributed (Fig. 2a).

3.2. Predictions and evaluation

3.2.1. Calibration sets and predictions

According to the laboratory analysis, the model calibration sets LD ($n = 60$), $\mathrm{LD}_{+\mathrm{rand}}$ ($n = 90$), and $\mathrm{LD}_{+\mathrm{lhs}}$ ($n = 90$) show similar patterns in central tendency and variability for both target soil properties. The average topsoil silt content amounts to 59.4% (SD = 9.2) in LD, to 59.9% (SD = 9.0) in $\mathrm{LD}_{+\mathrm{rand}}$, and to 60.8% (SD = 8.9) in $\mathrm{LD}_{+\mathrm{lhs}}$. The average topsoil clay content varies from 28.5% (SD = 6) in LD, to 29.6% (SD = 6) in $\mathrm{LD}_{+\mathrm{rand}}$, and to 28.9% (SD = 5.9) in $\mathrm{LD}_{+\mathrm{lhs}}$ (Fig. 3).

The prediction approaches (RFLD, $\mathrm{RFLD}_{+\mathrm{rand}}$, $\mathrm{RFLD}_{+\mathrm{lhs}}$) of silt and clay contents show slightly increased averaged values compared to the averaged observed values from laboratory analysis. The increase varies from 2% to 2.5% for silt and from 1% to 2.7% for clay. The predicted average silt content amounts to 61.6% (SD = 3.7) for RFLD, to 61.9% (SD = 4.4) for the $\mathrm{RFLD}_{+\mathrm{rand}}$, and to 63.3% (SD = 3.9) for $\mathrm{RFLD}_{+\mathrm{lhs}}$. The predicted average clay content shows an average of 29.9% for RFLD (SD = 2.6), 30.7% (SD = 2.8) for $\mathrm{RFLD}_{+\mathrm{rand}}$, and 29.6% (SD = 2.5) $\mathrm{RFLD}_{+\mathrm{lhs}}$ (Fig. 3).

All prediction approaches (RFLD, $\mathrm{RFLD}_{+\mathrm{rand}}$, $\mathrm{RFLD}_{+\mathrm{lhs}}$) show a similar trend with increasing silt contents from the north to the south. In the topographic depression lines, the silt contents are generally decreased (<50%). The predicted clay contents show lowest values in the very north of the study area (<30%). In the middle part, increased clay contents are homogeneously distributed. In the south, clay contents are generally increased compared to the very north, but decreased and less homogeneously distributed compared to the central part (Fig. 4).

3.2.2. Model performance and comparison

The prediction approaches show a quality improvement from RFLD, to $\mathrm{RFLD}_{+\mathrm{rand}}$, and $\mathrm{RFLD}_{+\mathrm{lhs}}$. This trend applies for both target soil properties with respect to the uncertainty ($\overline{ERR_{var}}$) and accuracy measures (R^2, *RMSE* based on *cv*, *boot*, and *val*). An exception occurs for the $\mathrm{RFLD}_{\mathrm{silt}}$ approach, which shows a slightly increased *RMSE* compared to $\mathrm{RFLD}_{+\mathrm{rand\_silt}}$ according to *cv* and *boot*. Besides, the trend is consistent with a gradual quality improvement between 2% and 9% for the accuracy measures. For the uncertainty measure, the trend is disproportional with a quality increase of 20% from RFLD to $\mathrm{RFLD}_{+\mathrm{rand}}$ for silt (clay: 11%), and 2% from $\mathrm{RFLD}_{+\mathrm{rand}}$ to $\mathrm{RFLD}_{+\mathrm{lhs}}$ for both target soil properties (Fig. 5).

When comparing the target soil properties, the quality of the clay predictions is slightly increased. This applies for the uncertainty measure ($\overline{ERR_{var}}$), as well as for *cv*- and *boot*- accuracy measures. Contrary, the quality difference between silt and clay predictions is more balanced with respect to *val* (Fig. 5).

Exemplary, the uncertainty ($\overline{ERR_{var}}$) for silt decreases from RFLD with $\overline{ERR_{var}} = 0.68$ (clay: 0.46), followed by $\overline{ERR_{var}} = 0.48$ for $\mathrm{RFLD}_{+\mathrm{rand}}$ (clay: 0.35) to $\overline{ERR_{var}} = 0.46$ for $\mathrm{RFLD}_{+\mathrm{lhs}}$ (clay: 0.33). The explained variance (R^2) for the silt predictions and according to *cv* amounts to $\mathrm{R}^2 = 0.39$ (clay: 0.47) for RFLD, $\mathrm{R}^2 = 0.44$ (clay: 0.49) for $\mathrm{RFLD}_{+\mathrm{rand}}$,

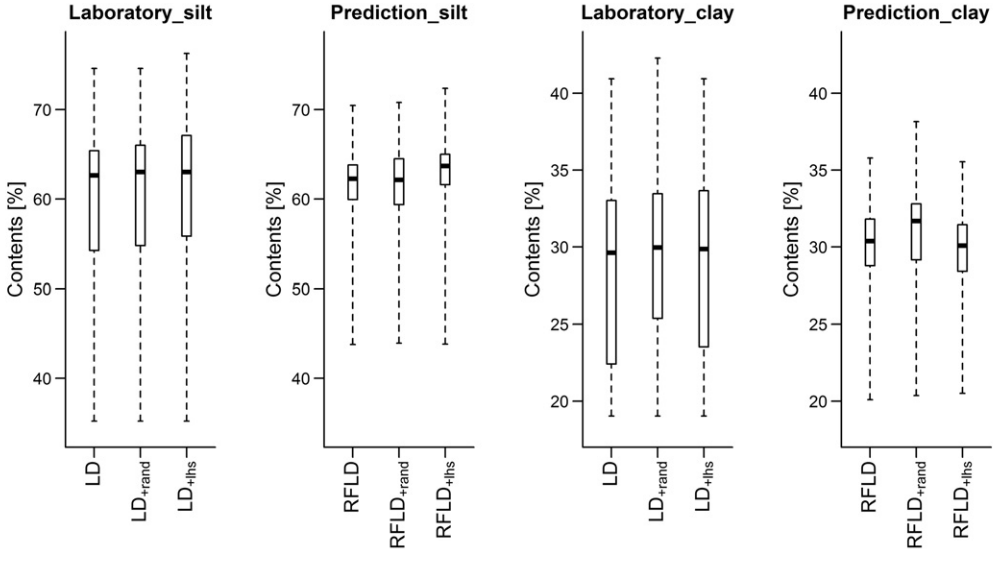

Fig. 3. Boxplots of topsoil silt and clay contents for all calibration sample sets (LD, $\mathrm{LD}_{+\mathrm{rand}}$, $\mathrm{LD}_{+\mathrm{lhs}}$) and associated predictions (RFLD, $\mathrm{RFLD}_{+\mathrm{rand}}$, $\mathrm{RFLD}_{+\mathrm{lhs}}$).

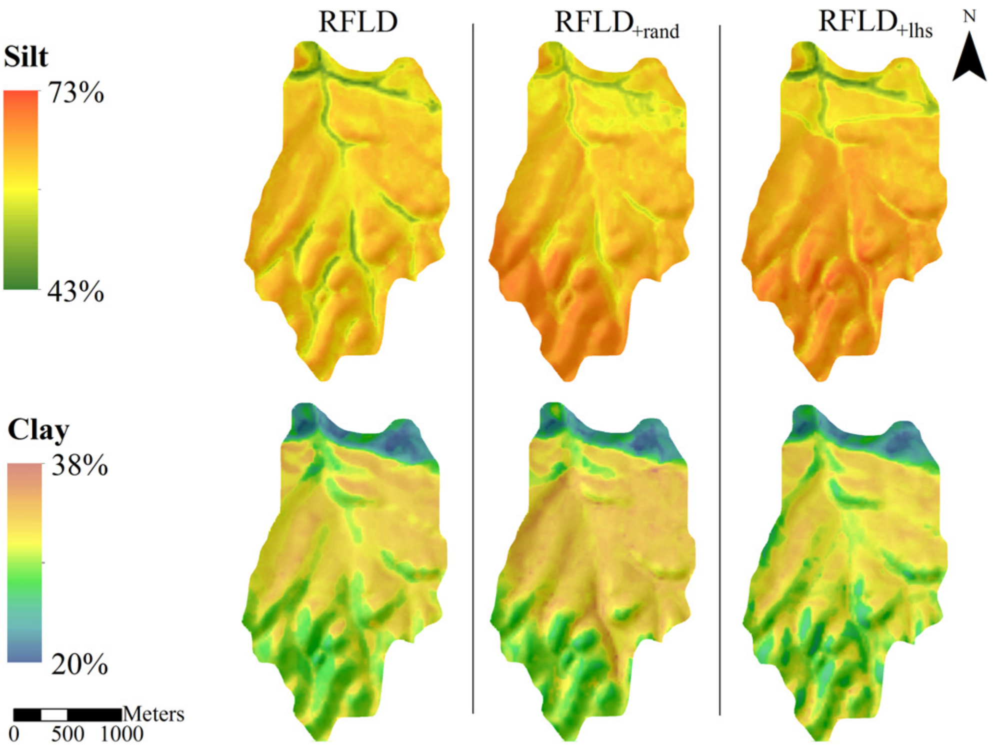

Fig. 4. Predictions of silt and clay contents according to three model approaches (RFLD, $RFDL_{+rand}$, $RFDL_{+lhs}$).

and $R^2 = 0.47$ (clay: 0.55) for $RFLD_{+lhs}$. Referring to *val*, the silt predictions show $R^2 = 0.45$ (clay: 0.44) for RFLD, $R^2 = 0.54$ (clay: 0.47) for $RFLD_{+rand}$, and $R^2 = 0.59$ (clay: 0.56) for $RFLD_{+lhs}$ (Fig. 5).

We compared the uncertainty of silt and clay predictions across the full spatial domain by using the pixel-wise measure err_{var}. The uncertainty maps reflect the aforementioned trend of a quality improvement from RFLD to $RFLD_{+rand}$ and $RFLD_{+lhs}$ in a spatial context. For all prediction approaches and both target soil properties, uncertainty is increased in the topographic depression lines of the study area.

Spatial uncertainties (err_{var}) for $RFLD_{silt}$ range from 0.63 to 0.72. Two coherent areas in the central-western and -eastern study area show decreased uncertainties ($err_{var} < 0.66$), while areas of increased uncertainties ($err_{var} > 0.69$) are distributed evenly all over the study area. Spatial uncertainties of $RFLD_{+rand\_silt}$ range from 0.4 to 0.6 We found a coherent uncertainty hotspot of $err_{var} > 0.55$ in the southern study area and less pronounced areas of increased uncertainties in the north. The spatial uncertainties for $RFLD_{+lhs\_silt}$ range from 0.4 to 0.59. Areas of increased uncertainties ($err_{var} > 0.55$) are mainly located in the north of the study area (Fig. 6).

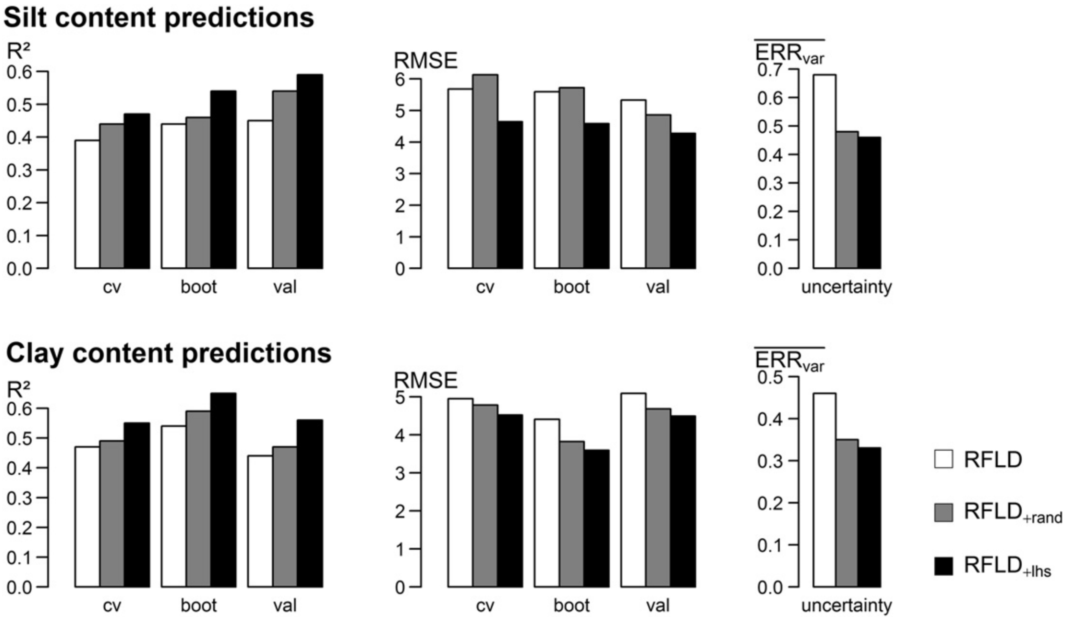

Fig. 5. Accuracies (R^2, RMSE) and uncertainty ($\overline{ERR_{var}}$) for silt and clay content predictions according to three model approaches (RFLD, $RFLD_{+rand}$, $RFLD_{+lhs}$). Cross- (cv; 10-fold), bootstrap- (boot; 10-fold), and external (*val*) validation were used as accuracy estimators.

Spatial uncertainties (err_{var}) for RFLD$_{clay}$ range from 0.41 to 0.5. For RFLD$_{+rand\_clay}$, they range from 0.29 to 0.46, and for RFLD$_{+lhs\_clay}$ from 0.31 to 0.38. We found increased uncertainties for all clay predictions in the north and less pronounced in the topographic depression lines in the central and southern study area (Fig. 6).

4. Discussion

One aim of the presented study was to derive spatially explicit uncertainty in context of a DSM approach using RF. For geostatistical soil property mapping, the kriging variance represents a well-established spatial error estimate (Bourennane et al., 2007; Carré and Girard, 2002; Diodato and Ceccarelli, 2006; Knotters et al., 1995; Qu et al., 2013; Sun et al., 2013). However, this kriging error depends on model-assumption for the variogram, the observed soil data, and their spatial configuration (Brus et al., 2011; Lark and Lapworth, 2012). Moreover, the kriging error relies on the use and limitations of geostatistical methods such as a relatively high sample density and the smoothing of local details in the predictions (Goovaerts, 1999). Malone et al. (2011) proposed a method to quantify spatial uncertainties based on PIs. Primarily, the intervals are derived from the residuals between predicted and observed data. Subsequently, the covariate space is clustered according to similar residuals. Then, a prediction interval is generated for each cluster based on the empirical distribution of residual observations of each cluster. According to the grade of membership to each cluster, a PI is attributed to each pixel in the covariate space. The method was discussed as statistically complex, thus exhibiting limited practicability. However, the PI-uncertainty accounts for all sources of uncertainty, only depending on the residuals derived from the model output and the observed data. Recent approaches in DSM proposed regionalized CIs to derive spatial uncertainty measures. Viscarra Rossel et al. (2014) computed pixel-wise CIs based on repetitively bootstrapping the calibration sample set and predicting for each set. Mulder et al. (2016) generated CIs for each final node of a regression tree model and regionalized the CIs by attributing each pixel to a final node using the tree splitting rules. Heuvelink (2014) proposed to use ordinary kriging for regionalizing residuals from predicted and observed values. Contrary, the presented spatial uncertainty is based on the pixel-wise variance, derived from the results of randomized tree models that in aggregation form the final RF predictions (cf. 2.2). The measure presents a by-product of RF predictions and does not require additional processing steps such as resampling the calibration sample set or spatial modeling. Thus, this method addresses the need for a practical uncertainty measure, while avoiding statistical complexity. However, the method implies the dependency to use a non-parametric ensemble learner, based on randomized simulations, for the spatial predictions. Nevertheless, these models are increasingly applied in context of DSM (Grimm et al., 2008; Heung et al., 2014; Ließ et al., 2012; Schmidt et al., 2014; Stumpf et al., 2016; Viscarra Rossel et al., 2014; Wiesmeier et al., 2011).

A further aim of the presented study was to identify areas relevant to acquire additional soil data. We set up a purposive design based on Latin Hypercube sampling and the spatial uncertainty to augment the initial calibration set (LD). In this context, Clifford et al. (2014) selected additional samples that, in combination with available legacy samples, cover the covariate space and approved the method by a simulation study. Carré et al. (2007) proposed a method to identify locations for additional samples by previously analyzing the distribution of legacy samples in the covariate space. Both methods refer to the covariate space and disregard spatial structures. Contrary, the presented sampling design incorporates the spatial structure of model uncertainties combined with the distribution of legacy samples in the covariate space (Figs. 1, 2).

The goal of the case study was to improve the initial soil property maps of silt and clay contents. In this context, Collard et al. (2014) resampled a legacy soil map for calibrating a regression model and improved the class purity by 10%. Other studies showed an accuracy improvement of 6% to 19% when using DSM approaches to upgrade legacy soil maps (Kempen et al., 2009; Rad et al., 2014; Yang et al., 2011). The presented method was evaluated by comparing three RF DSM protocols, each estimating topsoil silt and clay contents, according to conventional accuracy measures and the proposed uncertainty

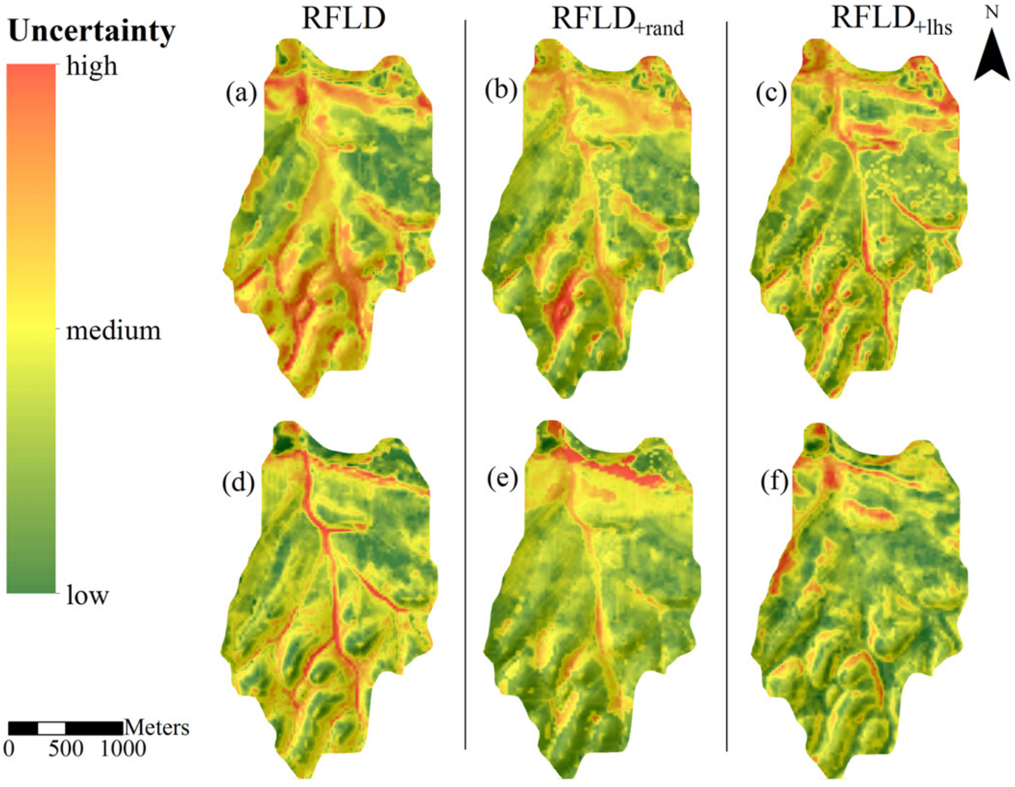

Fig. 6. Spatial uncertainty of silt (a to c) and clay (d to f) content predictions according to three models (RFLD, RFDL$_{+rand}$, RFDL$_{+lhs}$).

measure (cf. 2.4). The DSM protocols solely differ by their calibration sample sets of LD ($n = 30$), and the augmented sets of LD_{+rand} ($n = 90$), and LD_{+lhs} ($n = 90$). With respect to all quality estimation methods, $RFLD_{+lhs}$ performs best. Across the accuracy estimation methods, the results show consistency in terms of a gradual quality improvement (2% to 9%) from RFLD to $RFLD_{+rand}$ and $RFLD_{+lhs}$. According to the uncertainty measure, the improvement from RFLD to $RFLD_{+rand}$ is increased (11% and 29%); while the improvement between $RFLD_{+rand}$ and $RFLD_{+lhs}$ is decreased with 2% across both target soil properties (Fig. 5). The consistency between the results approves the validity of (i) the conventional accuracy measures and (ii) the proposed spatial uncertainty measure. Generally in DSM, accuracies with $R^2 > 70\%$ are unusual, while $R^2 < 50\%$ are common (Malone et al., 2009). The accuracy of the best performing RF DSM protocol ($RFLD_{+lhs}$) shows explained variances of $R^2 = 0.59$ for silt and $R^2 = 0.56$ for clay according to external validation (*val*). The successful application of the spatial uncertainty measure to obtain additional samples and to improve the quality of initial DSM products approves the practicability and validity of the method.

5. Conclusion

With this study we attempt to improve initial DSM approaches for mapping topsoil silt and clay contents in a small heterogeneous catchment in the TGRA, Central China. These maps are based on a non-parametric ensemble model, legacy soil data and continuous terrain attributes. For this standard DSM framework, we propose a method on deriving spatially explicit uncertainty and subsequently on using this information for guided sampling of additional soil data to upgrade the initial DSM protocols.

By means of a comparative uncertainty analysis, we are able to show that additional sampling decreases uncertainties in the final mapping. Preferably, the sampling should be conducted by using a purposive design following our proposed method, which incorporates the spatial structure of uncertainties and the distribution of the soil data in the covariate space.

Furthermore, our method on uncertainty-guided sampling is strongly supposed to be robust since the results outline a consistent trend of quality improvement from the initial maps to a second approach augmented by randomly sampled soil data, up to the proposed approach. This trend remains consistent across several independent statistical measures of accuracy and uncertainty.

The successful and robust improvement of the initial maps showing the spatial distribution of topsoil properties proofs the presented method to be a reliable and transferable tool for DSM. Thus, it is highly expected to support decision-making in context of soil management.

Acknowledgements

This study was conducted within the framework of the Sino-German joint research project YANGTZE GEO. The funding by the German Federal Ministry for Education and Research (BMBF, grant no. 03G 0827A) is highly acknowledged. We particularly thank all students from the China University of Geosciences in Wuhan (P.R. China) that were involved in field campaigns and laboratory analyzes.

References

Behrens, T., Scholten, T., 2006. Digital soil mapping in Germany – a review. J. Plant Nutr. Soil Sci. 169, 434–443.

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

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.X., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 2013, 578–588.

Bishop, T.F.A., McBratney, A.B., Whelan, B.M., 2001. Measuring the quality of digital soil maps using information criteria. Geoderma 103, 95–111.

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

Borra, S., Di Ciaccio, A., 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariate penalty methods. Comput. Stat. Data Anal. 54, 2976–2989.

Bourennane, H., King, D., Couturier, A., Nicoullaud, B., Mary, B., Richard, G., 2007. Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: an empirical comparison of simulation accounting for single attribute and a simulation for secondary information. Ecol. Model. 205, 323–335.

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., Howell, D., Moore, A., Hartemink, E., Kienast-Brown, S. (Eds.): Digital Soil Mapping Bridging Research, Environmental Application, and Operation. Springer, pp. 67–75.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation. Eur. J. Soil Sci. 62, 394–407.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extends: an example from the Murray-Darling basin of Australia. Geoderma 111, 21–44.

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

Carré, F., Girard, M.C., 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.

Collard, F., Kempen, B., Heuvelink, G.B.M., Saby, N.P.A., Richer de Forges, A.C., Lehmann, S., Nehlig, P., Arrouays, D., 2014. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Reg. 1, 21–30.

Diodato, N., Ceccarelli, M., 2006. Computational uncertainty analysis of groundwater recharge in catchment. Eco. Inform. 1, 377–389.

Efron, B., Tibshirani, R., 1993. An introduction to the bootstrap. NY Monographs on Statistics and Applied Probability 57. Chapman and Hall, London.

Everitt, B., 2002. The Cambridge Dictionary of Statistics. Cambridge University Press, Cambridge.

FAO, 2006. Food and Agricultural Organization of the United Nations: Guidelines for Soil Description. (ftp://ftp.fao.org/agl/agll/docs/guidel_soil_descr.pdf (accessed at May 14th, 2014)).

Finke, P.A., 2007. Quality Assessment of Digital Soil Maps: Producers and Users Perspectives. Chapter 39. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping, an introductory perspective. Developments in Soil Science Vol. 31. Elsevier, pp. 523–541.

Good, P.I., 2006. Resampling Methods - A Practical Guide to Data Analysis. third ed. Birkhauser, Boston.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3–26.

Goovaerts, P., Journel, A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. Eur. J. Soil Sci. 46, 397–414.

Grimm, R., Behrens, T., 2010. Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma 155, 154–163.

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.

Grün, B., Leisch, F., 2008. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28, 1–35.

Grunwald, S., 2005. Environmental Soil-landscape Modelling. Geographic Information Technologies and Pedometrics. CRC Press-Taylor & Francis Group, Boca Raton.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modelling approaches. Geoderma 152, 195–207.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12, 993–1001.

Hengl, T., Husnjak, S., 2006. Evaluating adequacy and usability of soil maps in Croatia. Soil Sci. Soc. Am. J. 70, 920–929.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a Random Forest approach. Geoderma 214, 141–154.

Heuvelink, G.B.M., 1996. Identification of field attribute error under different models of spatial variation. Int. J. Geogr. Inf. Sci. 10, 921–935.

Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press-Taylor & Francis Group, Boca Raton, pp. 335–340.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinominal logistic regression approach. Geoderma 151, 311–326.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., De Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76, 2097–2115.

Kim, J.H., 2009. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput. Stat. Data Anal. 53, 3735–3745.

Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246.

Krol, B., 2008. Towards a data quality management framework for digital soil mapping with limited data. Chapter 11. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.L. (Eds.), Digital Soil Mapping With Limited Data. Springer, pp. 136–149.

Kuhn, M., 2009. The caret package: (URL http://cran.r-project.org/web/packages/caret/index.html (accessed at October 2nd, 2014)).

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.

Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. Evaluating large-extent spatial modelling approaches: a case study for soil depth for France. Geoderma Reg. 7, 137–152.

Lagacherie, P., 2008. Digital Soil Mapping: A State of the Art. Chapter 1. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.L. (Eds.), Digital Soil Mapping With Limited Data. Springer, pp. 3–14.

Lagacherie, P., Legros, J., Burrough, P., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65, 283–301.

Lark, R.M., Lapworth, D.J., 2012. Quality measures for soil surveys by lognormal kriging. Geoderma 173, 231–240.

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Eco. Inform. 6, 228–241.

Liaw, A., Wiener, M., 2002. Classification and regression by Random Forest. R News 2, 18–22.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. Geoderma 170, 70–79.

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.

Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626.

McBratney, A.B., Inakwu, O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327.

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

McBratney, A.B., Minasny, B., Viscarra Rossel, R., 2006. Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis. Geoderma 136, 272–278.

McMillan, R.A., 2008. Experiences with Applied DSM: Protocol, Availability, Quality and Capacity Building. Chapter 1. In: Hartemink, A.E., McBratney, A.B., Mendonça-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer, pp. 113–135.

Mendonça-Santos, M.L., Santos, H.G., Dart, R.O., Pares, J.G., 2008. Digital mapping of soil classes in Rio de Janeiro State, Brazil: data, modelling and prediction. Chapter 34. In: Hartemink, A.E., McBratney, A.B., Mendonç-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer, pp. 381–396.

Minasny, B., Bishop, T.F.A., 2008. Analyzing Uncertainty. Chapter 24. In: McKenzie, N., Grundy, M., Webster, R., Ringrose-Voase, A. (Eds.), Guidelines for surveying soil and land resources. CSRO Publishing, pp. 383–393.

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

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264, 301–311.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301–3307 (Geoderma 32, 1378–1388).

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., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16–34.

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.E.C., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random Forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207, 304–318.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9, 181–199.

Qu, M., Li, W., Zhang, C., 2013. Assessing the spatial uncertainty in soil nitrogen mapping through stochastic simulations with categorical land use information. Eco. Inform. 16, 1–9.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (URL http://www.R-project.org (accessed May 5th, 2014)).

Rad, M.R.P., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps by using Random Forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232, 97–106.

Roudier, P., 2011. Clhs: a R package for conditioned Latin hypercube sampling. (URL http://cran.r-project.org/web/packages/clhs/index.html (accessed March 2nd, 2014)).

Ruhe, R.V., 1956. Geomorphic surfaces and the nature of soils. Soil Sci. 82, 441–456.

SAGA GIS, 2011. SAGA GIS (System for Automated Geoscientific Analyses). (Version 2.0.6: URL https://www.saga-gis.org (accessed August 8th, 2014)).

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T., 2010. A method to generate soilscapes from soil maps. J. Plant Nutr. Soil Sci. 173, 163–172.

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

Schönbrodt-Stitt, S., Behrens, T., Schmidt, K., Scholten, T., 2013. Degradation of cultivated bench terraces in the Three Gorges area – field mapping and data mining. Ecol. Indic. 34, 478–493.

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

Stoorvogel, J.J., Kempen, B., Heuvelink, G.B.M., De Bruin, S., 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149, 161–170.

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. 174, 4939–4960.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Wei, X., Scholten, T., 2016. Incorporating legacy soil samples and field operability using a conditioned Latin hypercube sampling design. J. Plant Nutr. Soil Sci. 179, 499–509.

Sun, X.L., Wu, S.C., Wang, H.L., Zhao, Y.G., Zhang, G.L., Man, Y.B., Wong, M.H., 2013. Dealing with spatial outliers and mapping uncertainty for evaluating the effects of urbanization on soil: a case study of soil pH and particle fractions in Hong Kong. Geoderma 195-196, 220–233.

Tutmez, B., Hatipoglu, Z., 2010. Comparing two data driven interpolation methods for modeling nitrate distribution in aquifer. Eco. Inform. 5, 311–315.

Viscarra Rossel, R.A., McBratney, A.B., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Aust. J. Exp. Agric. 38, 765–775.

Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang. Biol. 20, 2953–2970.

Webster, R., Oliver, M.A., 2006. Modeling Spatial Variation of Soil as Random Functions. Chapter 9. In: Grunwald, S. (Ed.)Environmental soil-landscape modeling: Geographic information technologies and Pedometrics. CRC, pp. 241–287.

Wellmann, F.J., 2013. Information theory for correlation analysis and estimation of uncertainty reduction in maps and models. Entropy 15, 1464–1485.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in semi-arid steppe ecosystems. Plant Soil 340, 7–24.

Yalamaç, E., Trapani, A., Akkurt, S., 2014. Sintering and microstructural investigation of gamma-alpha alumina powders. Eng. Sci. Technol. Int. J. 17, 2–7.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Soc. Am. J. 75, 1044–1053.