Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery

Philippe Lagacherie, D. Arrouays, H. Bourennane, Cecile Gomez, L. Nkuba-Kasanda

Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: a numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery

P. Lagacherie[1], D. Arrouays[2], H. Bourennane[3], C. Gomez[1,4], L. Nkuba-Kasanda [1]

1. LISAH, Univ Montpellier, INRAE, IRD, Montpellier SupAgro, Montpellier, France
2. Infosol, INRAE, 45075 Orléans, France
3. UR Sols, INRAE, 45075 Orléans, France
4. Indo-French Cell for Water Sciences, IRD, Indian Institute of Science, Bangalore 560012, India

Abstract

It has long been acknowledged that the soil spatial samplings used as inputs to DSM models are strong drivers – and often limiting factors – of the performances of such models. However, few studies have focused on evaluating this impact and identifying the related spatial sampling characteristics. In this study, a numerical experiment was conducted on this topic using the pseudo values of topsoil clay content obtained from an airborne Visible Near InfraRed-Short Wave InfraRed (Vis-NIR-SWIR) hyperspectral image in the Cap Bon region (Tunisia) as the source of the spatial sampling.

Twelve thousand DSM models were built by running a Random Forest algorithm from soil spatial sampling of different sizes and average spacings (from 200 m to 2000 m) and different spatial distributions (from clustered to regularly distributed), aiming to mimic the various situations encountered when handling legacy data. These DSM models were evaluated with regard to both their prediction performances and their ability to estimate their overall and local uncertainties. Three evaluation methods were applied: a model-based one, a classical model-free one using 25% of the sites removed from the initial soil data, and a reference one using a set of 100,000 independent sites selected by stratified random sampling over the entire region.

The results showed that: 1) While, as expected, the performances of the DSM models increased when the spacing of the sample increased, this increase was diminished for the smallest spacing as soon as 50% of the spatially structured variance was captured by the sampling, 2) Sampling that provided complete and even distributions in the geographical space and had as great spread of the target soil property as possible increased the DSM performances, while complete and even sampling distributions in the covariate space had less impacts, 3) Systematic underestimations of the overall uncertainty of DSM models were observed, that were all the more important that the sparse samplings poorly covered the real distribution of the target soil property and that the dense sampling were unevenly distributed in the geographical space, 4) The local uncertainties were underestimated for sparse sampling and over-estimated for dense sampling while being sensitive to the same sampling characteristics as overall uncertainty.

Such finding have practical outcomes on sampling strategies and DSM model evaluation that are discussed.

Keywords

Uncertainty, sampling methods, spatial distribution indicators, Quantile Random Forest

1. Introduction

Digital Soil Mapping has now emerged as a credible solution for providing soil data to decision-makers acting at global or local scales. Following the GlobalSoilMap specifications (Arrouays et al., 2014), a significant number of countries across the world are now covered by a high-resolution (90-m) grid documented with estimates of soil properties and their associated prediction uncertainties. To obtain this spatial soil information, Digital Soil Mapping models relating the targeted soil properties with exhaustively available covariates were built by running learning algorithms onto spatial sampling of available legacy soil profiles with measured soil properties. Numerous different algorithms have been proposed, mainly depending on the availability and type of soil data (e.g. point data (profiles, augerings), soil maps with various scales and detailed legends) and on the availability of covariates (Minasny and McBratney, 2010).

It has been largely acknowledged that the main limitation of such digital models is the spatial sampling of the legacy measured soil profiles. The average spacing of such soil spatial sampling used in most operational DSM applications have been very large, e.g. 28 km (Hengl et al., 2017) and 5.5 km (Mulder et al., 2016). Furthermore, these soils' spatial samplings have been constituted by the soil profiles of the different existing soil surveys within the mapped area, which means that their locations were not specifically selected for a DSM application and are often clustered at a few sub-areas of the study region, with large sub-areas without any soil information. Consequently, the performances of soil predictions are often severely limited, especially for soil properties whose pattern of variation is largely below the soil profiles' spacing (Vaysse and Lagacherie, 2016; Gomez and Coulouma, 2018). One way to increase DSM performance is therefore to increase the average densities and improve the spatial distribution of the measured soil profiles (Voltz et al., submitted).

Because of their high cost of acquisition, such new locations should be selected with care. For that, the prerequisites are: i) to accurately evaluate the uncertainty of the current soil property maps made from the unevenly distributed legacy measured soil profiles, ii) to determine the expected gain of an increase in the average spatial sampling densities and iii) to identify the spatial sampling characteristics that should be considered in the sampling strategies. These three prerequisites will be examined in this paper and are successively evoked in the following.

The uncertainty evaluation of the soil property maps is an intrinsic part of a DSM application. This was theorized very early (McBratney et al., 2003) and translated into specifications in the GlobalSoilMap project (Arrouays et al., 2014). Although there exist guidelines for the assessment of uncertainty of Digital Soil Maps (Heuvelink, 2014), their practical application is not straightforward, which makes the uncertainty evaluation itself uncertain. The spatial sampling from which uncertainty is evaluated is a crucial point: the probability sampling that is advocated to obtain an unbiased estimate of uncertainty (Brus et al., 2011) cannot be obtained with the already fixed legacy soil data locations. Furthermore, uncertain estimations of the uncertainty of soil maps can be non-negligible even when using independent probability sampling (Kempen et al., 2011, Lagacherie et al., 2019). Lastly, it is uncertain that the set of locations used as evaluation data well describes the real pattern of the soil cover that is to be mapped, especially the short-range variations. For all of these reasons, determining the uncertainty of a DSM map is still an open question.

Although there has been a large consensus in the DSM community that the density of soil data is a clear limiting factor for DSM (Lagacherie, 2008), few experiments showing their impacts on DSM mapping performances exist in the literature. Somarathna et al. (2017) and Wadoux et al. (2019) both observed an increase in the performances of Soil Carbon Mapping as the amount of input data increased, regardless the algorithms used to build the DSM models. However, Somarathna et al. (2017) observed that this increase in performances was lower for the highest sampling densities, which may suggest that, beyond a given threshold of density, it would not be worthwhile to add further new locations. Such a threshold was not observed by Wadoux et al. (2019), which means that it should be highly case-dependent. More case studies should therefore be studied to gain expertise regarding fixing this threshold.

Apart from their average densities, the soil spatial samplings may vary with regard to their spatial distribution across the study areas, which may also strongly affect the prediction performances of the DSM models that use such data as input. This was acknowledged early on (Brus and de Gruijter, 1993), and sampling algorithms for optimizing the completeness and evenness of the input soil locations within the geographical space (Brus et al., 2007) or soil covariate space (Minasny and McBratney, 2006) have been proposed. Similarly, Zhang and Zhu (2019) observed a slight increase in DSM performances when the spatial sampling was optimized with regard to its representativeness of the distribution of the soil covariates. Lark and Marchant (2018) showed that geostatistical predictions can be improved by adding 10% of the closely spaced locations to regular sampling to better represent the short-range spatial structures. Finally, Adamchuk et al. (2011) improved the soil property predictions from soil sensors by developing a sampling strategy that considered the spread among sensor output, the local homogeneity and the physical coverage across an entire field as target spatial sampling properties. All of these works converge towards the idea that sampling strategies can be leveraged to increase DSM performances. However, a recent experiment (Wadoux et al., 2019) showed that this increase could not be obtained by some sampling strategies, which means that the benefit of sampling strategies could be highly case-dependent.

This paper presents a numerical experiment that relates the characteristics of legacy spatial samplings used as input soil data (spatial density and spatial distribution characteristics) with the DSM performance and the accuracy of their ex-ante evaluation. The study used the pattern of the topsoil pseudo clay content derived from airborne Visible Near InfraRed-Short Wave InfraRed (Vis-NIR-SWIR) hyperspectral data acquired over the Cap Bon region (300 km^2, Tunisia) at five-metre resolution (Gomez et al., 2012). This pattern is composed of well-predicted clay values ($R^2 = 0.75$) that are free of visible artefacts and pedologically plausible, which allows it to be considered as a fair representation of the variations of a real soil property across the landscape (Lagacherie et al., 2019). Such a soil dataset provided a quasi-unlimited number of pseudo-measured sites, which enabled the testing of a large number (12,000) of spatial samplings. These spatial samplings were used as input data for a Quantile Random Forest algorithm that produced DSM models whose performances in predicting clay values and the associated accuracy were analysed with regard to the sampling characteristics.

Insert figure 1 here

2. Case study

This case study has already been described in a previous paper (Lagacherie et al., 2019). Large excerpts of this paper are used in the following.

2.1. The study area

The study area is in the Cap Bon region in northern Tunisia (36°24'N to 36°53'N; 10°20'E to 10°58'E), which is 60 km east of Tunis (Figure 1a). This 300-km^2 area includes the Lebna catchment, which is mainly rural (>90%). The Lebna catchment is devoted to the cultivation of cereals in addition to legumes, olive trees, vineyards and natural vegetation for animals. The region is characterized by its rolling hills and elevations between 0 and 226 m. The climate varies from humid to semi-arid, with an inter-annual precipitation of 600 mm and an inter-annual potential evapotranspiration of 1500 mm. The soil pattern of the Lebna catchment is mainly the result of variations in lithology. The variations in the bedrock between Miocene sandstone and Marl cause large variations in the physical and chemical soil properties (Zante et al., 2005). Furthermore, the distance between successive sandstone outcrops decreases significantly as the terrain changes from the ocean to the mountains, which also causes variations in the soil property patterns (Gomez et al., 2012). The soil materials were redistributed laterally along the slopes during the Holocene, which adds to the complexity of the soil pattern. The main soil types are Regosols (IUSS working group WRB, 2006), which are preferentially associated with sandstone outcrops, and Calcic Cambisols and Vertisols, which preferentially formed on marl outcrops and lowlands. The south-eastern region of the study area has a flatter landscape with sandy Pliocene deposits in which Calcosols and Rendzinas prevail.

2.2. Data

2.2.1. Hyperspectral image and derived topsoil clay content predictions

The numerical experiment uses an image of topsoil clay content as input. The topsoil clay contents were derived from a Vis-NIR-SWIR hyperspectral image (Gomez et al., 2012). The approach used to produce the data is summarized below. More details regarding the pre- and post-processing of the hyperspectral image can be found in Gomez et al. (2012).

On November 2, 2010, AISA-Dual airborne-based hyperspectral data were acquired over the study area with a spatial resolution of 5 m. The area of the image is approximately 12 km x 24 km. The AISA-Dual spectrometer measured the reflected radiance via 359 non-contiguous bands covering the 400- to 2450-nm spectral range, with 4.6-nm bandwidths between 400 and 970 nm and 6.5-nm bandwidths between 970 and 2450 nm. The radiance units were converted to reflectance units using ASD spectrometer measurements of uniform surfaces (parking lots, asphalt, concrete) that were collected at the same time during the over flight. Topographical corrections were performed using a digital elevation model built from ASTER data and ground control points.

To isolate the bare soil areas, the study masked pixels with normalized difference vegetation index (NDVI) values greater than an expert-calibrated threshold (0.20). Water and Urban areas were also removed. Finally, the bare soil represented 46.3% of our study area, that is, 5,889,847 measured AISA-Dual 5-m x 5-m pixels.

A Partial Least Square Regression (PLSR) technique (e.g. Tenenhaus, 1998) was then applied to estimate the topsoil clay contents from AISA-DUAL reflectance at each location. The PLSR was calibrated from 129 couples of AISA-DUAL Vis-NIR-SWIR reflectance spectra on bare soil surfaces associated with the topsoil clay content measured on a laboratory soil sample collected from the same bare soil surfaces. Before the PLSR model was built, the reflectance was converted into “absorbance” (log [1/reflectance]). In addition, a Savitzky–Golay filter with second-order polynomial smoothing and window widths of 30 nm (Savitzky and Golay,

1964) and a mean centring and variance scaling was applied to the spectra to reduce noise. The calibrated PLSR model was then validated using a leave-one-out cross-validation that showed successful predictions (R² = 0.75, Gomez et al., 2012, figure 3). The PLSR model was then applied to all bare soil pixels to estimate the topsoil clay content, thus providing the final predicted topsoil clay properties map (Figure 1b). These treatments were implemented in R (Version 1.17) using the signal and pls packages (Mevik and Wehrens, 2007).

2.2.2. Digital Elevation Model and derivatives

A 30-m ASTER digital elevation model (DEM) with specific ortho-rectification and mosaicking was produced for this area. The classical geomorphometric indicators found in the DSM literature were calculated. These include Elevation, Slope, Aspect, plan Curvature, Profile Curvature, Multi-Resolution Valley Bottom Flatness (MRVBF) and four variables describing the aspect: northness, easterness, northwesterness, and northeasterness. All of these indicators served as covariate candidates for representing the relationships between clay content and the relief.

2.2.3. Tunisian agriculture map

A set of layers of the Tunisian agriculture map (STUDI-SCOT-SODETEG, 2001) at 1:20,000 scale was considered as covariates. This includes two soil properties, soil colour classes and textural classes, which were mapped by considering existing detailed soil maps completed by manual interpretation of remote sensing images. The other covariates extracted from the Tunisian agriculture map were the parent material (14 classes) and the land use (7 classes).

2.2.4. Indices from Sentinel-2 images

Three spectral indices were used in this work, based on a Sentinel-2 image acquired on the 2nd of November 2016 over the study area. The acquisition date of Sentinel-2 data was chosen to fit in the period of more extensive bare soils. Sentinel-2 data are composed by multispectral data in 13 bands covering the Visible, Near InfraRed and Short Wave InfraRed spectral domain with spatial resolutions ranging from 10 to 60 m. The Sentinel-2 data acquired on the 2nd of November 2016 was downloaded from the Muscate platform of the French land data centre, called Theia (https://www.theia-land.fr/) in Level 2A, i.e., corrected from atmospheric effects, thanks to the three bands acquired at 60-m spatial resolution (coastal at 443 nm, water vapour at 1375-nm atmospheric correction, coastal at 443 nm, and cirrus at 1376 nm). The three spectral indices used in this work were the normalized difference vegetation index (NDVI), Redness index (RI) and Colour index (CI), calculated following Pouget et al. (1990) and Ghodalizeh et al. (2016):

$$NDVI = \frac{(B8-B4)}{(B8+B4)} \qquad [1]$$

$$CI = \frac{B4^2}{B3^3} \qquad [2]$$

$$RI = \frac{(B4-B3)}{(B4+B3)} \qquad [3]$$

where B8, B4 and B3 are the spectral bands centred at 842 nm, 665 nm, and 490 nm, respectively.

NDVI was selected as a proxy of the vegetation health and biomass that could be in relation with topsoil clay contents through the water and nutriment retention properties. CI and RI were selected as proxies of topsoil colour that could be related to clay content through different parent materials.

3. **Methods**

3.1. DSM modelling

Three criteria were considered to select the learning algorithm used to produce the DSM models: i) the algorithm should be one of the most used and the most efficient among the recent DSM applications, ii) the algorithm had to provide local uncertainty predictions to be able to test its ability to predict the associated uncertainty, and iii) the algorithm should be run without manual intervention and repeated a great number of times in the numerical experiment. Combining these three criteria resulted in selecting the Quantile Regression Forest as the learning algorithm used in this study.

Recent performance testing has found that the Random Forest, from which the Quantile Regression Forest is derived, was among the best algorithm for obtaining predictions of soil properties (Nussbaum et al., 2017), which confirmed a test performed on a wider range of machine learning applications (Caruana et al., 2006). The Quantile Regression Forest has also been demonstrated as efficient for predicting the uncertainty associated with soil property predictions (Vaysse and Lagacherie, 2017). Finally, many authors have adopted this machine learning algorithm for recent large-scale DSM applications (Hengl et al., 2017; Roman Dobarco et al., 2019).

3.1.1. Random Forests and Quantile Regression Forests

This section summarizes the main characteristics of Random Forests and Quantile Regression Forests, using excerpts of Meinshausen (2006). It was already presented in Lagacherie et al. (2019). More details on these two machine learning algorithms are given in the seminal papers by Breiman et al. (2001) and Meinshausen (2006), respectively.

Let *Y* be a real-valued response variable and *X* be a covariate or predictor variable that is likely high-dimensional. A standard goal of statistical analysis is to infer the relationship between *Y* and *X*. Random Forests grow a large (>500) ensemble of trees using *n* independent observations (Y_i, X_i), i = 1, . . ., n. Each tree grows via a recursive partitioning of the source set using one predictor variable X. At each step, the source set is split into two subsets following a test on the value of X. When Y is a quantitative variable, the selected test is the one that minimizes the within-subset variance of Y (Breiman et al., 1984). The recursive partitioning is limited by a stopping rule, and the subsets are produced by the last split being the leaves of the tree. The ensemble of trees is produced by using a random sample of the training data and a random subset of the predictor variables for each tree.

For the regression, the prediction $\widehat{Y}_\theta(x)$ of a single tree ?? of a Random Forest for a new data point x can be represented as the weighted average of the original observations Y_i, i = 1, . . ., n:

$$\widehat{Y}_\theta(x) = \sum_{i=1}^{n} w_{\theta i}(x,\theta) Y_i \qquad [4]$$

where $w_{\theta i}(x,\theta)$ is the weight vector given by a positive constant that is 1 if the observation Y_i is part of the same leaf and 0 otherwise.

By using Random Forests, the prediction is the average prediction of k single trees that were constructed as described above.

$$\widehat{Y_T}(x) = \sum_{i=1}^{n} w_{Ti}(x) Y_i$$

[5]

$$\text{with} \quad w_{Ti}(x) = k^{-1} \sum_{t=1}^{k} w_{\theta i}(x,\theta) \qquad [6]$$

One could assume that the weighted observations deliver a good approximation not only of the conditional mean but also of the full conditional distribution. This assumption is at the heart of the Quantile Regression Forest algorithm, which estimates the conditional distribution function of *Y* given *x* via:

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \qquad [7]$$

From this conditional distribution, it is possible to derive both the predicted value (the mean) and the bound of the 90% prediction interval that predicts the associated uncertainty (the 0.05 and 0.95 quantiles).

Numerous implementations of RF and QRF now exist. Ranger Package (Wright and Ziegler, 2017) was selected because it is a fast implementation of Random Forests that is suitable for multiple model building.

3.1.2. Tuning Random Forest Parameters

The Random Forest Algorithm has several hyperparameters that must be set by the user. Among them, three parameters may significantly impact the results and therefore should be tuned to improve the predictions (Probst et al., 2018): i) the number of observations drawn randomly for each tree, ii) the number of variables drawn randomly for each split and iii) the minimum number of samples that a node must contain. These parameters were tuned using one of the most established tuning strategies, sequential model-based optimization (Jones et al., 1998; Hutter et al., 2011). This tuning algorithm iteratively uses the results of the different already evaluated hyperparameter values and chooses future hyperparameters based on these results. It is implemented in the TuneRanger Package (Probst et al., 2018). After some trials, 100 iterations appeared to be a good compromise that ensured a fairly good convergence towards an optimized solution while being acceptable in terms of computing costs.

Insert figure 2 here

3.2. The numerical experiment

3.2.1. General approach

The workflow of the numerical experiment is presented in Figure 2. First, a master evaluation set of 100,000 locations with pseudo values of clay content was randomly selected from the total set of pixels. This master set served to determine the real performances of the tested models through a set of indicators described further in 3.2.3.

The remaining locations were used to build and evaluate 12,000 models, i.e., 1,000 models for each of the 12 considered sample sizes (100, 200, 300, 400, 500, 600, 800, 1,000, 2,000, 3,000, 5,000, and 10,000 locations), each corresponding to a given average spacing (1732 m, 1225 m, 1000 m, 866 m, 775 m, 707 m, 612 m, 548 m, 387 m, 316 m, 245 m and 173 m) using the following equation:

$$average\ spacing = \sqrt{\frac{total\ area}{size}} \quad . \quad [8]$$

For a given average spacing, the soil inputs of the 1000 models were selected by a specific sampling procedure (see the next section) that mimicked the more or less uneven spatial distributions observed when using legacy data as soil inputs. The soil inputs were then randomly divided into two sets, keeping 75% of the locations to produce DSM models using the QRF algorithm and 25% to obtain a so-called independent evaluation set as currently practised in DSM applications. The DSM models were obtained using QRF according to the method presented above and then evaluated from the independent datasets. The same performance indicators used for the master set were considered to enable statistical comparisons.

Insert figure 3 here

3.2.2. The sampling procedure

The sampling procedure was designed to randomly produce spatial samplings with contrasting degrees of unevenness. The study area was first stratified into 25 geographical strata of equal area using a K-means classification of the locations (Walvoort et al., 2010). Then, the following algorithm was applied:

1) Define a given size of sampling N
2) Select at random an integer P between 1 and 25 (the number of strata) and select at random P strata among the 25
3) Select N/2 locations by a stratified random sampling, using the P strata selected in step 2
4) Complete the spatial sampling by N/2 locations selected at random over the entire study area

For a given size N, steps 2 through 4 were repeated 1,000 times to obtain 1,000 spatial sampling that differed in unevenness thanks to different random selections of P. The procedure was repeated for each selected sample size.

Figure 3 shows four examples of different spatial samplings provided by this procedure.

3.2.3. Model performance indicators

Four model performance indicators were considered. The first two were the Mean Square Error (MSE) and the 90% Prediction Interval Coverage probability (PICP90), which were calculated using the master evaluation set of 100,000 sites.

$$\mathrm{MSE}_{ref} = \frac{1}{n} . (\hat{y}_i - \mathrm{y}_i)^2 \qquad [9]$$

with ŷ$_i$ as the predicted value, y$_i$ as the observed value of Clay, and n = 100,000

$$\mathrm{PICP90} = \frac{1}{n} . \mathrm{card}\,(\{\, \hat{y}_i \geq \mathrm{y_{pred05}} \text{ and } \hat{y}_i \leq \mathrm{y_{pred95}}\})^2 \qquad [10]$$

with $\mathrm{y_{pred05}}$ and $\mathrm{y_{pred95}}$ as the lower and upper bounds of the predicted 90% confidence interval PICP90 expressing the probability that all observed values fit within the 90% prediction limits provided by the DSM model.

The last two indicators aimed to quantify the relative errors of Mean Square Error provided either by the QRF algorithm as a by-product (MSE$_{\mathrm{mod}}$) or by removing 25% of the soil inputs taken as an independent evaluation set (MSE$_{\mathrm{rmv}}$). To calculate these errors, the MSE calculated on the master evaluation (MSE$_{\mathrm{ref}}$) set served as a reference.

$$Err_{mod} = \frac{MSE_{mod} - MSE_{ref}}{MSE_{ref}} \qquad [11]$$

$$Err_{rmv} = \frac{MSE_{rmv} - MSE_{ref}}{MSE_{ref}} \qquad [12]$$

3.2.4. The Spatial sampling indicators

Several spatial sampling indicators quantifying the characteristics of the spatial distributions of the tested spatial sampling were calculated as candidates to explain the performances of the model summarized by the performance indicator presented above. The first three indicators (Coverage Index, Kullback-Leibler Divergence and percentage of out of range) were calculated in the covariate space, in the geographical space and with regard to the target variable (Clay content). The last two (variance and spatially structured-Variance ratio) were calculated for Clay content only. These indicators are detailed below.

Coverage index

The Coverage index (CI) (Gunzburer and Burkdart, 2004) expresses the degree of unevenness of the spatial distribution of the locations of the spatial sampling.

$$CI = \sqrt{\frac{1}{\bar{d}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2\right)} \quad [13]$$

with d_i the distance of site *i* to its nearest neighbour and $\bar{d}$ the mean value of the d_i

Note that, for a regular mesh, CI = 0. Then, a small value of CI means that the spatial sampling has a spatial distribution close to that of a regular grid. The R package DiceDesign (Dupuy et al., 2015) was used to calculate CI.

Kullback-Leibler Divergence (KLD)

The Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951) measures how close two probability distributions are. It was used here to measure the distance between the calibration or evaluation sets and the master validation set. It is a measurement of the representativeness of the calibration and validation sets with regard to the master validation set, which is assumed to represent the real distribution of the different variables across the study region.

For distributions P and Q defined in the same probability space, the Kullback–Leibler divergence between P and Qis defined to be (Wikipedia, https://fr.wikipedia.org/wiki/Divergence_de_Kullback-Leibler)

$$KLD(P||\,Q) = -\sum_{x \in X} P(x)\,.\,log\left(\frac{Q(x)}{P(x)}\right) \qquad [14]$$

For distributions P and Q of continuous random variables, the Kullback–Leibler divergence is

$$KLD(P||Q) = \int_{-\infty}^{\infty} p(x).\log\left(\frac{q(x)}{p(x)}\right) \qquad [15]$$

The KLDs are first calculated for each variable and then averaged to obtain a unique value for the geographical and the covariate spaces.

Percentage of out of range

The percentage of out of range measures the proportion of locations of the study region having values of variables (in the covariates space, the geographical space of Clay values) that are out of the range of those of the sites included in the calibration or evaluation sets. It is also a measurement of representativeness but differs from the former due to its focus on extreme values.

$$\%Out\_of\_Range = \frac{card\left(\left\{y_i < \min_P y \;\&\; y_i > \max_P y\right\}\right)}{card(R)} \qquad [16]$$

with *P* the calibration or evaluation sets and *R* the set of locations in the study region

Clay Variance and percentage of spatially structured variance ratio

Variance of clay is a measure of dispersion within the calibration or validation dataset with regard to clay values.

A spatially structured variance ratio (SSVR) was proposed by Vaysse and Lagacherie (2015) as the complement to 1 of the nugget-to-sill ratio (Kerry and Oliver, 2008). It indicates the proportion of the spatially structured variance that is captured by the model.

SSVR = variance – nugget / variance.   [17]

Because it was impossible to fit variograms on 12,000 trials, the nugget could not be calculated directly. It was therefore approximated by computing the semi-variances at lags centred on the average spacing ([average spacing – 100 metres, average spacing + 100 metres])

Insert figure 4 here

4. Results

4.1. Impact of average spacing

Figures 4a through 4d show the evolution of the different indicators of DSM performances with the average spacing (Equation 8). The Mean Square error on the predicted Clay value (MSE_{ref}) covered a large range of values across the 12,000 models (Figure 4a). These values increased regularly with the average spacing: from MSE = 7,911 $g^2.kg^{-2}$ (68% of explained variance) to MSE = 18,882 g^2 kg^{-2} (22% of explained variance). The amount of variation of MSE_{ref} for a given size of sampling also increased regularly with the average spacing.

PICP90 exhibited a positive bias (overestimated uncertainty) with regard to the expected 90% value for the smallest average spacings (below 612 m) and a negative one (underestimated uncertainty) beyond this threshold (Figure 4b). However, the errors were only important (more than 1%) for the largest average spacing (- 2.5% for 1732 m) and for the smallest ones (between 1.2 and 1.7 for average spacing at and below 316 m). Apart from the influence of the average sampling variations, great variabilities of performances for

PICP90 estimations were observed for the largest average spacings (see the bars of Figure 4b).

The bagging procedure of the QRF algorithm and the external evaluation using 25% of removed sites both had a negative bias (Figures 4c and 4d) regardless of the average spacing, which revealed systematic underestimations of the overall uncertainty of the DSM models (e.g., -17% and -10% respectively for 1732-m spacing). This bias seemed not as closely related to the average spacing of sites as was observed previously for the other indicators, although slight decreases could be observed beyond 866-m spacing. The most important variations were observed within each sampling size, as shown by the large bars of Figures 4c and 4d.

Insert figure 5 here

4.2. The impacts of the spatial distributions of sites

The matrices of Figures 5a through 5d show, for each average spacing (the columns of the matrices), the correlations between the spatial sampling indicators (lines of the matrices) and the indicators of DSM performance (one matrix per indicator). The last lines of the matrices show the coefficients of determination of the stepwise regressions between the indicator of performance of interest and the set of indicators of spatial sampling, which allowed the strength of the relation between these two types of indicators to be appreciated.

Whatever the considered DSM performance indicators, the correlation coefficients were highly variable with regard to the spatial sampling indicators and the average spacing (between 0.00 to 0.91). These coefficients tended to increase with the decrease of average spacing, which is summarized by the increase in the strength of the relation between the indicators of performance and the spatial sampling indicators (last lines of the matrices). However, some noticeable exceptions occurred for some relations between performance indicators and spatial sampling indicators.

MSE$_{\text{ref}}$ (Figure 5a), which expressed the ability of the DSM models to predict the correct value of clay content, was strongly positively correlated with the *coverage_index* and *Kullback-Leibler divergence* calculated in the geographical space, which means that the performances of the DSM models were better for evenly geographically distributed and representative spatial samplings. These correlations increased greatly when average spacing decreased (until - 0.87 for the two indicators) but were already substantial for the largest spacing (-0.42 and -0.38). For all of the tested average spacings, MSE$_{\text{ref}}$ was also moderately correlated with *Variance_Clay* and *%-out-of-range_Clay* (between 0.27 and 0.40 and between – 0.27 and -0.56, respectively). This means that the performances of the DSM models tended to increase as the clay values included in the spatial sampling were largely dispersed, and this well covered the range of clay values of the study region. The three indicators calculated in the covariate space were only significantly correlated with MSE$_{\text{ref}}$ for the smallest average spacings. Furthermore, these correlations were always smaller than those obtained by the same indicators calculated in the geographical space. Finally, the spatially structured variance ratio (SSVR) exhibited moderate negative correlations for intermediate values of average spacing (between -0.27 and -0.42), which means that, for these values, the more the spatially structured variability (especially the short range one) was captured by the spatial sampling, the better the performances were.

PICP90, Err_{mod} and Err_{rmv} (Figures 5b, 5c and 5d), which all expressed the ability to predict the uncertainty associated with the predicted values given by the DSM models, behaved similarly to each other with regard to the correlations with the spatial sampling indicators, with however stronger overall correlations for Err_{rmv} than for PICP90 and for Err_{mod} than for Err_{rmv} (see the stepwise regression coefficient, the last lines of the matrices in Figures 5b, 5c and 5d). Contrary to MSE$_{ref}$, clear differences of correlation rankings were observed between the smallest and the largest average spacings. As far as the former are concerned, PICP90, Err_{mod} and Err_{rmv} were predominantly correlated with *Variance_Clay* and *%-out-of-range_Clay* (between 0.46 and 0.76 and between -0.38 and -0.62, respectively, for average spacing larger than or equal to 866 m). This means that the uncertainty was much better predicted when the clay values included in the spatial sampling were highly dispersed and covered well the range of clay values of the study region. At the smallest spacings, the strongest correlations were observed with the *coverage_index* and *Kullback-Leibler divergence* calculated in the geographical space (between -0.49 and -0.88 and between -0.48 and - 0.85, respectively, for average spacing smaller than or equal to 387 m), which means that evenly geographically distributed and representative spatial samplings enabled an accurate prediction of the DSM model uncertainty. The spatial sampling indicators calculated on the covariate space only exhibited substantial correlations with PICP90, Err_{mod} and Err_{rmv} at the smallest spacings. Finally, SSVR exhibited substantial correlations only for PICP90 (between 0.33 and 0.45).

Insert figure 6 here

5. Discussion

5.1. The impact of the average spacing

All of the results confirmed that the average spacing, related with the size of the calibration data sets used as input for the DSM approach, strongly impacted the results of a DSM approach. As already observed by Somarathna et al. (2017) and Wadoux et al. (2019), we observed (Figure 4a) a clear decrease of prediction errors (MSE_{ref}) when the average spacing is decreasing. It should be noticed that, thanks to the use of pseudo-values of clay content given by the hyperspectral image, we explored a more complete range of average spacing, which allowed us to analyse a larger range of model performances (from 22% to 68% of explained variance) that covered fairly well the ones cited in the literature. Figure 4a revealed that the decrease of prediction errors with average spacing was not linear. By substituting the average spacing X axis of Figure 4a by the spatially structured variance ratio (SSVR, see section 3.2.4), a new insight into this average proportion of clay variance was revealed (Figure 6): for the largest average spacing, that captured the least spatially structured variance (<= 800 sites, average spacing = 707 m), the average increase in performances was perfectly linear, whereas further increases of this ratio provided gains that were smaller and smaller than the previously observed linear trend. Therefore, this observed threshold separated two contrasting situations: below the threshold of average spacing of 707 m, the spacing of sampled sites was the only limiting factor, while beyond the threshold, other limiting factors, such as the precision of the covariates, also played a role in the quality of the results. This contrasting behaviour could explain why contradictory results have been obtained recently regarding the impact of the spatially structured variance ratio on DSM results, as observed by Vaysse and Lagacherie (2015) and not observed by Nussbaum et al. (2017). It may also explain why improving a covariate dataset often do not significantly improve the DSM products when overly sparse test datasets are used to calibrate the DSM model (Samuel-Rosa et al., 2015; Loiseau et al., 2019).

Apart from its influence on the prediction error, the average spacing of sites also had an impact on the estimations of the associated uncertainty, which, to our knowledge, has not been observed before. Decreasing the average spacing reduced the underestimation of the overall uncertainty provided either by the bagging procedure of the Random Forest (Figure 4c) or the model-free evaluation process (Figure 4d), although these reductions were not as clear and regular as the prediction error because larger variabilities of results were observed within each tested spacing (see the error bars in Figures 4c and 4d). A more complex behaviour was observed for the local estimation of the confidence interval by the model tested by PICP90 (Figure 4b). For the largest spacings, PICP90 estimates converged towards the nominal value of the confidence interval (90%) as the spacing decreased, whereas for the smallest spacings, PICP90 moved away from this nominal value. This latter unexpected result could be interpreted as the inclusion of outliers as the spacing decreased, which could perturb the estimates of the confidence interval bounds. It must be noticed that, as for the estimations of the overall uncertainty evoked before, a large variability in estimating PICP90 was observed within each tested spacing.

Finally, a final effect of the spatial sampling size is that it changes the amount and the drivers of the variations of performances observed within each sampling size. This will be developed in the next section.

5.2. The impact of the distribution of sites over the study region

The bars on Figures 4a through 4d show that the average spacing is not the only driver of DSM performance, especially with regard to the ability of DSM approaches to estimate overall (Figures 4 c and 4 d) and local uncertainties (Figure 4b). The matrices of Figure 5 confirmed many of the underlying hypotheses of the sampling strategies that have been proposed in the literature while providing new insights on the relation between spatial sampling and uncertainty estimation and nuancing the importance of some sampling characteristics according to the size of the spatial sampling.

From the matrices of Figure 5, it clearly appeared that the regularity of sampling and the representativeness in the geographical space improved the DSM results whatever the size of the spatial sampling and the considered indicators of performances. Therefore, the legacy soil data that are often characterized by both under-sampled and over-sampled sub-regions should be ideally completed using sampling strategies that could mitigate this irregularity of sampling in the geographical space (Brus et al., 2011; Adamchuk et al., 2011). This would require harmonizing the legacy and the new dataset techniques for removing biases caused by differences of dates, field protocols and laboratory methods (Baume et al., 2011; Ciampalini et al., 2013). An alternative to adding samples should be to better take into consideration in the DSM modelling the perturbing effects of the clusters of sites. This could be done by assigning different weights to the input sites according to their degree of remoteness (Bel et al., 2009), applying resampling techniques (Richer-de-Forges et al., 2017; Taghizadeh-Merjadhi et al., 2020) or restricting the predictions inferred from each cluster of sites to representative areas corresponding to well-identified and well-mapped soil systems (Lagacherie et al., 2001).

The spread of the spatial sampling with regard to the values of the target soil property (%out-of-range and variance of Clay in Figure 5) also seemed crucial for improving both the predictions of the soil property and the predictions of the associated overall and local uncertainties. However, correcting the existing legacy sample with regard to this characteristic is an uneasy task because it requires additional knowledge to anticipate the locations of the extreme values of the targeted soil property that should be preferentially sampled. A local pedological knowledge or a proxy of the soil property of interest (Adamchuk et al., 2011) could be mobilized for that.

It is interesting to note that the sampling characteristics that involved the soil covariates only had an impact on the results for the smallest spacing of spatial sampling. In these cases, this impact was increased by strong relationships established between the covariates and the predicted soil property, whereas for the largest spacing, these relationships were too weak, even if the most related covariates were selected (results not shown in this paper). This means that the strategies of sampling based on the regularity of coverage of the spatial sampling in the covariate spaces (Minasny and Mc Bratney, 2006; Carré et al., 2007; Zhang and Zhu, 2019) could not be effective for correcting overly sparse legacy datasets. Conversely, for large datasets for which more covariates were involved in the model, a fair distribution of the covariates values would be required. However, the coverages and KLD indices in the covariate space and the geographical one were found to be highly correlated for the smallest average spacing (r > 0.90 for average sampling >= 1225 m), which means that taking into account the covariate space would be of little interest if the regularity of sampling in the geographical space is already ensured. However, this result would not hold in particular pedological contexts characterized by small inclusions of land with contrasted values of covariates that could be missed by regular samplings in the geographical space.

Finally, the stepwise regression coefficients of determination given in the last lines of the matrices of Figure 4 clearly showed that the selected sampling characteristics could not alone explain the variations of performances that were observed across the 12,000 trials. This was particularly true for the sparsest spatial sampling when we considered the mean square error on predicted value (Figure 5a) or the biases of estimation of the latter (Figure

5c or 5d). The reverse tendency was observed for the error on PICP, for which the smallest average spacing of sites obtained the lowest coefficient of determination.

5.3. Uncertainty estimation biases

The two tested procedures of overall uncertainty estimation - the model-based and the model-free ones - exhibited non-negligible biases on uncertainty of clay content predictions (Figures 4c and 4d). Although the two procedures could be considered as intrinsically unbiased, some specific characteristics of the legacy spatial sampling to which they were applied were responsible of these biases. Although it could be observed (Figure 4) that the sparse samplings were more prone to bias than the dense ones, the average spacing did not seem to be a first-order driver because much more variability occurs within a given sampling size (see the bars of Figures 4c and 4d).

The correlations matrices (Figures 5c and 5d) provided some insights on the causes of such biases. As far as the largest spacing were concerned, biases were all the more great that the sampling underestimated the real variations of clay content and thus left aside their extreme values. This observation can be related with the general difficulty of the inference models, such as Random Forest, to predict values that are out of the range of their learning sample (Conn et al., 2014). Alternately, overly clustered datasets (see examples in Figure 3, left column) resulted in selecting evaluation sites that could be too close from the calibration sites for satisfying the condition of independence , which may induce underestimations of the prediction errors. Therefore, to estimate the overall uncertainty as well, it is important to mitigate the perturbing effects of the clusters of sites techniques cited above.

Finally, it is worth noting that the estimations of the local uncertainty through a confidence interval calculated by the QRF algorithm seemed to be more robust than the estimations of the overall uncertainty (Figure 3b). Only the largest spacing (1732 m) and, to a lesser extent, smallest ones gave unsatisfactory results. This confirmed the results obtained by Vaysse and Lagacherie (2017) using the same algorithm.

5.4. Limitations and open questions.

Although some clear and coherent tendencies could be retrieved from the results of this numerical experiment (see above), some open questions remain. First, the characteristics of the spatial sampling that were considered in this paper did not explain the entire variability of DSM performances. The weak statistical relations observed in Figure 5 for the largest spacing suggest that some hidden factors should be evoked. Among others, we hypothesize that the random process used for optimizing the hyperparameters of the Random Forest generated a noise on DSM performances, the best possible combination of parameters not always being reached because of local optimal solutions, especially when the size of the learning sample is small. This hypothesis is supported by the fact that the average variabilities of the optimal QRF parameters provided by the optimization process decreased as the size of the sampling increased (average Coefficient of Variation from 35% to 21%).

Second, although a large range of soil sampling spacings were explored in this case study, the size of the study areas limited the testing of the sparser soil datasets that fed the DSM applications conducted at national (e.g. Mulder et al., 2016), continental (Ballabio et al., 2016) or global scale (Hengl et al., 2017). Whether or not the trends exhibited in Figures 4 and 5 can be extrapolated to these applications remains an open question. With the next availability of hyperspectral VIS-NIR-SWIR satellite data (such as the French HYPerspectral X Imagery –HYPXIM-, Briottet et al., 2013; the Spaceborne Hyperspectral Applicative Land and Ocean Mission –SHALOM-, Bussoletti, 2012; the German Environmental Mapping and

Analysis Program –EnMAP-, Stuffler et al., 2007; Steinberg et al., 2016 and the Hyperspectral Infrared Imager -HyspIRI-, Lee et al., 2015), it could be envisaged to reproduce the same numerical experiment at a wider extent with, however, a loose spatial resolution. In the absence of such an experiment, the results obtained for the largest average spacing considered in this numerical experiment (1732 m) should help in orienting the future design of DSM approaches at these largest extents.

6. Conclusions

The main lessons of the numerical experiment are as follows

- User and producers of DSM products should be aware that the current methods of evaluation tend to underestimate the overall uncertainty, especially for sparse and unevenly distributed soil sampling
- Although decreasing the average spacing of soil inputs always brings improvements of DSM performances, one should be aware that, beyond a given threshold of average spacing, the improvement would need also to collect better soil covariates.
- The spatial distributions of the legacy data and the sampling strategies for correcting these distributions play a key role in reaching the best DSM performances. Sampling strategies that provide complete and even distributions in the geographical space and have as great a spread of the target soil property as possible should be privileged.
- Some hidden sampling characteristics that were not considered in this experiment seem to play a significant role, especially for sparse sampling. More research is required for identifying these characteristics.

7. Acknowledgements

This research was conducted within the "Centre d'Expertise Scientifique Cartographie Numérique des sols" granted by the CNES-TOSCA programme.

Dominique Arrouays is coordinator and Philippe Lagacherie and Hocine Bourennane are collaborators of the GLADSOILMAP Consortium supported by the LE STUDIUM Loire Valley Institute for Advanced Research Studies

8. References

Adamchuk, V, Viscarra, R.A., Marx, D.B., Samal, A.K., 2011. Using targeted sampling to process multivariate soil sensing data. Geoderma 163, 63–73.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014. The GlobalSoilMap project specifications, in: Arrouays, D., McKenzie, N.J., Hempel, J.W., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the global Spatial soil information system. CRC press & Taylor & Francis group, Boca Raton, USA, pp 1-12

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.

Baume, O., Skien, J.O., Heuvelink, G.B.M., Pebesma, E.J., 2011. A geostatistical approach to data harmonization — application to radioactivity exposure data. International Journal of Applied Earth Observation and Geoinformation 13 (3), 409–419.

Bel, L., Allard, D., Laurent, J.M., Cheddadi, R. and Bar-Hen, A.,2009. CART algorithm for spatial data: Application to environmental and ecological data. Computational Statistics and Data Analysis 53, 3082-3093. doi:10.1016/j.csda.2008.09.012.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Briottet, X., Marion, R., Carrere, V., Jacquemoud, S., Bourguignon, A., Chami, M., et al.,2013. HYPXIM: HYPXIM: A second generation high spatial resolution hyperspectral satellite for dual applications. 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), June 2013, Gainesville, Florida, USA.

Brus, D. J., De Gruijter, J. J., 1993. Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science. Environmetrics 4 (2), 123–152.

Brus, D. J., De Gruijter, J. J., Van Groenigen, J. W., 2007. Designing spatial coverage samples using the k-means clustering algorithm. Developments in Soil Science 31, 183–192.

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

Bussoletti, E.,2012. Space observations for agriculture and food support. Inter-Agency Meeting on Outer Space Activities: 2012. Thirty-second session, 7–9 March 2012. Rome, Italy.

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.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. Proc. 23rd Int. Conf. Mach. Learn. - ICML '06 161–168.

Ciampalini, R., Lagacherie, P., Gomez, C., Grünberger, O., Hamrouni, M.H., Mekki Insaf, Richard, A., 2013. Detecting, correcting and interpreting the biases of measured soil profile data: A case study in the cap bon region (Tunisia). Geoderma 192, 68-76.

Dupuy, D., Helbert, C., Franco, J.,2015. DiceDesign and DiceEval: Two R Packages for Design and Analysis of Computer Experiments. Journal of Statistical Software 65(11), 1-38. URL http://www.jstatsoft.org/v65/i11/.

Gomez, C., Lagacherie P., Bacha, S. 2012. Using a VNIR/SWIR hyperspectral image to map topsoil properties over bare soil surfaces in the Cap Bon region (Tunisia). In "Digital Soil Assessments and Beyond" Minasny B., Malone B.P., McBratney A.B. (Ed.). Springer, 387-392.

Gomez, C., Coulouma, G., 2018. Importance of the spatial extent for using soil properties estimated by laboratory VNIR/SWIR spectroscopy: Examples of the clay and calcium carbonate content. Geoderma 330, 244–253.

Gunzburer M., Burkdart J., 2004. Uniformity measures for point samples in hypercubes

https://people.sc.fsu.edu/%7Ejburkardt/. Consulted on 2019 august 21rd

Hengl, T., De Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, 1–40.

Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. In: Arrouays, D., McKenzie, N.J., Hempel, J.W., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the global Spatial soil information system. CRC press & Taylor & Francis group, Boca Raton, USA, pp 327–332.

Hutter, F., Hoos, H. H., Leyton-Brown, K., 2011. Sequential model-based optimization for general algorithm configuration, 507–523. Berlin, Heidelberg: Springer Berlin Heidelberg.

IUSS (International Union of Soil Scientists) Working Group WRB, 2006. World Reference Base for Soil Resources 2006, World Soil Resources Report No. 103. Food and Agriculture Organization of the United Nations, Rome, Italy.

Jones, D. R., Schonlau, M., Welch, W. J.,1998. Efficient global optimization of expensive black-boxfunctions. Journal of Global optimization 13, 455–492.

Kempen, B., Brus, D.J., Stoorvogel, J.J., 2011. Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. Geoderma 162, 107–123.

Kerry, R., Oliver, Æ.M.A., 2008. Determining nugget : sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Precis. Agric., 33–56.

Kullback, S., Leibler, R., 1951. On information and sufficiency », *Annals of Mathematical Statistics*22, 79-86.

Lagacherie, P., 2008. Digital soil mapping: A state of the art, in Digital Soil Mapping with limited soil data, Hartemink A., McBratney, A.B. and Mendonça-Santos, L. (eds). Springer. Pages 3-14

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., Barthès, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101 (3-4), 105-118.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N.P.A., 2019. How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 337, 1320–1328.

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

Lee, C.M., Cable, M.L., Hook, S.J., Green, R.O., Ustin, S.L., Mandl, D.J., Middleton E.M., 2015. "An introduction to the NASA Hyperspectral InfraRed Imager (HyspIRI) mission and preparatory activities", Remote Sensing of Environment 167, 6–19

Loiseau, T., Chen, S., Mulder, V.L., Dobarco, M.R., Lehmann, S., Bourennane, H., Saby, N.P.A., Martin, M.P., Vaudour, E., Gomez, C., 2019. Satellite data integration for soil clay content modelling at a national scale. Int J Appl Earth Obs Geoinf. 82.

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

Meinshausen, N., 2006. Quantile Regression Forests. J. of Machine Learning Res. 7, 983–999.

Mevik, B.-H., Wehrens, R., 2007. The pls Package: Principal Component and Partial Least Squares Regression in R; Journal of Statistical Software 18(2), 1—24

Minasny, B., McBratney, A. B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32 (9), 1378–1388.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping—Bridging Research, Environmental Application, and Operation. Progress in Soil Science, Springer, Dordrecht.

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

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. Sci. Total Environ. 573, 1352–1369.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2017. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL Discuss. 1–32.

Pouget, M., Madeira, J., Le Floch, E., Kamal, S., 1990. Caractéristiques spectrales des surfaces sableuses de la région Nord-Ouest de L'Égypte: Application aux données satellitaires SPOT. In: 2eme Journées de Télédétection: Caractérisation et suivi des milieux terrestres en régions arides et tropicales. 4–6/12/1990. ORSTOM, Collection Colloques et Séminaires, Paris, France.

Probst, P., Wright, M., Boulesteix, A., 2018. Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1–19.

Richer-de-Forges, A.C., Saby, N.P.A., Mulder, V.L., Laroche, B., Arrouays, D., 2017. Probability mapping of iron pan presence in sandy podzols in South-West France, using digital soil mapping. Geoderma Reg. 9, 39–46.

Roman Dobarco, M., Bourennane, H., Arrouays, D., Saby, N.P.A., Cousin, I., Martin, M.P., 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma 344, 14-30.

Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243-244, 214–227.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36(8), 1627–1639

Somarathna, P.D.S.N., Minasny, B., Malone, B.P., 2017. More Data or a Better Model? Figuring what Matters Most for the Spatial Prediction of Soil Carbon. Soil Sci. Soc. Am. J. 81, 1413–1426.

Steinberg, A., Chabrillat, S., Stevens, A., Segl, K. Foerster, S. 2016. "Prediction of Common Surface Soil Properties Based on Vis-NIR Airborne and Simulated EnMAP Imaging Spectroscopy Data: Prediction Accuracy and Influence of Spatial Resolution", Remote Sensing, 8, 613; doi:10.3390/rs8070613

Stuffler, T., Kaufmann, H., Hofer, S., Förster, K. -P., Schreier, G., Müller, A., et al., 2007. The EnMAP hyperspectral imager — An advanced optical payload for future applications in Earth observation programmes. Acta Astronautica 61(1–6), 115–120.

STUDI-SCOT-SODETEG, 2001. Etude des cartes agricoles régionales. Ministère de l'agriculture, de l'environnement et des ressources hydrauliques. Report and map.

Taghizadeh-Mehrjardi, R., Schmidt, K., , Eftekhari, K, Behrens, T, Jamshidi, M, Davatgaar, N, Toomanian, N, Scholten, T. (in press). Synthetic resampling strategies and machine learning for digital soil mapping in Iran. European J. of Soil Sciences.

Tenenhaus, M., 1998. La régression PLS. Editions Technip, Paris. 254 pp

Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg. 4, 20-30

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55-64

Voltz, M., Arrouays, D., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B, Richer-de-Forges, Sauter, J., Schnebelen, N., 2020. Disseminating Digital Soil Mapping in national soil mapping programmes: a prospective analysis in France. Submitted in Geoderma Regional

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

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

Wright, M.N., Ziegler, A., 2017. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C ++ and R 77.

Zhang, G., Zhu, A.X., 2019. A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil.pdf. Geoderma 351, 140–163.

Figure 1: Location of the study area (a) and the spatial pattern of pseudo values of topsoil clay content (b) (after Lagacherie et al, 2009)

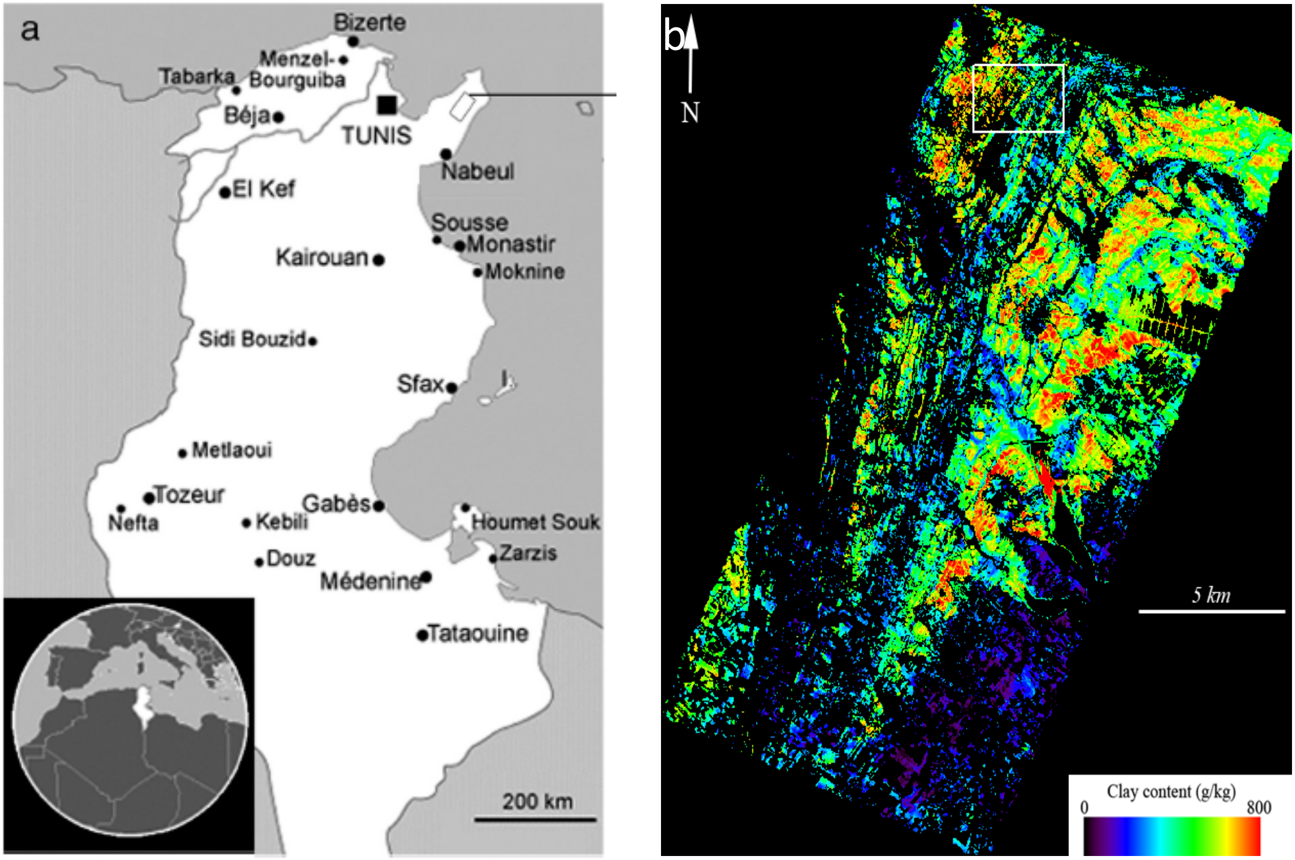
Figure 2: General approach for the numerical experiment

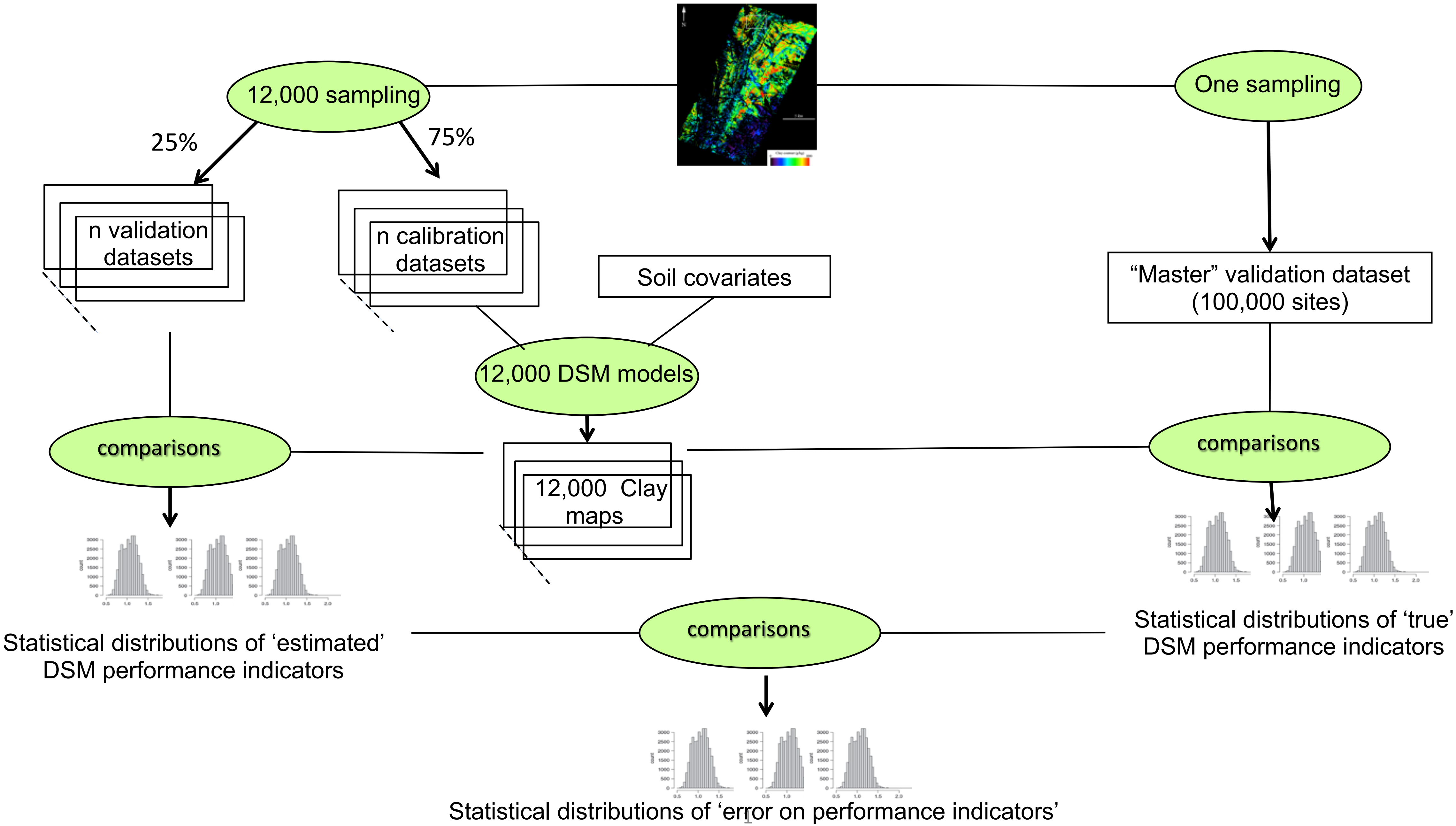
Figure 3: Examples of tested spatial sampling in the numerical experiment. Left column: Three clusters, right column: Fifteen clusters, top row : 75 calibration sites and 25 evaluation sites, bottom row : 750 calibration sites and 250 evaluation sites. Calibration sites are in black. Evaluation sites are in red.

Figure 4: Evaluation of DSM models (quantile Random Forests) using different size of soil input data (size is expressed by spacing):
a) Mean Square error on predicted Clay value (MSE_{ref}), (in g^2/kg^2).The green line is the total variance of Clay content over the study area
b) 90% Prediction Interval Coverage Index (PICP90) (in %). The green line is the expected value of 90%.
c) error on the QRF based estimation of MSEref (in % MSEref). The green line is 0 (no error)
d) error on the model-free estimation of MSEref (removing 25% of the soil inputs for validation) (in % MSEref),). The green line is 0 (no error).
Red dots are averaged values per spacing and bars are +- the standard deviations (1000 models per spacing)

Figure 5: Correlation coefficients (CC) between the indicators of performances and the indicators of spatial distribution of sampling for different sizes of spatial sampling (1000 simulations per size): a) Mean Square error (MSE_{ref}), b) PICP90 c) error on MSE estimated by the random forest bagging procedure d) error on MSE estimated by removing 25% of sample.

Figure 6 : Evolution of the average performances of predictions (mean MSEref) with the spatially-structured variance ratio (SSVR). The red line is the linear regression using the seven spatial sampling (out of 12) having the smallest values of SSVR.


a
Bizerte
Menzel-
Bourguiba
Tabarka
Béja
TUNIS
Nabeul
El Kef
Sousse
Monastir
Kairouan
Moknine
Sidi Bouzid
Sfax
Metlaoui
Tozeur
Nefta
Gabès
Kebili
Houmet Souk
Zarzis
Douz
Médenine
Tataouine
200 km
b
N
5 km
Clay content (g/kg)
0
800



12,000 sampling

25%

75%

n validation datasets

n calibration datasets

Soil covariates

One sampling

"Master" validation dataset (100,000 sites)

12,000 DSM models

comparisons

12,000 Clay maps

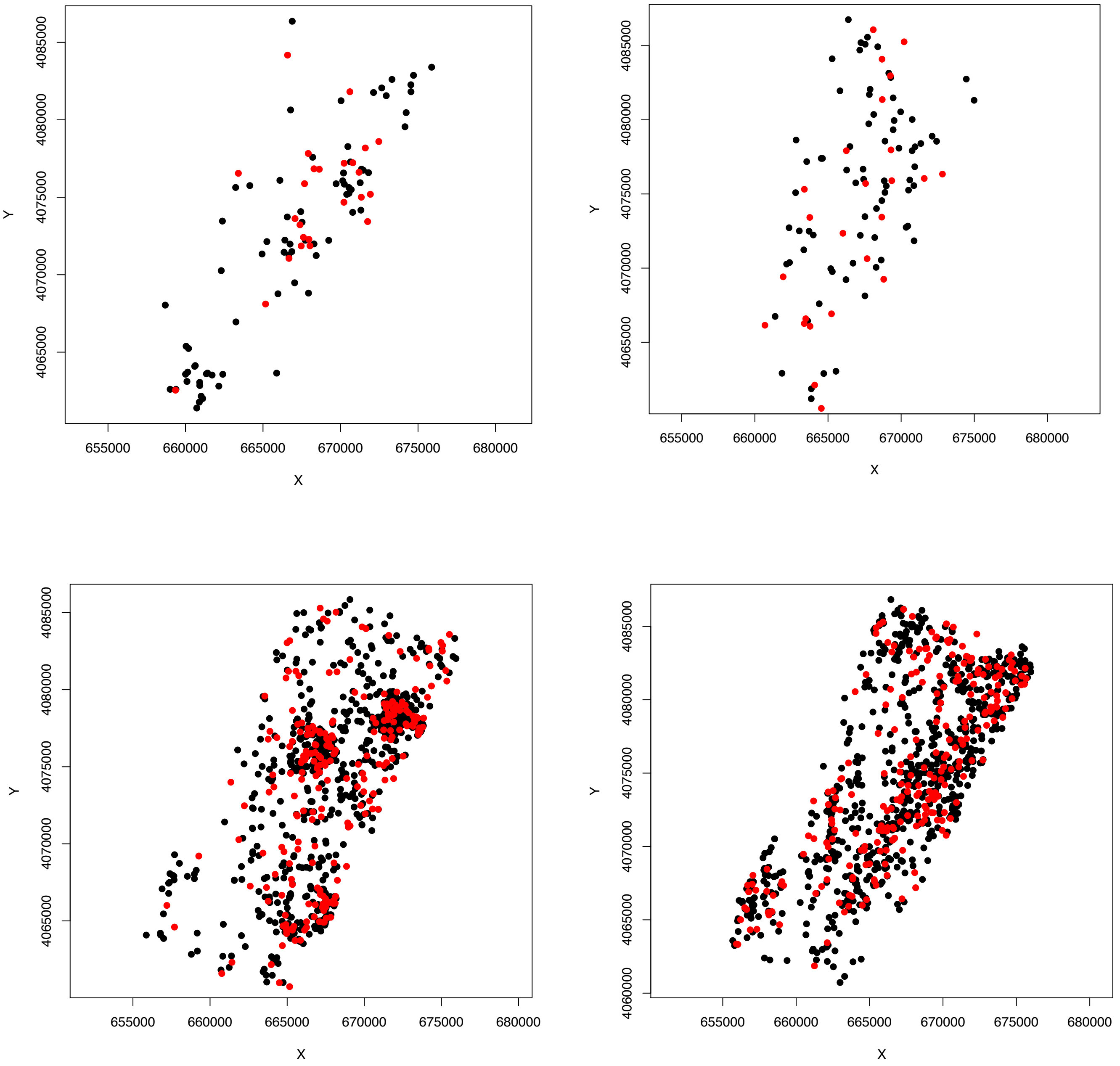
comparisons

Statistical distributions of 'estimated' DSM performance indicators

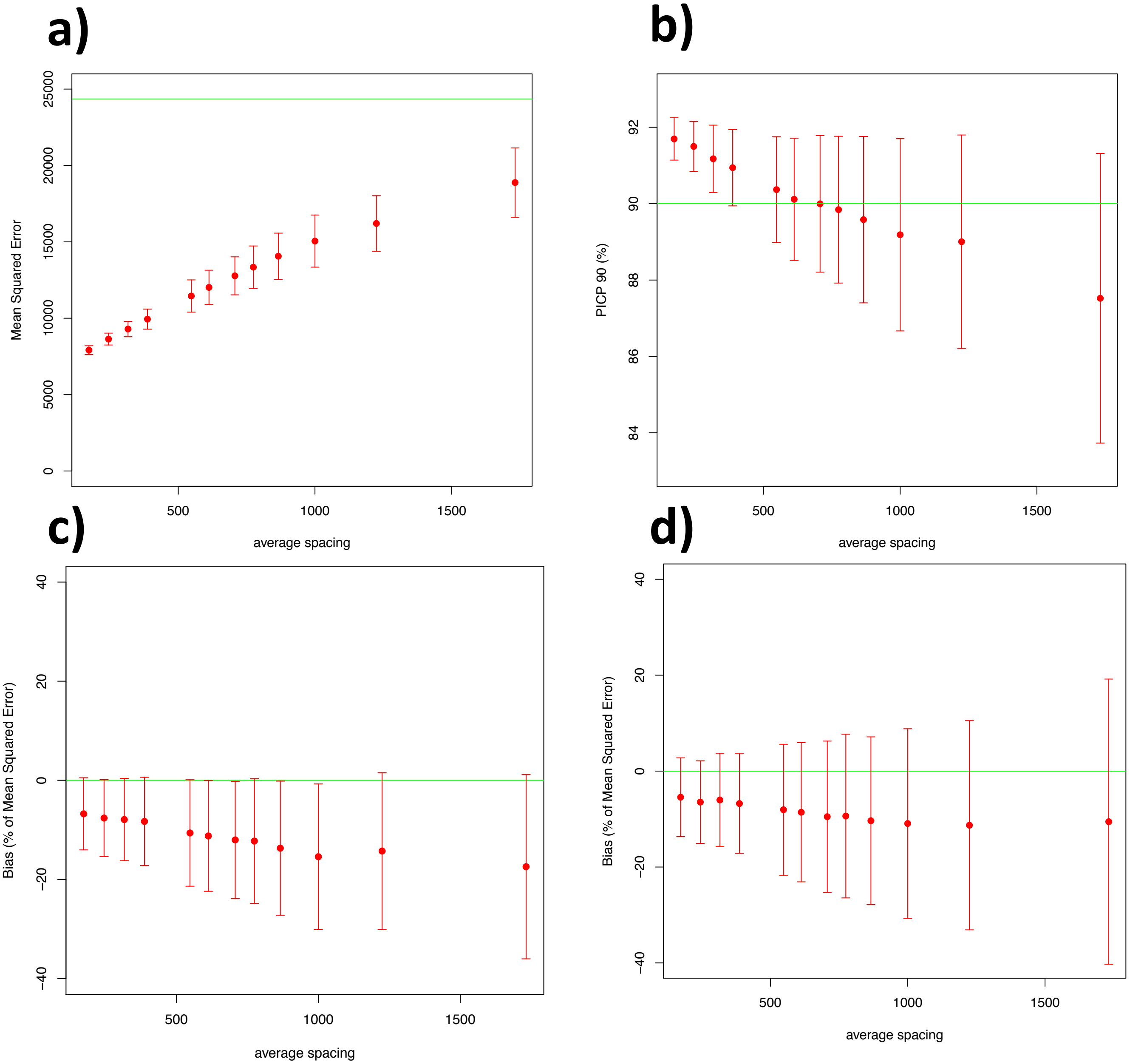
comparisons

Statistical distributions of 'true' DSM performance indicators

Statistical distributions of 'error on performance indicators'



Y
4065000
4070000
4075000
4080000
4085000
655000
660000
665000
670000
675000
680000
X
Y
4065000
4070000
4075000
4080000
4085000
655000
660000
665000
670000
675000
680000
X
Y
4065000
4070000
4075000
4080000
4085000
655000
660000
665000
670000
675000
680000
X
Y
4060000
4065000
4070000
4075000
4080000
4085000
655000
660000
665000
670000
675000
680000
X


a)
Mean Squared Error
0
5000
10000
15000
20000
25000
500
1000
1500
average spacing
b)
PICP 90 (%)
84
86
88
90
92
500
1000
1500
average spacing
c)
Bias (% of Mean Squared Error)
−40
−20
0
20
40
500
1000
1500
average spacing
d)
Bias (% of Mean Squared Error)
−40
−20
0
20
40
500
1000
1500
average spacing

a)

		Average spacing (m)											
space	Indicator	1732	1225	1000	866	775	707	612	548	387	316	245	173
Covariates	coverage	0.11	0.17	0.15	0.16	0.23	0.24	0.38	0.39	0.58	0.67	0.79	0.84
	KLD	0.05	0.01	0.07	0.03	0.05	0.08	0.30	0.09	0.11	0.12	0.36	0.29
	%out of range	0.29	0.32	0.40	0.40	0.40	0.42	0.47	0.46	0.54	0.53	0.64	0.60
geographical	coverage	0.42	0.46	0.53	0.54	0.57	0.58	0.68	0.68	0.74	0.79	0.85	0.87
	KLD	0.38	0.48	0.52	0.54	0.56	0.61	0.67	0.67	0.73	0.78	0.83	0.87
	%out of range	0.20	0.24	0.32	0.30	0.33	0.25	0.29	0.31	0.31	0.37	0.26	0.41
Clay	coverage	0.06	0.10	0.23	0.12	0.18	0.17	0.17	0.08	0.18	0.05	0.04	0.02
	KLD	0.08	0.22	0.12	0.19	0.27	0.25	0.25	0.31	0.26	0.37	0.36	0.37
	%out of range	0.34	0.37	0.36	0.38	0.35	0.39	0.36	0.36	0.27	0.40	0.34	0.29
	variance	-0.38	-0.47	-0.53	-0.54	-0.53	-0.41	-0.56	-0.38	-0.40	-0.25	-0.33	-0.42
	Var/semiVar ratio	-0.18	-0.31	-0.33	-0.38	-0.37	-0.28	-0.42	-0.30	-0.27	-0.13	-0.14	-0.28
Stepwise multiple linear regression R^2		**0.31**	**0.42**	**0.48**	**0.47**	**0.52**	**0.50**	**0.57**	**0.56**	**0.61**	**0.68**	**0.77**	**0.80**

b)

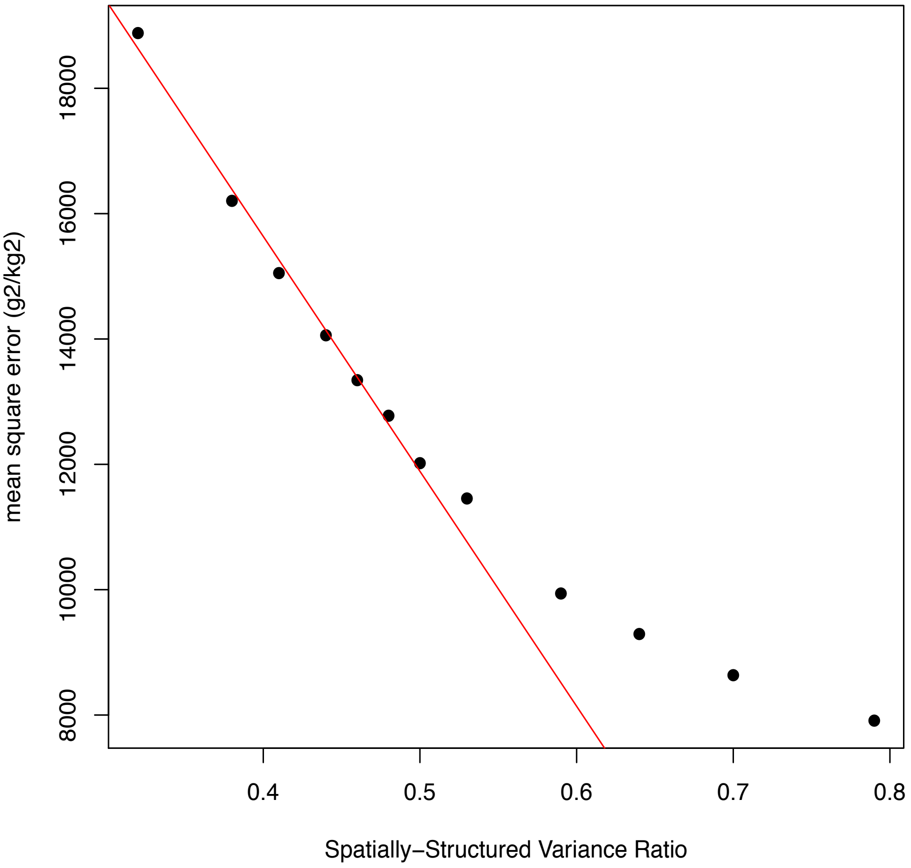
		Average spacing (m)											
space	Indicator	1732	1225	1000	866	775	707	612	548	387	316	245	173
Covariates	coverage	-0.06	-0.03	-0.08	-0.12	-0.17	-0.13	-0.27	-0.23	-0.36	-0.36	-0.42	-0.43
	KLD	0.04	0.12	0.02	0.00	0.02	0.13	-0.24	0.07	-0.06	0.06	-0.16	-0.16
	%out of range	0.20	0.25	0.34	0.32	0.30	0.29	0.37	0.33	0.33	0.42	0.39	0.38
geographical	coverage	-0.27	-0.26	-0.42	-0.41	-0.36	-0.33	-0.50	-0.44	-0.51	-0.49	-0.49	-0.51
	KLD	-0.26	-0.29	-0.42	-0.43	-0.36	-0.35	-0.49	-0.44	-0.49	-0.49	-0.48	-0.49
	%out of range	0.19	0.16	0.25	0.27	0.21	0.16	0.22	0.20	0.25	0.26	0.13	0.26
Clay	coverage	-0.11	-0.16	-0.22	-0.09	-0.15	-0.19	-0.15	-0.14	-0.15	-0.04	-0.05	-0.08
	KLD	-0.10	-0.09	-0.12	-0.11	-0.14	-0.09	-0.15	-0.17	-0.16	-0.18	-0.12	-0.17
	%out of range	-0.62	-0.55	-0.52	-0.53	-0.44	-0.47	-0.47	-0.44	-0.38	-0.38	-0.34	-0.25
	variance	0.76	0.71	0.68	0.68	0.64	0.62	0.65	0.57	0.54	0.49	0.41	0.54
	Var/semiVar ratio	0.34	0.33	0.35	0.41	0.38	0.38	0.40	0.41	0.35	0.34	0.30	0.45
Stepwise multiple linear regression R^2		**0.31**	**0.42**	**0.48**	**0.47**	**0.52**	**0.50**	**0.57**	**0.56**	**0.61**	**0.68**	**0.79**	**0.85**

c)

		Average spacing (m)											
space	Indicator	1732	1225	1000	866	775	707	612	548	387	316	245	173
Covariates	coverage	-0.13	-0.15	-0.17	-0.20	-0.27	-0.29	-0.41	-0.46	-0.67	-0.75	-0.85	-0.91
	KLD	-0.05	-0.01	-0.12	-0.07	-0.07	-0.12	-0.30	-0.12	-0.15	-0.15	-0.37	-0.24
	%out of range	-0.29	-0.37	-0.44	-0.43	-0.43	-0.42	-0.49	-0.47	-0.51	-0.51	-0.58	-0.55
geographical	coverage	-0.41	-0.45	-0.58	-0.57	-0.58	-0.59	-0.68	-0.69	-0.78	-0.81	-0.86	-0.88
	KLD	-0.36	-0.46	-0.57	-0.56	-0.58	-0.60	-0.67	-0.68	-0.76	-0.78	-0.82	-0.85
	%out of range	-0.21	-0.21	-0.32	-0.28	-0.29	-0.25	-0.28	-0.31	-0.31	-0.34	-0.22	-0.33
Clay	coverage	-0.04	-0.07	-0.18	-0.04	-0.15	-0.13	-0.13	-0.08	-0.13	-0.01	-0.03	0.01
	KLD	-0.12	-0.25	-0.22	-0.24	-0.32	-0.26	-0.30	-0.36	-0.29	-0.45	-0.50	-0.42
	%out of range	-0.49	-0.46	-0.40	-0.43	-0.36	-0.41	-0.39	-0.35	-0.30	-0.32	-0.28	-0.26
	variance	0.57	0.54	0.52	0.54	0.54	0.38	0.56	0.36	0.35	0.13	0.18	0.28
	Var/semiVar ratio	0.18	0.18	0.19	0.28	0.27	0.16	0.31	0.21	0.15	-0.03	-0.07	0.08
Stepwise multiple linear regression R^2		**0.49**	**0.50**	**0.52**	**0.52**	**0.55**	**0.51**	**0.60**	**0.58**	**0.67**	**0.73**	**0.83**	**0.87**

d)

		Average spacing (m)											
space	Indicator	1732	1225	1000	866	775	707	612	548	387	316	245	173
Covariates	coverage	-0.01	-0.06	-0.08	-0.05	-0.09	-0.17	-0.23	-0.27	-0.44	-0.53	-0.65	-0.83
	KLD	-0.02	-0.03	-0.02	-0.02	-0.06	-0.08	-0.07	-0.07	-0.13	-0.14	-0.45	-0.38
	%out of range	0.13	0.26	0.28	0.32	0.33	0.39	0.36	0.40	0.41	0.44	0.44	0.47
geographical	coverage	-0.25	-0.30	-0.35	-0.42	-0.45	-0.49	-0.53	-0.60	-0.66	-0.71	-0.78	-0.83
	KLD	-0.23	-0.26	-0.33	-0.39	-0.40	-0.45	-0.53	-0.59	-0.63	-0.68	-0.74	-0.81
	%out of range	0.18	0.24	0.27	0.22	0.26	0.23	0.35	0.30	0.30	0.32	0.33	0.31
Clay	coverage	0.08	-0.02	-0.02	-0.03	-0.03	-0.05	-0.16	-0.10	-0.07	-0.07	-0.03	-0.01
	KLD	-0.10	-0.07	-0.15	-0.15	-0.25	-0.22	-0.23	-0.27	-0.29	-0.39	-0.47	-0.39
	%out of range	-0.46	-0.43	-0.38	-0.46	-0.40	-0.38	-0.35	-0.34	-0.35	-0.23	-0.19	-0.26
	variance	0.56	0.51	0.46	0.50	0.49	0.39	0.51	0.34	0.33	0.14	0.16	0.26
	Var/semiVar ratio	0.11	0.11	0.12	0.16	0.12	0.07	0.16	0.09	0.12	-0.05	0.05	0.04
Stepwise multiple linear regression R^2		**0.39**	**0.36**	**0.34**	**0.42**	**0.42**	**0.40**	**0.45**	**0.46**	**0.51**	**0.59**	**0.69**	**0.80**


mean square error (g2/kg2)
8000
10000
12000
14000
16000
18000
0.4
0.5
0.6
0.7
0.8
Spatially−Structured Variance Ratio