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

Philippe Lagacherie, Dominique Arrouays, Hocine Bourennane, Cécile Gomez, Manuel P. Martin, Nicolas P.A. Saby

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

Philippe Lagacherie[1], Dominique Arrouays[2], Hocine Bourennane[3], Cécile Gomez[1], Manuel Martin[2], Nicolas P.A. Saby[2]

1. LISAH, Univ Montpellier, INRA, IRD, Montpellier SupAgro, Montpellier, France
2. Infosol, INRA, 45075 Orléans, France
3. UR Sols, INRA, 45075 Orléans, France

Corresponding author: Philippe Lagacherie, LISAH, INRA, 2 place Viala, 34060 Montpellier (France). philippe.lagacherie@inra.fr

Abstract

Digital Soil Map uncertainty is usually evaluated from a set of independent soil observations that are used to determine various uncertainty indicators. However, the number and locations of the sites that constitute these evaluations may impact the value of these indicators.

In this paper, a numerical experiment on uncertainty indicators was performed using the pseudo values of topsoil clay content obtained from an airborne hyperspectral image in the Cap Bon region (Tunisia). These pseudo values form a soil pattern with a large extent (46% of 300 km^2), high resolution (5 m) and good accuracy (R^2_{val} = 0.75) while being free of visible artefacts and pedologically plausible. Therefore, the dataset was considered a fair representation of reality while providing a quasi-unlimited choice of sites.

The numerical experiment considered three Quantile Regression Forests as examples of DSM models by using inputs from relief soil covariates and geographical locations that were calibrated from 200, 2,000 and 100,000 individuals respectively (low, medium and high quality models). Their uncertainty indicators were first evaluated by calculating four uncertainty indicators (ME, MSE, SS_{MSE} and PICP) from a large independent validation set of 100,000 sites. These uncertainty indicators were then computed from independent evaluation sets of different sizes (from 50 to 500 sites) and from different locations (500 evaluation sets of each size). The independent evaluation sets were selected following a stratified random sampling using compact geographical strata.

The numerical experiment showed that the values of the uncertainty indicators were highly variable across numbers and locations of sites. The largest variations were observed for evaluation sets with fewer than 100 sites, but non-negligible variations remained for larger evaluation datasets. This result suggested that evaluations from independent sets convey a non-negligible error on the uncertainty indicators, which increases as the number of sites decrease.

Evaluations of DSM models from independent evaluation sets should be interpreted with care and uncertainty on validation results should be systematically estimated. For that, numerical experiments for benchmarking DSM models on known soil patterns across the world would be a valuable complement to the analytical expressions for the uncertainty indicators and the many DSM applications for which these analytical expressions are not valid. This would serve also to improve the sampling techniques for the calibration and evaluation datasets to reduce the error when estimating the uncertainty of a DSM product.

Keywords: Soil mapping, Uncertainty, Hyperspectral imagery, Random forest, Sampling

1. Introduction

Soil maps are simplified representations of more complex and partially unknown patterns of soil variations. Therefore, any prediction of a soil property that can be derived from these soil maps has an irreducible and often substantial error that is the difference between the true and estimated values of the soil property. Since there is no way to systematically measure this error, we are uncertain about the true value of the soil property at most of the locations, where uncertainty refers to the state of mind of a person who expresses a lack of confidence about reality (Heuvelink, 2014).

Getting accurate estimates of this uncertainty is of paramount importance for end-users to make enlightened decisions on the utility and limitations of the soil data products delivered to them. An example of a concrete translation of this exigence is provided by the GlobalSoilMap specifications (Arrouays et al., 2014). Following these specifications, each soil property estimate should be provided under the form of a 90% prediction interval (PI), which reports the range of values in which the true value is expected to occur 9 times out of 10.

To provide this uncertainty information, a rigorous assessment of the uncertainty is considered a mandatory component of any Digital soil mapping application. Two main groups of methods can be used to accomplish this task (Heuvelink, 2014). The first is a model-based approach that involves spatial stochastic models that can provide estimates of the uncertainty of their own predictions. The main drawback of these model-based approaches is that their uncertainty evaluations are only valid under certain assumptions of the stochastic models. They also use the same data that were used to calibrate the models, which may cause an

72 underestimation of the uncertainty. The second group of uncertainty evaluation methods
73 avoids this restriction by undertaking a model-free statistical evaluation with independent
74 sites selected by probability sampling. The comparison between the predicted and actual
75 values of the soil attributes of interest yields a set of uncertainty indicators, the most common
76 of which are the mean error (ME), the mean squared error (MSE), the root mean square error
77 (RMSE) and the coefficient of determination (R^2). The latter is confusingly referred to in the
78 DSM literature either as the goodness-of-fit of a linear regression between the predicted and
79 observed values, or to how close the paired prediction-observation points are to the 1:1 line.
80 For DSM models that deliver predictions via probability distributions, accuracy plots
81 (Goovaerts, 2001) or Prediction Interval Coverage probability (PICP) (Shrestha and Solomatine
82 2006), it can also be calculated to evaluate the accuracy of the uncertainty attached to each
83 prediction.
84 Various methods for producing sets of independent sites (denoted further 'evaluation sets'),
85 including the data-splitting of the available dataset, cross-validation over the available dataset
86 and the probabilistic sampling of additional sites, were reviewed by Brus et al (2011). They
87 recommended the latter to ensure unbiased uncertainty estimations and provide a more
88 complete view of the spatial distribution of the error via the Spatial Cumulative Distribution
89 Function (SCDF).
90 The main drawback of any model-free statistical evaluation is that the calculated uncertainty
91 indicators are themselves prone to uncertainty. Indeed, similar to any statistical parameters
92 that are derived from a set of individuals, the uncertainty indicators (e.g., R^2, ME, and PICP)
93 are sensitive to the number and the locations of the soil observations used for calculating
94 them. Under the condition that a probabilistic sampling is applied, Brus et al (2011) provided
95 an analytical expression of the mean error (ME) and the mean square error (MSE) variances

using the sampling fraction (the ratio between the number of samples and the number of possible sample locations) and the estimated variance of the error over a given area as inputs. Such analytical expressions of the ME and MSE variances measure the uncertainty of these uncertainty indicators, which could be very useful. For example, they could be used for determining whether the differences between the ME and MSE of two different DSM models reveal significantly different performances or for computing the size of the validation sample that is required for estimating the ME and MSE at a given precision level. However, it must be noted that i) for some indicators other than ME and MSE and ii) for some specific sampling designs (such as two-stage random sampling), there might not be readily available statistical estimates of the uncertainty of the uncertainty indicators (Brus et al. (2011)).

Although the uncertainty of the ME and MSE is easily computable for some sampling designs, it is far from being currently computed in DSM applications. This is because the soil datasets used as inputs of DSM models are rarely suitable for applying a probabilistic sampling. Indeed, these soil datasets are often undersized with regard to the size of the study area and the complexity of the soil cover to be modelled. This leads to substantial losses of predictive performances as soon as the sampling effort for collecting calibration sites is depleted for populating the set of independent sites required by probabilistic sampling. Furthermore, DSM applications most often use legacy data that do not respect the randomness and evenness required for a probabilistic sampling. For all these reasons, the DSM mappers most often disregard calculating any uncertainty in their uncertainty indicators and therefore neglect this issue when evaluating the DSM products.

This paper presents a numerical experiment for assessing the uncertainties of the uncertainty indicators of the three DSM models with contrasted predictive performances by using different probabilistic samplings of different sizes. To overcome the above-evoked limitations

of the current soil input data, the study used the virtual pattern of the topsoil pseudo clay content derived from airborne Vis-NIR-SWIR hyperspectral data acquired over the Cap Bon region (300 km^2, Tunisia) at a five-meter resolution (Gomez et al., 2015). This pattern is constituted 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. Such a soil dataset provided a quasi-unlimited number of pseudo-measured sites that made probabilistic sampling (and therefore ME and MSE variance calculations) applicable without any effect on the predictive performances. It also enabled the calculation of any uncertainty indicators from their empirical distributions obtained by repeating the validation process n times, which means selecting n different validation sets of a given size and determining the uncertainty indicators each time.

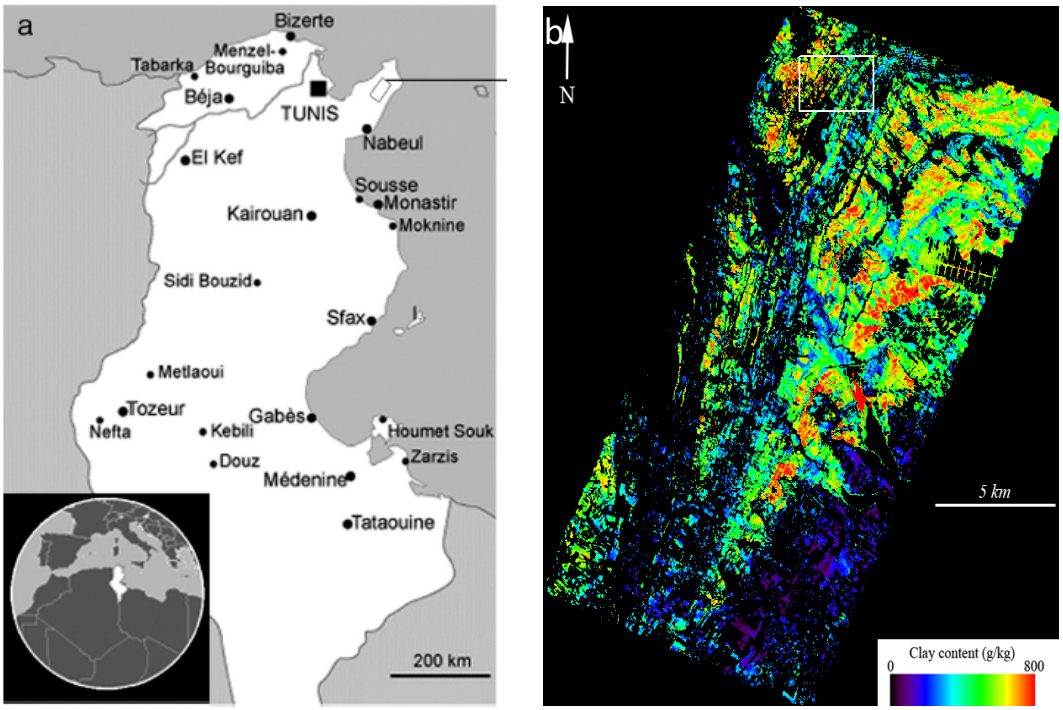


Figure 1: Location of the study area (a) and the spatial pattern of pseudo values of topsoil clay content (b)

2. Material and methods

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 km² 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., 2012b). 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 southeastern region of the study area has a flatter landscape with sandy Pliocene deposits in which calcosols and rendzinas prevail.

159 2.2. Data

160

161 2.2.1. Hyperspectral image and derived topsoil clay content predictions

162

163 The numerical experiment uses an image of topsoil clay content as input. The data were
164 derived from a Vis-NIR-SWIR hyperspectral image (Gomez et al., 2012b). The approach used
165 to produce the data is summarized below. More details about the pre- and post-processing of
166 the hyperspectral image can be found in Gomez et al (2012b).
167 On November 2, 2010, AISA-Dual airborne-based hyperspectral data were acquired over the
168 study area with a spatial resolution of 5 m. The area of the image is approximately 12 km x 24
169 km. The AISA-Dual spectrometer measured the reflected radiance via 359 non-contiguous
170 bands covering the 400 to 2450 nm spectral range, with 4.6 nm bandwidths between 400 and
171 970 nm and 6.5 nm bandwidths between 970 and 2450 nm. The instantaneous field of view
172 (IFOV) was 24 degrees. Topographical corrections were performed using a digital elevation
173 model built from ASTER data and ground control points.
174 To isolate the bare soil areas, the study masked pixels with normalized difference vegetation
175 index (NDVI) values greater than an expert-calibrated threshold (0.20). Water and Urban areas
176 were also removed. Finally, the bare soil represented 46.3% of our study area and potentially
177 5,889,847 measured AISA-Dual 5 m x 5 m pixels.
178 A Partial Least Square Regression (PLSR) technique (Tenenhaus, 1998) was then applied to
179 estimate the topsoil clay contents from the 280 reflectance bands provided by the AISA-DUAL
180 airborne sensor at each location. The PLSR was calibrated from 129 couples of Vis-NIR-SWIR
181 reflectance spectra acquired by the AISA-DUAL sensor on bare soil surfaces associated with
182 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 centering 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 ($R^2 = 0.75$). 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), which is denoted "pseudo values of Clay content". 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 and Multi-Resolution Valley Bottom Flatness (MRVBF). Sine and cosine transformations were applied to the 'aspect' to obtain four indices with a continuous gradient: 'northness', 'easterness','north-westerness' and 'north-easterness'. Finally, the X and Y coordinates (the n of "scorpan" in McBratney et al., 2003) were also used as soil covariates.

2.3. Checking the plausibility of the predicted soil patterns

We conducted a prior check to confirm the reliability of using the pseudo values of the clay content determined above as a realistic example of a soil property pattern. It was particularly

important to check the absence of any distortion of the spatial pattern due to the spectral measurements by remote sensing. Three experimental variograms showing the spatial structure of clay variations were calculated and then fitted with an exponential model using the weighted least square method (Cressie, 1993) using three different data sets: i) the 129 sites with laboratory measurements of the clay content, ii) the same sites with pseudo values of the clay content and iii) 100,000 randomly selected sites with pseudo values of the clay content. Comparisons of the variograms were performed (see section 3.1.).

2.4. Sampling technique

We used probability sampling for selecting the calibration set used to build the DSM model, evaluating the performance of these models and undertaking the numerical experiment. All of the probability sampling techniques followed the same sampling approach (the stratified random sampling technique) using compact geographical strata (Walvoort et al., 2010) recommended by Brus et al. (2011). The main advantage of this technique is that it ensures an even distribution of the samples over the studied area and is simple to apply.

The stratified random sampling approach was applied as follows. The study area was first stratified into 25 geographical strata of equal area using a K-means classification of the X and Y coordinates of each locations. We then randomly selected pixels of the grid within each stratum with a fixed number of locations in accordance with the total of samples required.

2.5. Uncertainty indicators

We considered four uncertainty indicators among the possible ones that could have been examined. The first two were the mean error (ME) and the mean squared error (MSE) that were selected because these are the classical indicators whose variance can be calculated from analytical expressions (Brus et al, 2011). The last two were the mean square error skill score (SS_{mse}) (Wilks, 2011 p 359, cited by NussBaum et al, 2017)) and the Prediction Interval Coverage probability (PICP) (Shrestha and Solomatine 2006). SS_{MSE} is similar to the R^2 reported in some studies (Vaysse and Lagacherie, 2015, Viscarra-Rossel, 2015) as the percentage of variance explained by the model:

$$SS_{mse} = 1 - \frac{\sum_{i=1}^{i=n}(z_i - z_i^*)^2}{\sum_{i=1}^{i=n}(z_i - \hat{\bar{z}}_i)^2} \qquad (1)$$

Where z_i and z_i^* are the respective observed and predicted values of property z at location I, and $\hat{\bar{z}}_i$ is the mean value of *z.*

The PICP expresses the probability that all observed values fit within the 90% prediction limits provided by the DSM model (see section 2.7.1.).

It must be noted that the first three uncertainty indicators (ME, MSE and SS_{MSE}) are measurements of the accuracy of predictions, whereas PICP is a measurement of the accuracy of the uncertainty prediction.

The calculations of these indicators should take into account the fact that a stratified random sampling technique is applied for selecting the independent sites. Following Brus et al (2011), these calculations are as follows.

The estimations of the ME, MSE, SS_{MSE} and PICP correspond to a global mean that was estimated by design-based inference, particularly by the usual estimator for stratified random sampling.

$$\hat{\bar{y}} = \sum_{h=1}^{H} w_h \hat{\bar{y}}_h \qquad (2)$$

$$\hat{\bar{y}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \qquad (3)$$

Where H is total number of strata ($H = 25$), w_h is the weight of stratum h quantified by the relative area, $\hat{\bar{y}}_h$ is the estimated mean of the stratum h, n_h is the number of sampling points in stratum h, and y_{hi} is the measurement of the indicator at location i in stratum h. For the ME, y_{hi} can be replaced by the difference between the actual value z_{hi} and the prediction ${z^*}_{hi}$. The MSE can be estimated by replacing y_{hi} with the squared difference between the actual value z_{hi} and the prediction ${z^*}_{hi}$. The SS_{MSE} and the PICP can be estimated by replacing y_{hi} by the SS_{MSE} and the PICP of the stratum h.

2.6. Analytical calculations of standard errors of ME and MSE

If probabilistic sampling is applied, it is possible to calculate the standard errors of the ME and MSE. In the case of a stratified random sampling, the equations are (from De Gruijter, 2006)

$$\widehat{SE}(\bar{y}) = \sqrt{\sum_{h=1}^{H} a_h^2 \hat{V}(\hat{\bar{y}}_h)} \qquad (4)$$

Where $\hat{V}(\hat{\bar{y}}_h)$ is the sampling variance of the stratum mean $\hat{\bar{y}}_h$, which is estimated by :

$$\hat{V}(\hat{\bar{y}}_h) = \frac{s_h^2}{n_h} \qquad (5)$$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \hat{\bar{y}}_h)^2 \qquad (6)$$

2.7. DSM modelling

Since the validation process tested in this paper is model free, any model used in DSM could have been selected as an example of a DSM model. However, only two criteria were

considered: i) the model had to provide local uncertainty predictions for being able to test PICP, and ii) the model should be run without manual intervention and repeated a great number of times in the numerical experiment. We combine these two criteria results in selecting the Quantile Regression Forest as the example DSM model.

2.7.1. Random Forests and Quantile Random Forests

This section describes Random Forests and Quantile Random Forests. More details on these two machine learning algorithms are given in the seminal papers by Breiman et al (2001) and by 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), $i = 1, \ldots, n$. 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 $\boldsymbol{\theta}$ of a Random Forest for a new data point x can be represented as the weighted average of the original observations Y_i, $i = 1,\ldots,$ n :

$$\hat{Y}_\theta(x) = \sum_{i=1}^{n} w_{\theta i}(x,\theta) Y_i \quad [7]$$

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

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

$$\widehat{Y_T}(x) = \sum_{i=1}^{n} w_{Ti}(x) Y_i \quad [8]$$

With $$w_{Ti}(x) = k^{-1} \sum_{t=1}^{k} w_{\theta i}(x,\theta) \quad [9]$$

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

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \quad [10]$$

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).

2.7.2. Calibration and Validation of QRFs

Three different QRF were calibrated with 200, 2,000 and 50,000 sites with known pseudo

values of clay content. The locations were selected according to the stratified sampling techniques described above. The increasing number of sites was selected to obtain contrasted predictive performances (see section 3.2.)

After removing the calibration sites, we selected a master evaluation set of 100,000 independent sites by applying the stratified random sampling technique using the compact geographical strata described in section 2.5. The reference values of the three uncertainty indicators of interest were computed from this master evaluation set. This set was then removed from the set of possible sites to ensure the independence of the further numerical experiment.

2.8. Empirical simulation

The empirical simulation aims to evaluate the amount of variation in the four uncertainty indicators (ME, MSE, SS_{MSE} and PICP) when different evaluation sets are selected. This variation can then be used as an estimate of the uncertainty caused by relying on the choice of a single specific evaluation set, as is always the case in reality.

The empirical simulation proceeds as follows:

1. Sample a set of *n* evaluation sites using a stratified random sampling technique using compact geographical strata,
2. Calculate the uncertainty indicators over the *n* sites,
3. Repeat steps 1 and 2 500 times, and
4. Compute the distributions and their summary statistics from the 500 values of the uncertainty indicators.

346 The tested numbers of sites ranged between n = 50 and n = 500 with an increment of 25. This
347 represents the densities of the observations ranging between 1/2.67 km^2 and 1/0.27 km^2.
348
349 2.9. Software
350
351 The software for Random Forests and Quantile Random Forest are made available in R (R
352 Development Core Team, 2008) with the packages RandomForest (Liaw and Wiener, 2002)
353 and quantregForest (Meinshausen and Schiesser, 2015), respectively. Stratified sampling
354 using compact geographical strata is implemented in the R package "spcosa" (Walvoort et al.,
355 2010). Variogram studies (section 3) were performed with the gstat package (Pebesma, 2004).
356
357 **3. Results**
358
359 3.1. Check of the predicted soil patterns
360
361 To check the plausibility of using the pseudo values of clay content derived from hyperspectral
362 data, we compared the experimental variograms and fitted model variograms obtained using
363 real clay content measurements at measured sites, the pseudo values of the clay content at
364 the same locations, and a set of 10,000 sites with the pseudo values of the clay content.
365 Figure 2 shows the experimental and the fitted variograms of the topsoil clay content
366 calculated from different inputs. The parameters of the variogram estimated from the pseudo
367 values of clay content (figure 2b) were similar to those estimated with the real clay content
368 measurements at the same locations (figure 2a). Indeed, the shapes, ranges and sills were

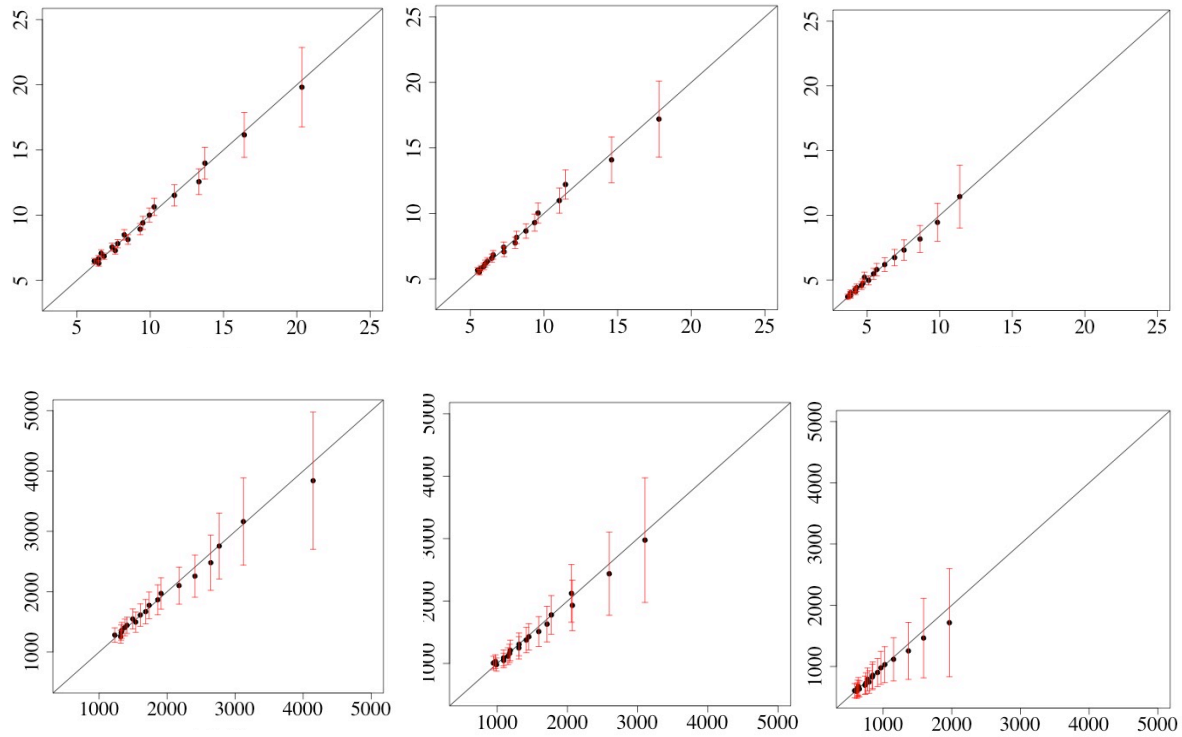
close to one another. The only noticeable difference was a smaller nugget value exhibited by the variogram of the pseudo values of the clay contents.

The experimental variogram obtained from 100,000 sites was much less noisier than but very similar to the previous one.

Figure 2: Variograms of clay content obtained from a) 129 sites with clay content laboratory measurements, b) the same 129 sites with pseudo values of clay content c) 100,000 sites with pseudo values of clay content.

	Uncertainty indicators			
Number of calibration sites	**ME (g/kg)**	**MSE(g^2/kg^2)**	**SS_{MSE}**	**PICP (%)**
200	-11.3 (0.5)	19555 (89)	0.29	88.5
2,000	-7.0 (0.4)	13550 (69)	0.51	90.3
50,000	-2.7 (0.3)	6090 (42)	0.78	90.3

381 *Table 1: Estimated ME (standard error), MSE (standard error), SS_{MSE}, and PICP from the master set of*
382 *independent sites*
383
384 3.2. DSM model performances
385
386 Table 1 shows the values of the uncertainty indicators calculated from the master set of
387 independent sites (100,000 sites) for the DSM models obtained by calibrating the quantile
388 regression Forests with three sizes of calibration sets. As expected, the overall accuracy of the
389 measured predictions (as measured by SS_{MSE}) increased significantly as the number of
390 calibration sites increased, while the bias measured by the ME and mean squared error (MSE)
391 decreased. The PICP values were found to be close to the expected value of 90 for the two
392 models with the greatest numbers of calibration sites. Meanwhile, the model built from 200
393 calibration sites exhibited a PICP below 90, which revealed a slight underestimation of the
394 uncertainty. Finally, the results obtained by the three models well covered the large range of
395 performances of DSM models that can be encountered in the literature. It is also worth noting
396 that the standard errors of the ME and MSE that were calculated from the variances given by
397 equations 8 and 9 are very small, which means that the performances of the three models are
398 significantly different from each other.
399 In the following, the QRFs calibrated from 200, 2,000 and 50,000 sites are denoted,
400 respectively, as “low-quality QRF”, “medium-quality QRF” and “high-quality QRF”.



401

402 Figure 3: Comparisons between the estimations of standard errors of ME and MSE derived from the
403 numerical experiment (X-axis) and their analytical calculations (Y-axis). ME (first row), MSE (second
404 row), low quality QRF (first column), medium quality QRF (second column) and high-quality QRF (third
405 column). Dots : mean values of the calculated standard errors on ME and MSE over the 500 iterations,
406 bars: twice the standard deviations of the calculated standard errors on ME and MSE over the 500
407 iterations)

408

409 3.3. Comparisons of estimated standard errors of ME and MSE

410 For each DSM model, the standard deviations of the ME and MSE observed over the 500
411 iterations for the different sizes of evaluation sets were compared with the standard errors of
412 the ME and MSE calculated following equations 4, 5 and 6 (figure 3).

413 The dots of figure 3 representing the mean values over the 500 iterations of the calculated
414 standard errors of the ME and MSE were found to be very close to the 1:1 line. This denoted
415 good agreement between the analytical calculations of the ME and MSE and the simulation
416 outputs, which was expected. Interestingly, the error bars around the dots of figure 3

representing twice the standard deviation over the 500 iterations of the analytical calculations of the standard errors of the ME and MSE were great for the largest standard errors and decreased as the standard errors decreased, regardless of the indicator and the DSM model. This revealed the residual impacts of the evaluation sites' locations that were selected by the spatial sampling, the DSM model and the size of the evaluation set being fixed.

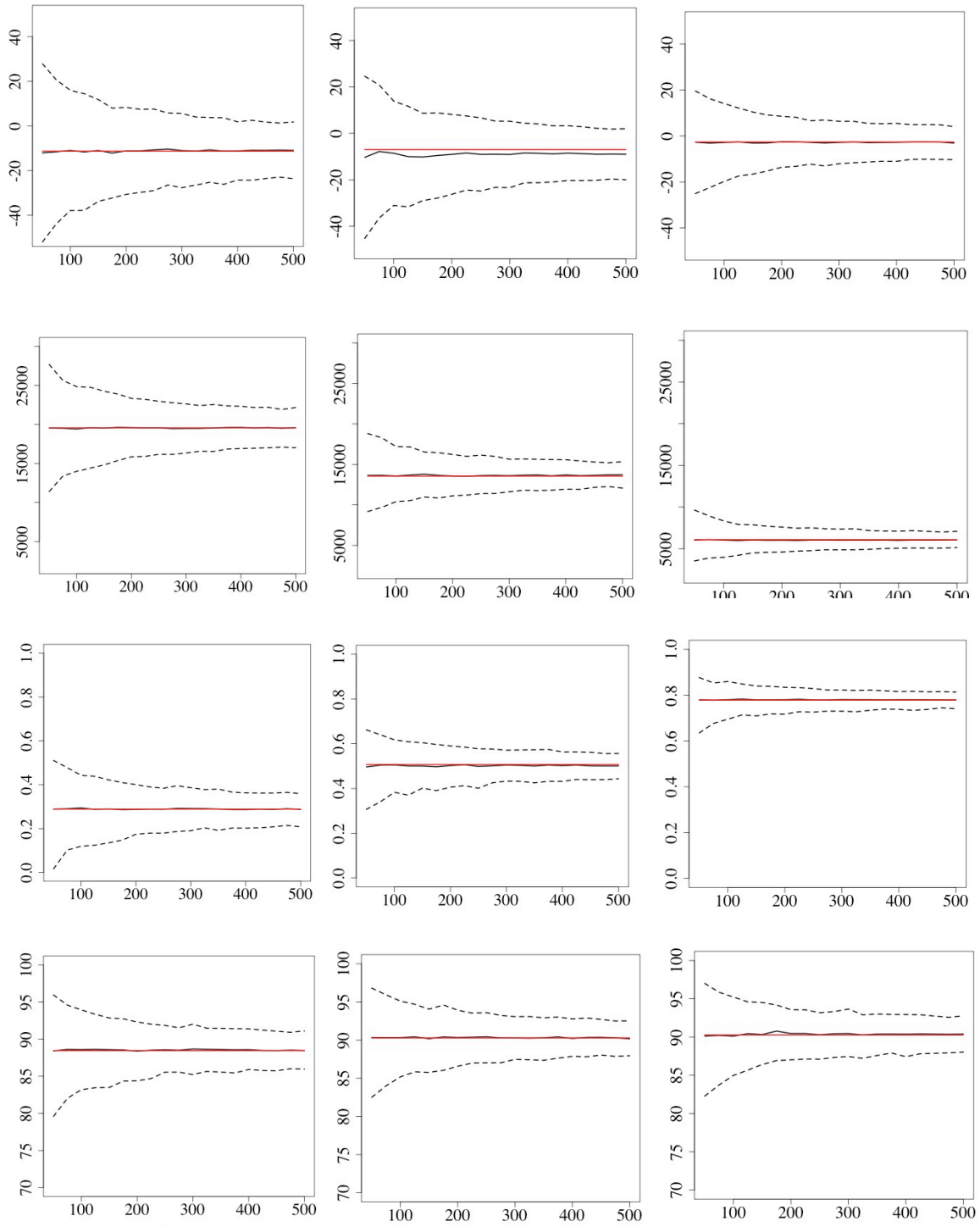


Figure 4: Values of uncertainty indicators obtained using validation sets of different sizes (from 50 to 500 sites): ME (first row), MSE (second row),SS_{MSE} (third row) and PICP (fourth row), Low quality QRF (first column), medium quality QRF (second column) and high-quality QRF (third column). Black line:

mean value of the uncertainty indicators over 500 trials. Dotted black lines: 0.05 and 0.95 quantiles of the uncertainty indicators over 500 trials.. Red line: reference values of the uncertainty indicators

3.4. Numerical experiment results

Figure 3 shows the result of the numerical experiment on the three tested QRFs and the four uncertainty indicators. Each result consists of three graphical curves showing the evolution of the median (black line) and the 0.05 and 0.95 quantiles (dotted black lines) of the uncertainty indicator values, with the size of evaluation sets expressed by its number of sites (in abscissa). A red line shows the reference value of the indicator, as calculated from the master evaluation set.

All the results exhibited similar patterns. The mean values of the uncertainty indicators were all close to the reference values. The differences between the quantiles (or the confidence interval widths) were generally large, which reveals imprecise estimations of the uncertainty indicators.

An increase in precision was observed as the sizes of the evaluation sets increased. For small sets (fewer than 100 sites), the confidence interval widths were so great that the estimations of the uncertainty indicators were weakly informative. For the largest sets (500 sites), the confidence interval widths were much smaller but still conveyed a non-negligible imprecision. For example, the confidence interval widths for the ME, SS_{MSE} and PICP were, respectively, 17 g/kg, 0.11 and 5% for the medium precision QRF.

Some differences in the results across uncertainty indicators and models had to be noted. The high precision QRF exhibited less difference between quantiles than did the two other QRFs for the ME, MSE and SS_{MSE}. A “better” model would also be a model whose performance can

449 be more easily assessed. However, this was not the case for PICP, which exhibited differences
450 in quantile values as large as those observed for the two other models.
451
452
453 **4. Discussion**
454
455 4.1. Suitability of hyperspectral data for testing DSM models
456
457 The comparisons of the variograms showed that the pseudo values of the clay content
458 obtained from hyperspectral data represented the spatial structure of a soil property well,
459 apart from a smoothing of the small range variability revealed by a decrease in the nugget
460 value. It must be noted that similar sill and range closeness and nugget decreases were
461 observed in a previous study (Gomez et al, 2012a). This smoothing is attributable to several
462 factors that may perturb the spectral signature of the topsoil clay content (Lagacherie et al.,
463 2008): atmospheric conditions, changes in support (a square block of 5 m side) compared to
464 soil sampling, or variations in the stoniness, vegetation and rugosity of the soil surface. This
465 may be also the result of using the partial least square regression that as a linear model,
466 smooths the variations of the predicted variable. This slight underestimation of the variability
467 of clay content could lead to slightly underestimated uncertainty indicators. However, this
468 artefact cannot compromise the results obtained from the variations across the sampling of
469 these indicators.
470 The spatial pattern of the pseudo values of clay content (as obtained from hyperspectral data)
471 can be considered a good approximation of a real pattern of soil properties while providing a
472 quasi-unlimited set of possible sites with soil property measurements. In this paper, we exploit

these advantages for experimentally assessing the quality of estimations of the usual uncertainty indicators of DSM models, which, to the best of our knowledge, has not been done before. Furthermore, such data make it possible to accurately validate and compare DSM models thanks to the large size of the validation sets, from which the uncertainty indicators can be computed.

Although it cannot be envisaged that spatial sets of pseudo values of soil properties derived from airborne hyperspectral imagery could be collected for each DSM application, several study areas with such soil datasets across the world can be used to enlarge the range of tested soil properties and pedological contexts (Schwangart and Jarmer, 2011; Stevens et al., 2010; Ben Dor et al., 2002, Gomez et al., 2012a, Vaudour et al, 2016). Furthermore, other study areas could be added to this initial set with the aim of building national, regional or global benchmarks for DSM models, which exist in other disciplines (Rosensweig et al., 2013, Luo et al., 2012).

4.2. Uncertainty of the evaluation process

Our results (figure 4) revealed that the uncertainty indicators can vary across evaluations sets, from different sample counts, and between evaluation sets of the same number of samples. This highlights that the uncertainty indicators calculated from statistical validations are themselves (depending on the evaluation setup) prone to non-negligible uncertainty.

It therefore can be claimed from these results that too low of a number of evaluation sites cannot accurately estimate the performance of DSM models since the values of the uncertainty indicators may vary a lot with the locations of the evaluation sites. Even with large numbers of evaluation sites and an unbiased probability sampling, there can still be an

497 imprecision that prevents the models from being ranked in terms of the calculated uncertainty
498 indicators if the differences in their performances are too small. For example, two models with
499 differences in SS_{MSE} below 0.05 may have confidence intervals that overlap each other by more
500 than half their width. However, it must be noted that the ME, MSE and SS_{MSE}^2 calculated for a
501 "high quality" model could be less uncertain. The decrease in the uncertainty of predictions
502 also correspond with a decrease in the spatial variability of the errors, which in turn may
503 correspond with a decrease of the sensitivity of these uncertainty indicators to the evaluation
504 dataset. The question then is how does one know for sure that a model is of high quality;
505 because of the variability of the indicators for low and average quality models, some models
506 might mistakenly be considered high quality models. This seems particularly true when
507 considering indicators such as PICP and MSE. This emphasizes the usefulness of estimating
508 models using a set of complementary indicators.
509
510 4.3. Benefit of the numerical experiment
511 It must be noted that the impact of the sample size on the uncertainty of some uncertainty
512 indicators (ME and MSE) is already well established in analytical expressions (equations 4, 5
513 and 6) that have been applied in a few DSM studies (e.g. Kempen et al, 2011). As expected,
514 our numerical experiment well reproduced the results obtained from the analytical
515 expressions both for estimating the uncertainty of the ME and MSE (figure 3) and for
516 reproducing the effect of the sample sizes (Figure 4). Beyond reproducing these results, the
517 large number of evaluation procedures performed in a same study (27,000) provided a
518 comprehensive understanding of the respective impacts of the DSM model quality, the
519 number of evaluation sites and the location of these sites on the values of uncertainty
520 indicators, including those for which no analytical expression still exist (SS_{MSE}, PICP). It also

521 revealed that the analytical calculations of the ME and MSE uncertainties were themselves
522 prone to uncertainty that could be important for the evaluation sets having the largest
523 variances (Figure 3), which corresponded to those with the smallest sized evaluation sets
524 (Figure 4).
525 In the future, such a numerical experiment can also be used for obtaining references about
526 the loss of precision in evaluating the uncertainty of DSM models when it is not possible to
527 perform a probabilistic sampling and thus to calculate ME and MSE by the analytical
528 expressions. This occurs in most of the current DSM applications that use legacy soil data
529 whose locations have been selected by a soil surveyor following a non-probabilistic process
530 ("free survey").
531
532 4.4. Improving the evaluation process
533
534 Our results clearly showed that the uncertainty of Digital Soil Mapping products cannot be
535 estimated with a great precision. This must be better taken into account in the practices of
536 soil mapping evaluations. A first recommendation is to systematically assess the standard
537 error of the uncertainty indicators using the available analytical formulations when possible
538 or by bootstrapping the validation set. Furthermore, better attention should be paid to the
539 sampling techniques used to select the evaluation sites. Indeed, stratified random sampling
540 using compact geographical strata ensures an even distribution of sites across space but does
541 not avoid the error on the uncertainty indicators. Reducing this error by using more
542 sophisticated sampling techniques is a priority. In this perspective, our case study provides a
543 quasi-infinite number of validation sets that exhibited differences with the master validation
544 set regarding the values of the uncertainty indicators. Analysing the variability of these

differences would permit the sampling criteria to be found that would ensure more accurate estimations. The new sampling techniques could also be extended to the calibration datasets of DSM models that provide a priori estimations of their errors or that are used for evaluations with cross-validation techniques (Brus et al., 2011). In addition, using uncertainty indicators that are less sensitive to outliers (Nusbaum et al., 2014) would be a complementary way to reduce the uncertainty of the uncertainty revealed by this paper. Finally, although a quantitative assessment of uncertainty represents a great progress over the current evaluation practices of traditional soil surveys, it should be completed by an expert-based assessment that could check the plausibility of the predicted spatial patterns with regard to the available pedological knowledge.

5. Conclusion

Different evaluation sets obtained by probabilistic sampling were tested for their ability to assess the prediction uncertainty of DSM models using (as a case study) a spatial pattern of pseudo-values of topsoil clay content obtained from airborne hyperspectral imagery. The main lessons are summarized as follows:

- The spatial patterns of pseudo-values of some soil properties that could be available in some study areas across the world constitutes a relevant network for experimental assessments of the uncertainty of validation results. This is because i) it allows the DSM model to be evaluated by using many sites that could not be envisaged if only real soil data were used, and ii) it allows different numbers and locations of possible evaluation sets to be tested. Thus, it may provide an useful complement to the analytical

569 expressions (Brus et al, 2011) for the indicators and the many DSM applications for
570 which these analytical expressions are not valid.
571 • Any evaluation from independent sets conveys a non-negligible error on the
572 uncertainty indicators that is greater when the number of sites is low. Such evaluations
573 should therefore be interpreted with care and the uncertainty on validation results
574 must be systematically estimated.
575 • The sampling techniques used for the calibration and evaluation datasets should be
576 improved to reduce this error.
577

578 **6. Acknowledgments**

579

580 This research was conducted within the “Centre d’Expertise Scientifique Cartographie
581 Numérique des sols” granted by the CNES-TOSCA program. We thank David Rossiter and an
582 anonymous referee for their useful comments for improving the paper.
583

584 **7. References**

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

590 Ben-Dor E., Patkin K., Banin A., Karnieli A., 2002. Mapping of several soil properties using DAIS-
591 7915 hyperspectral scanner data – a case study over clayey soils in Israel. International
592 Journal of Remote Sensing, 23, p. 1043–1062

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

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

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

598 Cho, E, Cho, M.J., Eltinge, J., 2004. The variance of sample variance from a finite population.
599 Proceedings of Joint American Statistical Association and International Statistical Institute
600 Conference Toronto Canada

601 Cochrane, W.G., Snedecor, G.W., 1989. Statistical methods. Eigth edition. Iowa State
602 University Press, Ames (Iowa).

603 Cressie, N.J.A., 1993. Statistics for spatial data. Revised edition. Wiley interscience publication.
604 New York.

605 Gomez, C., Coulouma, G., Lagacherie, P., 2012a. Regional predictions of eight common soil
606 properties and their spatial structures from hyperspectral Vis–NIR data, Geoderma, 189–
607 190, 176-185.

608 Gomez, C., Lagacherie P., Bacha, S. 2012b. Using an VNIR/SWIR hyperspectral image to map
609 topsoil properties over bare soil surfaces in the Cap Bon region (Tunisia). In “Digital Soil
610 Assessments and Beyond” Minasny B., Malone B.P., McBratney A.B. (Ed.).Springer, 387-
611 392.

612 Gomez, C., Oltra Carrio, R., Lagacherie, P., Bacha, S., and Briottet, X. 2015. Sensitivity of soil
613 property prediction obtained from Hyperspectral Vis-NIR imagery to atmospheric effects
614 and degradation in image spatial resolutions. Remote Sensing of Environment 164, 1-15.

615 Goovaerts, P., 2001. Geostatistical modeling of uncertainty in soil science. Geoderma 103, 3–
616 26.

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

621 ISSS, ISRIC, FAO, 1998. World Reference Base for Soil Resources. 84 World Soil Resources
622 report. FAO Rome Italy.

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

626

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

629 Lagacherie, P., Baret, F., Feret, J.-B., Madeira Netto, J., Robbez-Masson, J.M., 2008. Estimation
630 of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral
631 measurements. Remote Sensing of Environment 112, 825–835. DOI
632 10.1016/j.rse.2007.06.014

633 Lagacherie, P., Gomez, C., in press. Vis-NIR-SWIR Remote Sensing Products as New Soil Data
634 for Digital Soil Mapping. in pedometrics A.B. McBratney et al. (eds.), Chapter 13. Progress
635 in Soil Science, Springer. DOI 10.1007/978-3-319-63439-5_13

636 Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2(3), 18--
637 22.

638 Luo, Y.Q., Randerson, J.T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P.,
639 Dalmonech, D., Fisher, J.B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F.,
640 Huntzinger, D., Jones, C.D., Koven, C., Lawrence, D., Li, D.J., Mahecha, M., Niu, S.L., Norby,
641 R., Piao, S.L., Qi, X., Peylin, P., Prentice, I.C., Riley, W., Reichstein, M., Schwalm, C., Wang,
642 Y.P., Xia, J.Y., Zaehle, S., Zhou, X.H., 2012. A framework for benchmarking land models.
643 Biogeosciences 9, 3857–3874.

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

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

647 Meinshausen, N., Schiesser, L., 2015. quantregForest: Quantile Regression Forests. R package.
648 https://cran.r-project.org.

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

651 Nussbaum, M., Papritz, A., Baltensweiler, A., Walthert, L., 2014. Estimating soil organic carbon
652 stocks of Swiss forest soils by robust external-drift kriging. Geosci. Model Dev. 7, 1197–
653 1210.

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

657 Pebesma, E.J., 2004. Multivariable geostatistics in S: the *gstat package*. *Computers &*
658 *Geosciences*, 30: 683-691.

659 R Development Core Team (2008). R: A language and environment for statistical computing.
660 R Foundation for Statistical Computing, Vienna, Austria.

661 Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M.,
662 Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria,
663 G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project
664 (AgMIP): Protocols and pilot studies. Agric. For. Meteorol. 170, 166–182.

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

667 Schwanghart, W., Jarmer, T., 2011. Linking spatial patterns of soil organic carbon to
668 topography — A case study from south-eastern Spain. Geomorphology 126, 252–263.

669 Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of
670 prediction interval for the model output. Neural Networks 19 (2), 225–235.

671 Stevens, A., Udelhoeven, T., Denis, A., Tychon, B., Lioy, R., Hoffman, L., Van Wesemael, B.,
672 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging
673 spectroscopy. *Geoderma* 158, 32-45.

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

675 Vaudour, E., Gilliot, J.M., Bel, L., Lefevre, J., Chehdi, K., 2016. Regional prediction of soil organic
676 carbon content over temperate croplands using visible near-infrared airborne
677 hyperspectral imagery and synchronous field spectra. Int. J. Appl. Earth Observ. Geoinf.,
678 49, 24-38. DOI: 10.1016/j.jag.2016.01.005

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

682 Zante, P., Collinet, J., Pepin, Y., 2005. Caractéristiques pédologiques et hydrométéorologiques
683 du bassin versant de Kamech, Cap Bon, Tunisie. *UMR LISAH IRD Tunis, DG ACTA Direction*
684 *des Sols Tunis, INRGREF Tunis*. (21 p. + 6 annexes).

685

686