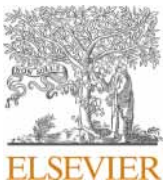
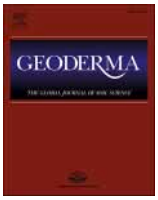
Geoderma 433 (2023) 116467



Contents lists available at ScienceDirect

Geoderma

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



Integrating additional spectroscopically inferred soil data improves the accuracy of digital soil mapping

Songchao Chen [a,b,c,*], Nicolas P.A. Saby [c], Manuel P. Martin [c], Bernard G. Barthès [d], Cécile Gomez [e], Zhou Shi [b], Dominique Arrouays [c]

[a] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China
[b] Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] INRAE, Unité Info&Sol, Orléans 45075, France
[d] Eco&Sols, University of Montpellier, CIRAD, INRAE, IRD, Institut Agro, Montpellier 34060, France
[e] LISAH, University of Montpellier, IRD, INRAE, Institut Agro, Montpellier 34060, France

ARTICLE INFO

Handling Editor: Morgan Cristine L.S.

Keywords:
Proximal soil sensing
Vis-NIR spectroscopy
MIR spectroscopy
Digital soil mapping
Measurement error

ABSTRACT

Digital soil mapping has been increasingly advocated as an efficient approach to deliver fine-resolution and up-to-date soil information in evaluating soil ecosystem services. Considering the great spatial heterogeneity of soils, it is widely recognized that more representative soil observations are needed for better capturing the soil spatial variation and thus to increase the accuracy of digital soil maps. In reality, the budget for the field work and soil laboratory analysis is commonly limited due to its high cost and low efficiency. In the last two decades, being an alternative to wet chemistry, soil spectroscopy, such as visible-near infrared (Vis-NIR), mid-infrared (MIR) spectroscopy has been developed in measuring soil information in a rapid and cost-effective manner and thus enable to collect more soil information for digital soil mapping (DSM). However, spectroscopically inferred (SI) data are subject to higher uncertainties than reference laboratory analysis. Many DSM practices integrated SI data with soil observations into spatial modelling while few studies addressed the key question that whether these non-errorless soil data improve map accuracy in DSM. In this study, French Soil Monitoring Network (RMQS) and Land Use and Coverage Area frame Survey Soil (LUCAS Soil) datasets were used to evaluate the potential of SI data from Vis-NIR and MIR in digital mapping of soil properties (i.e. soil organic carbon, clay, and pH) at a national scale. Cubist and quantile regression forests were used for spectral predictive modelling and DSM modelling, respectively. For both RMQS and LUCAS Soil dataset, different scenarios regarding varying proportions of SI data and laboratory observations were tested for spectral predictive models and DSM models. Repeated (50 times) external validation suggested that adding additional SI data can improve the performance of DSM models regardless of soil properties (gain of R^2 proportion at 3–19%) when the laboratory observations are limited ($\leq$50%). Lower proportion of SI data used in DSM model and higher accuracy of spectral predictive models led to greater improvement of DSM. Our results also showed that a greater proportion of SI data lowered the prediction intervals which may result in an underestimation of prediction uncertainty. The determination of accuracy threshold on SI data for the use in DSM needs to be explored in future studies.

1. Introduction

In the 21st century, soils are at the nexus for ensuring ecosystem services and achieving sustainable development goals (McBratney et al., 2014; Keesstra et al., 2016; Bouma et al., 2019). Up-to-date and fine-resolution soil information is urgently needed to support relevant scientific research and evidence-based decision-making (Sanchez et al., 2009; Arrouays et al., 2014). As the way to produce conventional soil maps is rather labor- and cost-intensive, time-consuming, and hard to be updated, digital soil mapping (DSM, McBratney et al., 2003) has been developed based on Jenny's soil-forming theory (Jenny, 1941). Under the conceptual framework of *Scorpan*, soil classes or soil attributes can be predicted at unvisited positions by their relationships to environmental covariates, such as other soil information, climate, organisms, relief, parent materials, age and spatial position. With the significant advances in Geographic Information Systems, remote sensing,

* Corresponding author at: ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China.
E-mail address: chensongchao@zju.edu.cn (S. Chen).

https://doi.org/10.1016/j.geoderma.2023.116467
Received 29 November 2021; Received in revised form 29 March 2023; Accepted 3 April 2023
Available online 5 April 2023

geostatistics, machine learning, and high-performance computing capacity, the DSM technique has been increasingly used for delivering high-resolution and recent soil information across scales over the last two decades (e.g., Minasny and McBratney, 2016; Arrouays et al., 2017; Chen et al., 2022; Liu et al., 2022; Zhang et al., 2023).

Under the same sampling protocol, it is widely admitted that a larger number of soil samples could capture more spatial heterogeneity, thus improving the map accuracy when using the DSM technique. However, constrained by the budget for the soil measurement, the number of soil samples would be greatly limited when using standard laboratory analysis, which is usually expensive and time-consuming (Stenberg et al., 2010; Viscarra Rossel et al., 2016). As an alternative, soil spectroscopic techniques, such as visible-near infrared (Vis-NIR, 350–2500 nm) and mid-infrared (MIR, 400–4000 cm^{-1}) spectroscopy, have shown great potential in measuring soil information (e.g., clay, soil organic carbon, iron content) in a rapid and cost-effective manner under laboratory condition (e.g., Grinand et al., 2012; Stevens et al., 2013; Shi et al., 2014; Nocita et al., 2015; Ji et al., 2016; Demattê et al., 2019; Chen et al., 2020b). In light of the advantages shown by soil spectroscopic techniques, soil surveyors can afford greater soil sampling density to provide a better understanding of soil spatial variation for DSM practices (Somarathna et al., 2018).

As shown in Table 1, SI soil data from Vis-NIR or MIR measurement has been used as a data source together with laboratory soil observations in DSM from field to national scales in previous studies. Somarathna et al. (2018) and Wadoux et al. (2019) noted that SI soil data has a greater measurement error, which is associated with the spectral predictive model, than the laboratory analysis. Here comes a question: can additional SI soil data improve map accuracy in DSM? Under the context of the on-going build-up of soil spectral libraries worldwide, this issue is becoming critical to evaluate the practical potential of SI data in DSM. Nevertheless, most relevant studies (Table 1) have not addressed this issue, while the results from Somarathna et al. (2018) and Paul et al. (2019) were rather opposite. Therefore, more efforts are still needed to answer the question mentioned above. To this end, the French Soil Monitoring Network (RMQS) (Jolivet et al., 2006) and Land Use and Coverage Area frame Survey Soil (LUCAS Soil) (Tóth et al., 2013) datasets where both laboratory physico-chemical measurements and spectral data are available for more than 2000 sites in France, were used to assess the potential of SI data in mapping soil organic carbon (SOC), clay and pH in mainland France. Using 50 times repeated consistent external validation sets from two datasets, we assessed (1) whether the ratio share between SI data and laboratory observations will affect the map accuracy and uncertainty estimates? (2) whether the results are specific for different soil properties and datasets?

2. Materials and methods

2.1. Soil data and environmental covariates

Considering the potential effect of sampling design, the RMQS (grid sampling) and LUCAS Soil (stratified sampling) datasets were tested in this study (Fig. 1).

The sites of RMQS program are located following a systematic square grid of 16 km which resulted in around 2,200 sampling sites collected from 2001 and 2009 covering France under different pedo-climatic, relief, and land cover conditions (Jolivet et al., 2006). From a 20 m square located at the centre of each 16 km grid, topsoil (0–30 cm) and subsoil (30–50 cm) layers were collected by merging 25 sub-samples based on an unaligned sampling design (Chen et al., 2018a). After air-drying and sieving to <2 mm, SOC was measured by the dry combustion method using an automated C:N analyzer, pH was determined in a 1:5 soil:water mixture (AFNOR, 1994), and clay (0–2 μm) was measured with the pipette method (AFNOR, 2003). A total of 2036 topsoil samples

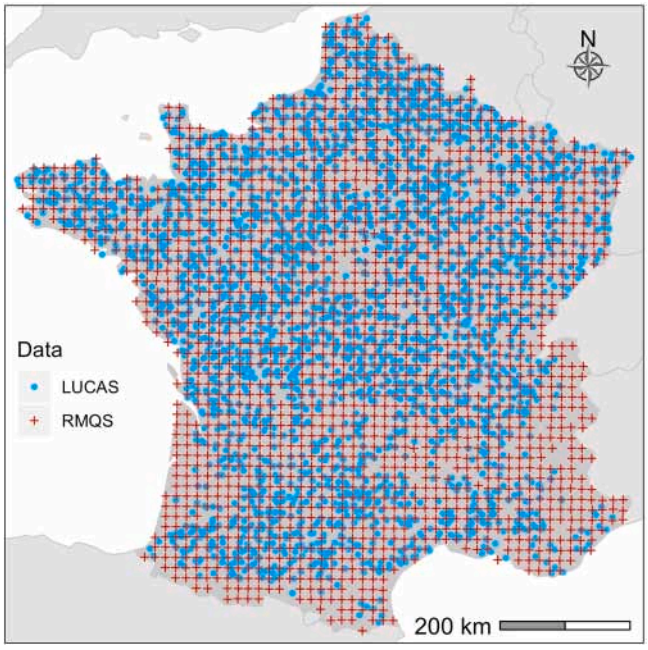


Fig. 1. Sampling sites from RMQS (red cross) and LUCAS Soil (blue point) located in mainland France.

Table 1
Summary of previous studies integrating spectroscopically (vis-NIR and/or MIR) inferred data in DSM.

Reference	Country	Scale	Soil property	No. samples (Obs/Spec)*	Accuracy improvement
Cambule et al. (2013)	Mozambique	Regional	SOC	104/281	NA
Viscarra Rossel et al. (2014)	Australia	National	SOC stock	4487/1101 sites	NA
Viscarra Rossel et al. (2015)	Australia	National	Sand, silt, clay, BD, SOC, TN, TP, pH, ECEC, AWC	NA	NA
Knadel et al. (2015)	Demark	Field	SOC, clay, silt, sand	30/12000	NA
Priori et al. (2016)	Italy	Field	SOC stock	36/176	NA
Ramifehiarivo et al (2017)	Madagascar	National	SOC	NA	NA
Somarathna et al. (2018)	Australia	Regional	SOC	681/998 (Topsoil) 43/987 (Subsoil)	No
Wadoux et al. (2019)	Australia	Regional	TOC	645/1743	NA
Paul et al. (2019)	Canada	Field	Sand, silt, clay, pH, EC, SOM, TN	62/308	Yes
Gray et al. (2019)	Australia	Regional	SOC fractions	427/372 profiles	NA
Zhang et al. (2020)	Canada	Local	SOM, clay, soil moisture	32/148 profiles	NA
Chatterjee et al. (2021)	USA	Local	SOC, pH, clay, silt, sand, TN	25/25 profiles	NA
Ma et al. (2021)	Australia	Regional	SOC	100/2682	NA
Sanderman et al. (2021)	USA	Regional	SOC fractions	659/8500	NA
Filippi et al. (2021)	Australia	Regional	SOC	353/118	NA
Takoutsing et al. (2022b)	Cameroon	Regional	SOC, clay, pH	48/432	NA

* Obs and Spec represent for soil observations (laboratory analysis) and spectroscopically inferred soil data (spectral prediction) respectively.

located in mainland France were used in this study.

The LUCAS Soil data was sampled in 2009, and it contains a total of 19,967 samples collected in topsoil (0–20 cm) from 25 EU member states by stratified random sampling according to topography and land use (Tóth et al., 2013). All the soil samples were air-dried and sieved to a fraction of <2 mm, and then SOC, pH and clay (0–2 μm) were determined using ISO standard methods (Stevens et al., 2013). In this study, 2685 LUCAS Soil samples located in mainland France were used.

Fig. 2 shows the statistics of SOC, clay and pH in the RMQS and LUCAS Soil. SOC had similar distributions (Mann-Whitney test, $p = 0.76$) in the RMQS and LUCAS Soil with 1st quartile (Q1) of 13.1 and 13.2 g kg^{-1}, median of 19.5 and 19.8 g kg^{-1}, 3rd quartile (Q3) of 30.3 and 30.70 g kg^{-1}, and mean of 25.5 and 24.4 g kg^{-1}, respectively. The RMQS had a slightly wider SOC range (1.5 to 243.0 g kg^{-1}) than the LUCAS Soil (2.2 to 165.7 g kg^{-1}) because the first LUCAS Soil survey did not fully cover the mountainous regions (e.g., the Alps, the Pyrénees) shown in Fig. 1. Significant differences ($p = 0.02$) were observed for clay between RMQS and LUCAS Soil (Q1 of 15.3 and 15.0%, median of 21.2 and 21.0%, Q3 of 32.2 and 30.0%, and mean of 24.6 and 23.2 %, range of 0.2–81.5% and 2.0–77.0%). Divergent distributions ($p < 0.001$) were found for pH between RMQS and LUCAS Soil, resulting from the greater proportion of forest samples in RMQS. With similar pH ranges, the RMQS had a much larger interquartile range (Q3-Q1, 2.4) than the LUCAS Soil (1.78), and its median pH (6.2) was also much lower than that of the LUCAS (6.75).

Based on previous DSM studies in France, fifteen environmental covariates (Table 2) related to five *Scorpan* factors, namely, soil, climate, organisms, relief, parent material, were used in this study (Mulder et al., 2016; Chen et al., 2019; Chen et al., 2020a). These environmental covariates were reprojected to the Lambert 93 coordinate system and resampled into 90 m resolution raster images.

Table 2
Environmental covariates used for DSM.

Covariates	Scorpan factors	Resolution/ scale	Reference
Soil type	Soil	1:1 M	IUSS Working Group WRB (2006)
Erosion rates	Soil	1:1 M	Cerdan et al. (2010)
Mean Annual Precipitation	Climate	1 km	Hijmans et al. (2005)
Mean Annual Temperature	Climate	1 km	Hijmans et al. (2005)
Net Primary Production	Organisms	1 km	NASA LD (2001)
Corine Land Cover 2006	Organisms	250 *m*	Feranec et al. (2010)
SRTM DEM	Relief	90 *m*	Jarvis et al. (2008)
Aspect	Relief	90 *m*	Jarvis et al. (2008)
Slope cosines	Relief	90 *m*	Jarvis et al. (2008)
Curvature	Relief	90 *m*	Jarvis et al. (2008)
Exposition	Relief	90 *m*	Jarvis et al. (2008)
Roughness	Relief	90 *m*	Jarvis et al. (2008)
Compound Topographic Index	Relief	90 *m*	Jarvis et al. (2008)
Topographic Wetness Index	Relief	90 *m*	Jarvis et al. (2008)
Parent material	Parent material	1:1 M	King et al. (1995)

2.2. Spectral measurement, pre-processing and modelling

For the RMQS dataset, soil samples were further sieved to <0.2 mm before measuring MIR absorbance spectra using a Nicolet 6700 Diffusive Reflectance Fourier Transform Spectrophotometer (Thermo Fisher Scientific Inc., USA) (Grinand et al., 2012). The MIR spectra ranged from 4000 to 400 cm with a spectral resolution of 3.86 cm. Thirty-two scans of spectra were recorded for each soil sample, and the average was taken as the representative spectra. The Vis-NIR absorbance spectra (350–2500 nm with a special resolution of 1 nm) for RMQS dataset were measured on the soil samples sieved to <1 mm using a NIRSystems 6500 spectrophotometer (Foss Analytical, Sweden) (Gogé et al., 2012; Gogé et al., 2014). The spectral ranges of 350–399 nm and 2451–2500 nm were removed due to their low signal-to-noise ratio. Based on previous studies on spectral modelling for RMQS dataset and after initial comparison, Savitzky-Golay algorithm (window size of 11 and 21 points for MIR and Vis-NIR, and a 2nd order polynomial) with the 1st derivative followed by Standard Normal Variate was chosen to smooth both MIR and Vis-NIR spectra and enhance the signal. The processed MIR and Vis-NIR spectra were resampled to a spectral resolution of 20 cm^{-1} and 10 nm, respectively, for speeding the computation efficiency while not losing the predictive performance (Yang et al., 2012).

While in the LUCAS Soil database, the vis–NIR absorbance spectra were measured on the air-dried and sieved (<2 mm) soil samples using a FOSS XDS rapid content analyzer (FOSS NIRSystems Inc., Denmark) (Nocita et al., 2014). The Vis-NIR spectra had a range of 400–2500 nm with a spectral resolution of 0.5 nm. The 400 to 500 nm spectra were removed due to instrument artefacts (Stevens et al., 2013). The Savitzky–Golay algorithm (window size of 101 points and a 2nd order polynomial) with the 1st derivative was adopted for spectral pre-processing as suggested by previous studies (Notica et al., 2014; Chen et al., 2021b).

Cubist is a commonly used spectroscopic modeling approach to modelling Vis-NIR and MIR data (Viscarra Rossel et al., 2017; Liu et al., 2019). Cubist iteratively splits the target variable into several partitions within which the predictor variables are similar. Within a given

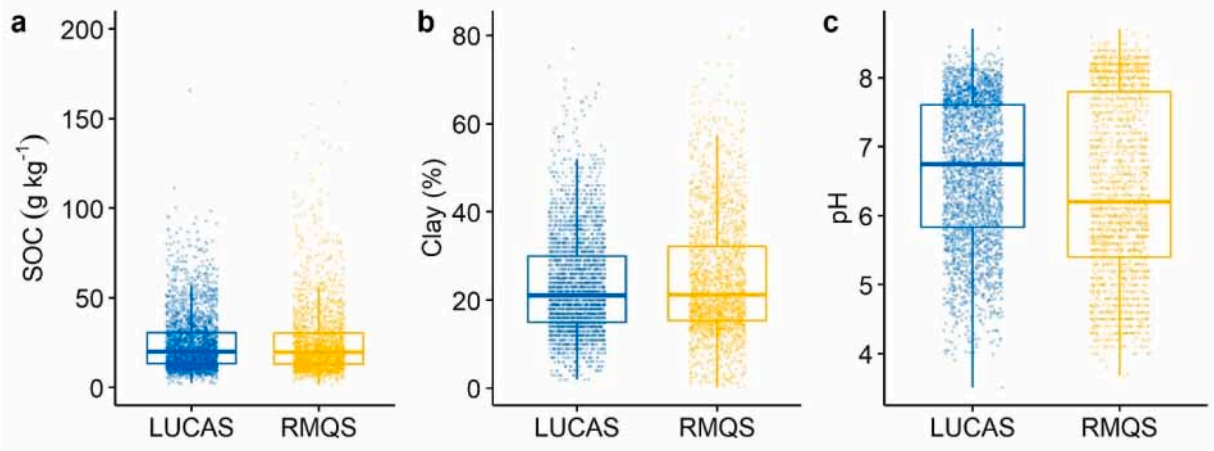


Fig. 2. Boxplots of laboratory-measured SOC (a), clay (b) and pH (c) in RMQS and LUCAS Soil located in mainland France.

partition, the standard deviation of the target variable is used as a node splitting criterion. The split that maximizes the reduction in the standard deviation is chosen in Cubist. Afterwards, pruning and smoothing processes are conducted to build the final model. We refer to Quinlan (1992) for more details. A list of hierarchically structured rules defines partitions in the final Cubist model. The form of the rule is listed as below:

IF {condition}, **THEN** {linear regression model}.

ELSE {apply next rule}.

Here, the linear regression model is applied to predict the target variable when the case satisfies the condition defined in the rule. The "Cubist" package (Kuhn and Quinlan, 2020) was used for Cubist modelling in R (R Core Team, 2019).

The measurement error variances of spectral predictions were estimated from residual variance of the Cubist model (Takoutsing and Heuvelink, 2022). The Cubist residual variance was estimated by the equation below:

$$var_{Cubist} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\widehat{y}_i - y_i\right)^2\right) \quad (1)$$

where n is the number of samples to be predicted by Cubist model, y_i and $\widehat{y}_i$ are the measured and spectral predicted values for sample i.

2.3. Predictive model for DSM

Quantile Regression Forests (QRF, Meinshausen, 2006) has been increasingly used for DSM modelling from regional to global scales as it enables to provide uncertainty estimates at users' defined prediction intervals (e.g., 90%) with reasonable accuracy (e.g., Vaysse and Lagacherie, 2017; Loiseau et al., 2019; Nauman and Duniway, 2019; Chen et al., 2021a; Kasraei et al., 2021; Poggio et al., 2021).

We define X and Y as the predictor variables and target variables, QRF generates a large number of trees (b) using bootstrapping (random sampling with replacement) from p training samples (X_i, Y_i), $i = 1, \ldots, p$. A random subset of the predictor variables is then used to select split-point for each node of the bootstrap tree. For a new sample $N = X_n$, its prediction for each bootstrap tree is the conditional mean estimate of Y. The mean predictions of b bootstrap trees are used to represent the final prediction of the new sample N. Using the weighted samples, QRF can also derive a conditional distribution from which the probability of Y being lower than a given percentile can be determined and thus to calculate the prediction intervals. We refer to Meinshausen (2006) for more details relevant to the calculation of conditional distribution. The number of trees (num.trees) was set to 500 which was large enough to generate a stable model performance. Based on 5-fold cross-validation, number of variables to possibly split in each node (mtry) and minimal node size (min.node.size) were optimized to 4 and 5 respectively for both RMQS and LUCAS Soil. The "caret" (Kuhn, 2020) and "ranger" (Wright and Ziegler, 2017) packages were used for optimizing and running QRF in R (R Core Team, 2019).

Since the spectral prediction are subjected to prediction error, the measurement error-filtered QRF (MEF-QRF) approach proposed by van der Westhuizen et al. (2022) was used for the measurement error of SI soil data. In traditional QRF, each calibration sample has the same weight in model training process because all the calibration samples are treated as error-free and the residual variance of QRF model is assumed as constant. Therefore, the QRF model is trained by minimizing the sum of squared prediction errors (SSPE) of the calibration data, and each calibration sample has the same contribution to SSPE. In MEF-QRF, the measurement error variance is included in the loss function and each calibration sample is assigned a weight by the inverse of the sum of the residual variance and the measurement error variance. In this study, we regarded the laboratory observations as errorless, so the measurement error variance was set to 0 for laboratory observations. While for the SI data, measurement error variance for each soil property was estimated from the residual variance of Cubist model on external validation data according to Takoutsing and Heuvelink (2022), Takoutsing et al. (2022) (Eq. (1)). Proposed by van der Westhuizen et al. (2022), the residual variance was estimated by an iterative procedure: (1) calibrate the standard QRF to estimate the model parameters; (2) estimate the residual variance from standard QRF using the conditional log-likelihood; (3) update the model parameters by minimizing the weighted sum of the squared residuals; (4) repeat step 1 to step 3 until convergence. For more details about the MEF-QRF, we refer to van der Westhuizen et al. (2022). Once the weight is assigned for each calibration sample, the "case. weights" in the "ranger" R package can be used to set the probability of each calibration sample to be used in the bootstrapping procedure. That means a calibration sample with a high measurement error will have a low probability to be sampled in the bootstrapping.

In summary, a traditional QRF was used when all the calibration data were laboratory observations while MEF-QRF was applied when integrating SI data with laboratory observations.

2.4. Model evaluation

Modeling efficiency (R^2), which indicates the amount of explained variance, and root mean square error (RMSE) were used to evaluate both spectral predictive models and DSM models. Prediction interval coverage percentage (PICP) and prediction interval width (PIW) were used to validate the quantifications of uncertainty for DSM models. PICP describes the percentage of the validation observations located between pre-defined lower and upper limits of PIs. As 90% PIs was used in this study, we should expect PICP around 90%. PIW (95% quantile minus 5% quantile) gives a measure of absolute model uncertainty and a greater PIW indicates a higher model uncertainty. The average of these indicators calculated from 50 repeats was taken for final model evaluation. As mentioned in the general framework (section 2.2), model evaluation in both spectral predictive models and DSM models was performed on the consistent external validation data (D2, see more details in section 2.5).

$$R^2 = 1 - \frac{\sqrt{\sum_{i}^{n}\left(\widehat{y}_i - y_i\right)^2}}{\sqrt{\sum_{i}^{n}\left(y_i - \overline{y}\right)^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2} \quad (3)$$

where n is the number of external validation samples, y_i and $\widehat{y}_i$ are the measured and predicted values for external validation sample i, and $\overline{y}$ is the mean of measured values for all external validation samples.

2.5. General framework

The general framework of this study is present in Fig. 3 (created with BioRender.com), which involves five steps as follows.

(1) RMQS and LUCAS Soil were both randomly divided into D1 (75%) and D2 (25%). Here D2 was used as consistent external data to evaluate the accuracy of spectral predictive models and DSM models while D1 was used to create four scenarios (see details in step 2). This random split was repeated 50 times for a robust evaluation (Xiao et al., 2022);

(2) Based on D1 dataset, four scenarios were created to evaluate the effect of ratio share between SI data and laboratory observations. Scenario 0 was a benchmark where no data was SI and all D1 was used for DSM modelling. The proportion of soil samples assumed with spectra only (to be predicted by the spectroscopic model) increased from 1/3 in scenario 1 (S1U) to 1/2 in scenario 2 (S2U) and 2/3 in scenario 3 (S3U) by random sampling (50 repeats as

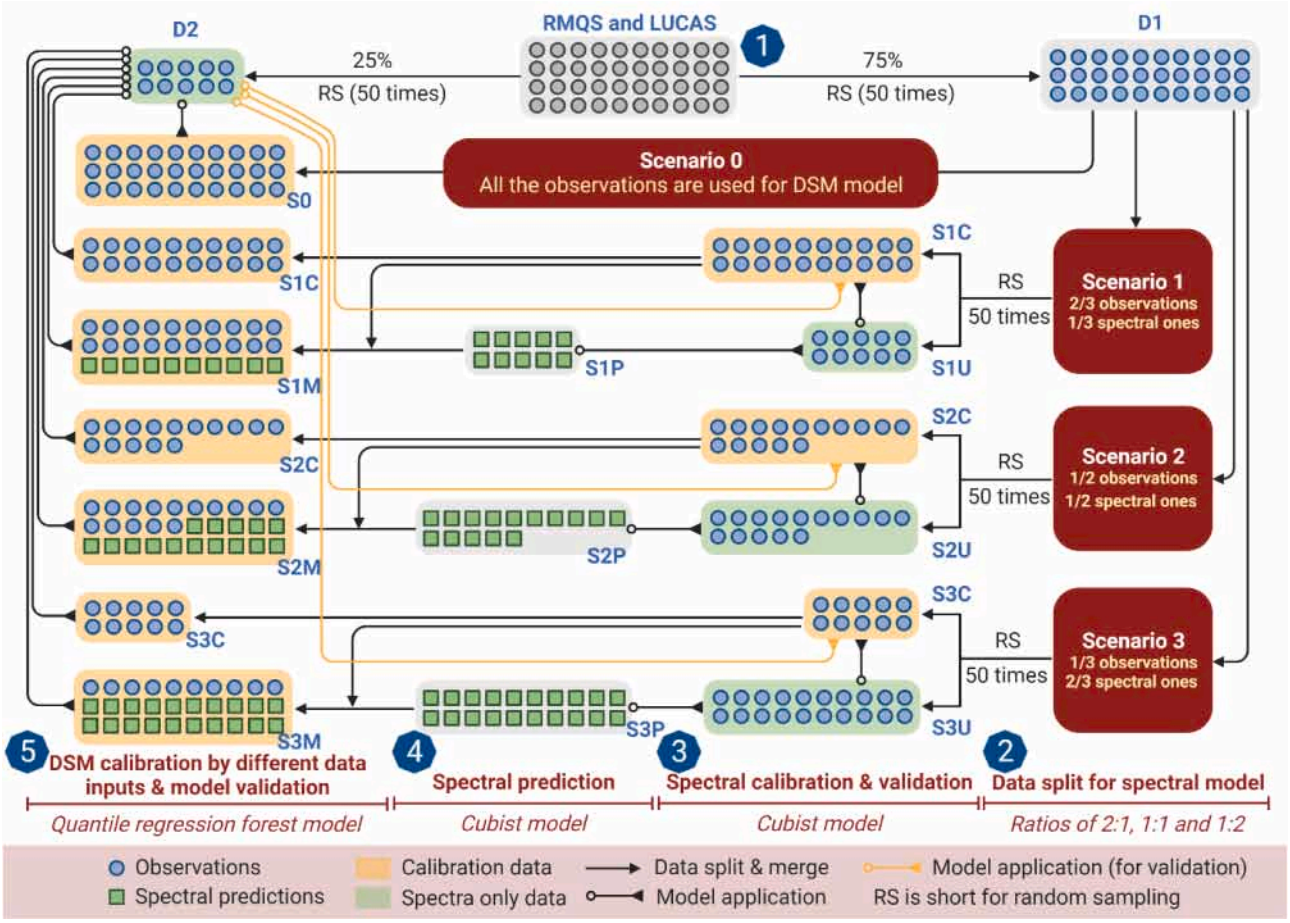


Fig. 3. General workflow of this study. Five steps are numbered in the octagons. The data included in different scenarios are listed below: S0 (all observations), S1M (2/3 observations and 1/3 spectroscopically inferred data), S1C (2/3 observations), S2M (1/2 observations and 1/2 spectroscopically inferred data), S2C (1/2 observations), S3M (1/3 observations and 2/3 spectroscopically inferred data), S3C (1/3 observations).

we did in step 1), while the remaining data in D1 was used to the calibrate the spectral models accordingly (S1C, S2C and S3C);

(3) The spectral predictive models were calibrated using Cubist model, and the model performance was evaluated by the consistent external data D2;

(4) The soil samples assumed with spectra only were predicted by relevant Cubist models (S1P, S2P, and S3P);

(5) In scenarios 1, 2 and 3, two strategies were compared. Strategy 1 only used a part of the laboratory observations (2/3, 1/2 and 1/3 of observations for S1C, S2C, and S3C, respectively) for DSM modelling, while strategy 2 used both laboratory observations and SI data (S1M, S2M, and S3M). It should be noted that S1M, S2M and S3M had the sample size to S0, while S1C, S2C, and S3C comprised of 2/3, 1/2 and 1/3 samples of S0. This allowed us to check whether the additional SI data can improve the DSM model. QRF was used for while DSM modelling, and the model performance was evaluated by the consistent external data D2. Please note that traditional QRF was used when all the data was laboratory observations while the MEF-QRF was used when including SI data with laboratory observations.

3. Results

3.1. Performance evaluation for spectral models

Based on the consistent external validation set D2 (Fig. 3), the performance of spectral models for SOC, clay, and pH under different scenarios is shown in Table 3. Regarding spectral modelling, the

Table 3
The performance (in R^2 and RMSE) of spectral models for SOC, clay, and pH under different scenarios using data from RMQS and LUCAS Soil located in mainland France. The proportion of calibration samples decreases from 2/3 of all laboratory observations in Scenario 1 to 1/3 in Scenario 3. The numbers in brackets indicates the width of 90% confidence intervals of 50 repeats. The units of RMSE for SOC and clay are g kg^{-1} and %.

Dataset	Soil property	Scenario 1		Scenario 2		Scenario 3	
		R^2	RMSE	R^2	RMSE	R^2	RMSE
RMQS Vis-NIR	SOC	0.73 (0.11)	10.32 (3.60)	0.68 (0.20)	11.20 (4.22)	0.67 (0.18)	11.42 (4.12)
	Clay	0.69 (0.11)	7.35 (1.38)	0.68 (0.11)	7.56 (1.43)	0.65 (0.12)	7.85 (1.43)
	pH	0.81 (0.05)	0.57 (0.07)	0.80 (0.06)	0.58 (0.08)	0.78 (0.05)	0.62 (0.08)
RMQS MIR	SOC	0.93 (0.05)	5.26 (2.42)	0.93 (0.05)	5.53 (2.63)	0.91 (0.07)	5.96 (2.41)
	Clay	0.88 (0.05)	4.61 (0.90)	0.87 (0.05)	4.79 (0.96)	0.85 (0.04)	5.12 (0.82)
	pH	0.92 (0.03)	0.37 (0.07)	0.91 (0.03)	0.38 (0.07)	0.90 (0.03)	0.41 (0.07)
LUCAS	SOC	0.71 (0.08)	8.51 (1.26)	0.69 (0.11)	8.81 (1.78)	0.67 (0.12)	9.05 (1.49)
	Clay	0.70 (0.09)	5.96 (0.98)	0.68 (0.07)	6.20 (0.67)	0.63 (0.06)	6.63 (0.67)
	pH	0.87 (0.03)	0.37 (0.04)	0.86 (0.03)	0.38 (0.04)	0.85 (0.03)	0.41 (0.03)

performance for both RMQS (Vis-NIR and MIR) and LUCAS Soil (Vis-NIR) showed a slightly decreasing trend when the spectral calibration size reduced from scenario 1 to scenario 3, and this decreasing trend was more evident for LUCAS Soil and RMQS Vis-NIR data. The spectral predictive models for the RMQS MIR data (R^2 of 0.91–0.93, 0.85–0.88, 0.90–0.92 and RMSE of 5.26–5.96 g kg^{-1}, 4.61–5.12%, 0.37–0.41 for SOC, clay and pH, respectively) performed better than these for the RMQS Vis-NIR (R^2 of 0.67–0.73, 0.65–0.69, 0.78–0.81 and RMSE of 10.32–11.42 g kg^{-1}, 7.35–7.85%, 0.57–0.62 for SOC, clay and pH, respectively) and LUCAS Soil (R^2 of 0.67–0.71, 0.63–0.70, 0.85–0.87 and RMSE of 8.51–9.05 g kg^{-1}, 5.96–6.63%, 0.37–0.41 for SOC, clay and pH, respectively).

3.2. Performance evaluation for DSM models

Table 4 presents the performance of the DSM model for SOC, clay, and pH under different scenarios. When using all the soil observations for DSM modelling (S0), the performance of the RMQS data (R^2 of 0.41, 0.36 and 0.53 for SOC, clay and pH respectively) was generally better than the LUCAS Soil (R^2 of 0.33, 0.27 and 0.46 for SOC, clay and pH, respectively). Table 4 also shows that, regardless of datasets (i.e. LUCAS Soil, RMQS) and soil properties (i.e. SOC, clay, pH), including additional SI data to laboratory observations (i.e. S1M, S2M, S3M) in DSM models slightly improved performance (gain of R^2 proportion at 3–19%) when compared to these models using only equal-sized laboratory observations (i.e. S1C, S2C, S3C). In addition, these improvements resulting from additional SI data were more evident for both RMQS and LUCAS Soil datasets when more SI data and less laboratory observations were used in modelling from S1M (2/3 observations and 1/3SI) to S2M (1/3 observations and 2/3SI). It is also clear that these models using both laboratory observations and SI data (i.e. S1M, S2M, S3M) had close performance to those using all the laboratory observations (S0) while a slight decrease of performance can be seen when the proportion of SI data increased (e.g. S1M, S2M, S3M for RMQS Vis-NIR in clay prediction). When comparing the two spectroscopic techniques for RMQS dataset, we found that the DSM models including MIR inferred predictions always performed better than models including Vis-NIR inferred predictions (i.e. S1M, S2M, S3M).

3.3. Uncertainty qualification for DSM models

As shown in Table 5, all the median PICP indicators derived from QRF models were located between 86.8% and 94.4%, which were quite close to the pre-defined 90% PIs. The predictive models of LUCAS Soil using all the soil observations (S0) for DSM modelling had similar uncertainty quantifications to those of RMQS data for SOC and clay, and a slightly greater PICP (94.3%) was found for RMQS than that of LUCAS Soil (92.2%). The PICP for the DSM models using different proportions of soil observations (i.e. S1C, S2C, S3C) were close to the DSM models using all the soil observations (S0) for both RMQS and LUCAS Soil datasets, and a slightly decreasing PICP trend was found when less soil observations were used for DSM (S1C to S2C and then to S3C). When the proportion of RMQS MIR inferred data was low (S1M), the PICP for these DSM models was quite close to S0 (ΔPICP <0.7%), while a greater decrease of PICP (ΔPICP of 0.7–1.7%) was found for the DSM models with higher proportion of RMQS MIR inferred data (S2M, S3M). In comparison to S0, RMQS Vis-NIR related DSM models showed a greater decrease of PICP (ΔPICP of 0.5–4.0%), and this decreasing trend was more evident for the DSM models with more Vis-NIR inferred data (S3M). LUCAS Soil dataset showed a similar pattern of PICP to RMQS Vis-NIR with a greater decreasing trend (ΔPICP of 0.7–5.9%). The most significant decrease of PICP for LUCAS Soil and RMQS (Vis-NIR and MIR) datasets was found for these DSM models in predicting clay.

As shown in Table 5, PIW gradually increased when the number of observations decreased from S0 (all observations) to S1C (2/3 observations) and then to S3C (1/3 observations), including an increasing model uncertainty. When gradually increasing the proportion of SI data increased from S1M (2/3 observations and 1/3SI) to S3M (1/3 observations and 2/3SI), it was evident that PIW showed a decreasing trend, implying a deceasing absolute model uncertainty.

3.4. The difference of spatial pattern for soil properties

To better visualize the effect of additional SI data on the spatial distribution of soil properties using DSM modes, we present the SOC, clay and pH maps for S0 and their difference to other three scenarios (S1, S2, S3) using data from RMQS and MIR inferred data in Fig. 4, Fig. 5 and Fig. 6 (maps for RMQS Vis-NIR and LUCAS are present in Figs. S1 to S6).

Table 4
Model performance of DSM models for SOC, clay, and pH under different scenarios using data from RMQS and LUCAS Soil located in mainland France. The numbers in brackets indicates the width of 90% confidence intervals of 50 repeats. The units of RMSE for SOC and clay are g kg^{-1} and %. The data included in different scenarios are listed below: S0 (all observations), S1M (2/3 observations and 1/3 spectroscopically inferred data), S1C (2/3 observations), S2M (1/2 observations and 1/2 spectroscopically inferred data), S2C (1/2 observations), S3M (1/3 observations and 2/3 spectroscopically inferred data), S3C (1/3 observations).

Dataset	Soil property	S0		S1M		S1C		S2M		S2C		S3M		S3C	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
RMQS Vis-NIR	SOC	0.41 (0.12)	15.39 (5.35)	0.39 (0.12)	15.67 (5.71)	0.39 (0.12)	15.68 (5.28)	0.38 (0.10)	15.78 (5.80)	0.38 (0.10)	15.80 (5.49)	0.38 (0.10)	15.89 (5.75)	0.37 (0.14)	15.95 (5.28)
	Clay	0.36 (0.10)	10.64 (1.64)	0.35 (0.10)	10.75 (1.61)	0.35 (0.10)	10.75 (1.63)	0.34 (0.11)	10.82 (1.66)	0.33 (0.11)	10.90 (1.59)	0.33 (0.11)	10.92 (1.70)	0.30 (0.12)	11.08 (1.60)
	pH	0.53 (0.08)	0.90 (0.08)	0.52 (0.08)	0.91 (0.09)	0.51 (0.09)	0.92 (0.09)	0.52 (0.08)	0.91 (0.08)	0.50 (0.08)	0.93 (0.08)	0.51 (0.09)	0.92 (0.09)	0.47 (0.09)	0.95 (0.09)
RMQS MIR	SOC	0.41 (0.12)	15.39 (5.35)	0.40 (0.11)	15.59 (5.47)	0.39 (0.12)	15.68 (5.28)	0.39 (0.11)	15.73 (5.67)	0.38 (0.13)	15.80 (5.49)	0.39 (0.11)	15.69 (5.55)	0.37 (0.14)	15.95 (5.28)
	Clay	0.36 (0.10)	10.64 (1.64)	0.35 (0.10)	10.69 (1.58)	0.35 (0.10)	10.75 (1.63)	0.35 (0.11)	10.71 (1.64)	0.33 (0.11)	10.90 (1.59)	0.35 (0.10)	10.72 (1.54)	0.30 (0.12)	11.08 (1.60)
	pH	0.53 (0.08)	0.90 (0.08)	0.52 (0.08)	0.91 (0.08)	0.51 (0.09)	0.92 (0.09)	0.52 (0.08)	0.91 (0.09)	0.50 (0.08)	0.93 (0.08)	0.52 (0.08)	0.91 (0.09)	0.47 (0.09)	0.95 (0.09)
LUCAS Soil	SOC	0.33 (0.09)	12.90 (1.98)	0.32 (0.08)	12.98 (2.03)	0.32 (0.08)	13.00 (2.02)	0.32 (0.07)	13.01 (1.92)	0.31 (0.09)	13.07 (1.87)	0.31 (0.06)	13.06 (2.00)	0.30 (0.08)	13.16 (2.06)
	Clay	0.27 (0.08)	9.35 (1.06)	0.26 (0.07)	9.40 (1.10)	0.25 (0.07)	9.48 (1.04)	0.26 (0.07)	9.43 (1.12)	0.24 (0.08)	9.57 (1.06)	0.25 (0.07)	9.49 (1.14)	0.23 (0.09)	9.65 (1.20)
	pH	0.46 (0.06)	0.77 (0.06)	0.46 (0.06)	0.77 (0.05)	0.45 (0.07)	0.78 (0.05)	0.46 (0.06)	0.77 (0.06)	0.43 (0.07)	0.79 (0.05)	0.46 (0.07)	0.77 (0.06)	0.41 (0.07)	0.80 (0.06)

Table 5
Uncertainty qualification of DSM models for SOC, clay, and pH under different scenarios using data from RMQS and LUCAS Soil located in mainland France. The numbers in brackets indicates the width of 90% confidence intervals of 50 repeats. The data included in different scenarios are listed below: S0 (all observations), S1M (2/3 observations and 1/3 spectroscopically inferred data), S1C (2/3 observations), S2M (1/2 observations and 1/2 spectroscopically inferred data), S2C (1/2 observations), S3M (1/3 observations and 2/3 spectroscopically inferred data), S3C (1/3 observations).

Dataset	Soil property	S0		S1M		S1C		S2M		S2C		S3M		S3C	
		PICP	PIW	PICP	PIW	PICP	PIW	PICP	PIW	PICP	PIW	PICP	PIW	PICP	PIW
RMQS Vis-NIR	SOC	91.8 (4.3)	42.8 (3.7)	91.3 (4.6)	42.5 (5.0)	91.7 (3.9)	43.2 (4.6)	90.8 (4.4)	41.7 (6.6)	91.5 (4.4)	43.4 (6.4)	90.1 (5.7)	41.1 (7.4)	91.5 (4.6)	43.9 (7.1)
	Clay	92.3 (4.0)	36.1 (1.5)	90.8 (4.8)	34.7 (2.1)	92.4 (4.2)	36.7 (2.0)	89.7 (5.3)	33.5 (2.1)	92.2 (4.3)	37.1 (2.4)	88.3 (6.0)	32.4 (3.3)	91.8 (4.6)	37.6 (3.6)
	pH	94.3 (3.7)	3.1 (0.1)	93.5 (3.9)	3.1 (0.1)	94.4 (3.9)	3.2 (0.1)	92.8 (4.7)	3.0 (0.1)	94.3 (4.3)	3.3 (0.2)	92.0 (6.1)	3.0 (0.1)	94.3 (4.1)	3.3 (0.2)
RMQS MIR	SOC	91.8 (4.3)	42.8 (3.7)	91.5 (4.0)	43.6 (3.8)	91.7 (3.9)	43.2 (4.6)	91.1 (4.4)	43.5 (4.5)	91.5 (4.4)	43.4 (6.4)	90.8 (5.0)	43.5 (4.7)	91.5 (4.6)	43.9 (7.1)
	Clay	92.3 (4.0)	36.1 (1.5)	91.6 (4.1)	35.7 (1.6)	92.4 (4.2)	36.7 (2.0)	91.2 (3.9)	35.2 (1.8)	92.2 (4.3)	37.1 (2.4)	90.8 (4.5)	34.9 (2.5)	91.8 (4.6)	37.6 (3.6)
	pH	94.3 (3.7)	3.1 (0.1)	93.7 (3.8)	3.1 (0.1)	94.4 (3.9)	3.2 (0.1)	93.2 (4.3)	3.1 (0.1)	94.3 (4.3)	3.3 (0.2)	92.6 (4.5)	3.1 (0.2)	94.3 (4.1)	3.3 (0.2)
LUCAS Soil	SOC	91.2 (3.8)	38.9 (2.1)	90.5 (4.8)	37.9 (2.7)	91.1 (4.3)	39.4 (2.7)	89.7 (4.9)	36.1 (3.1)	91.2 (4.0)	39.7 (2.6)	89.1 (4.0)	35.6 (3.6)	90.8 (4.3)	39.9 (3.8)
	Clay	92.7 (3.4)	31.3 (1.0)	90.8 (4.1)	29.6 (1.4)	92.6 (3.2)	31.6 (1.7)	89.2 (4.3)	27.7 (1.9)	92.4 (3.7)	31.8 (1.9)	86.8 (4.7)	25.9 (2.1)	92.1 (3.7)	32.1 (2.3)
	pH	92.2 (3.1)	2.6 (0.1)	91.2 (3.4)	2.6 (0.1)	91.9 (3.7)	2.7 (0.1)	90.3 (4.6)	2.5 (0.1)	91.8 (3.2)	2.7 (0.1)	89.6 (4.3)	2.4 (0.1)	91.6 (4.4)	2.7 (0.2)

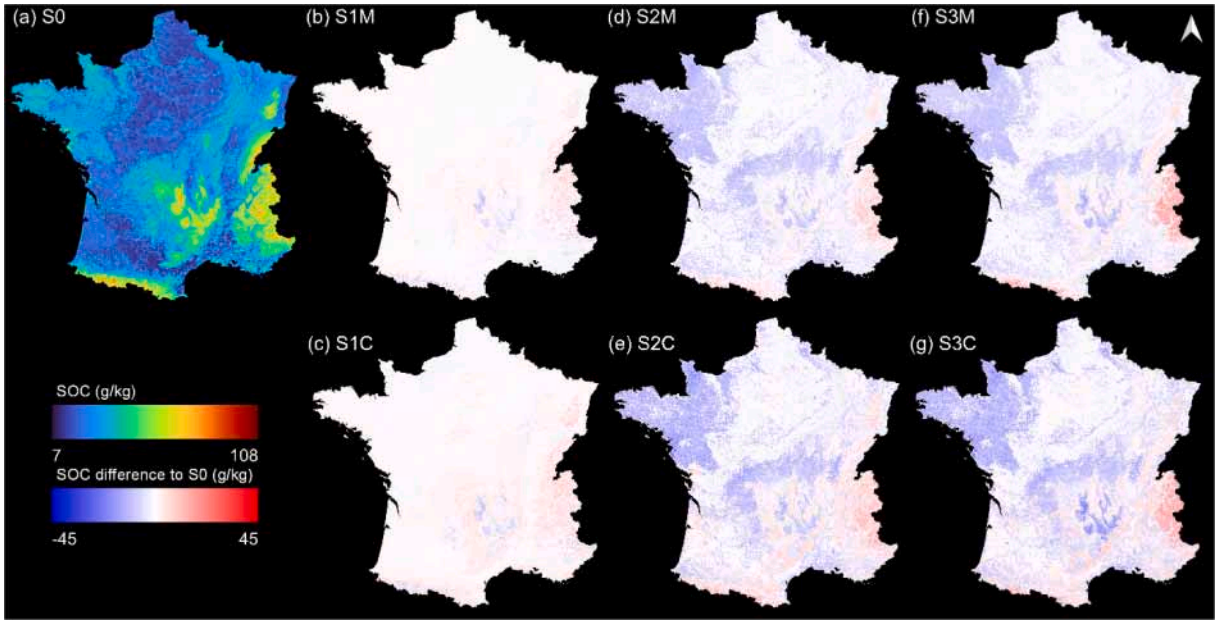


Fig. 4. Spatial distribution of SOC in mainland France for scenario 0 (a) and its difference to other scenarios (b to g) using data from RMQS and MIR inferred data.

As shown in Fig. 4a, high SOC was found in mountainous regions (eastern, southwestern and central France) while low SOC was mainly located in agricultural region (northern and southwestern France). The southwestern and central France had low clay content and highly clayey soil was sparsely distributed in western France and southern France near Mediterranean (Fig. 5a). Acid soil was mainly located in mountainous regions (eastern, southwestern and central France) and alkaline soil could be found in the region near Mediterranean and northern France (Fig. 6a).

The result showed that the difference to S0 gradually increased from S1 to S3 for three soil properties using data from RMQS and MIR inferred data (Fig. 4, Fig. 5 and Fig. 6). The difference was much more evident for pH, SOC than for clay. It was clear that including additional SI data into DSM model (S1M, S2M and S3M) led to closer spatial pattern to S0 than that of using different proportion of observations only (S1C, S2C and S3C).

4. Discussion

4.1. The impact of spectroscopic techniques in spectral prediction

We observed a similar spectral predictive performance between LUCAS Soil and RMQS when using Vis-NIR spectra for modelling. The spectral predictive performance for RMQS using MIR spectra performed substantially better than LUCAS Soil and RMQS using Vis-NIR spectra, especially for SOC and clay (Table 3). Our result is in line with the

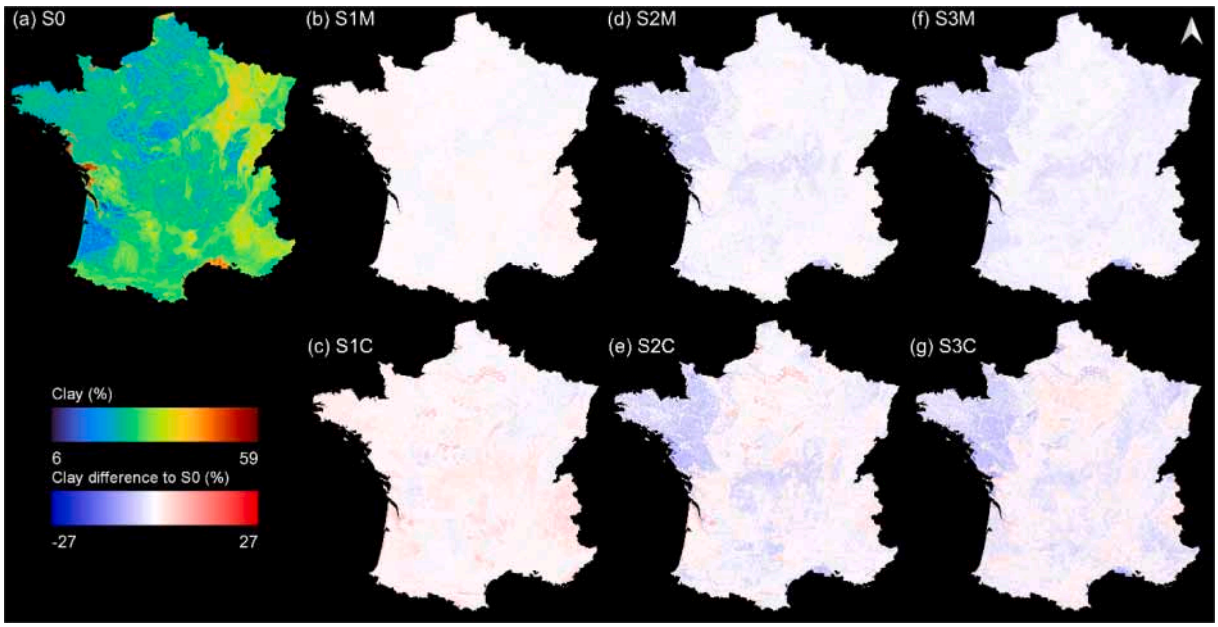


Fig. 5. Spatial distribution of clay in mainland France for scenario 0 (a) and its difference to other scenarios (b to g) using data from RMQS and MIR inferred data.

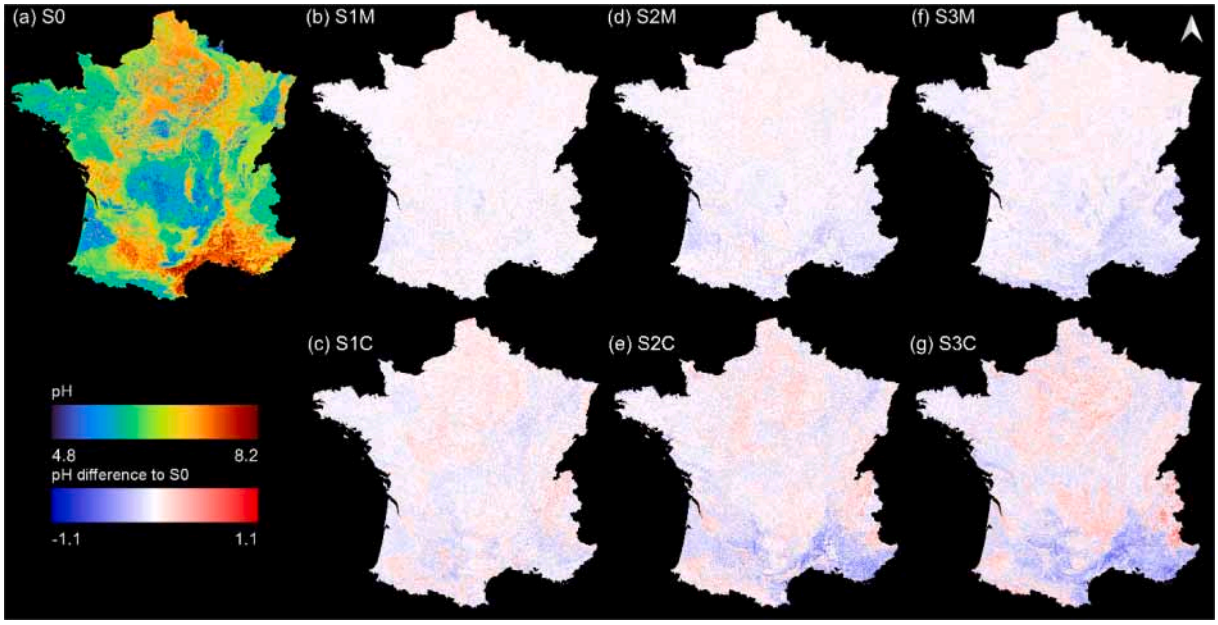


Fig. 6. Spatial distribution of pH in mainland France for scenario 0 (a) and its difference to other scenarios (b to g) using data from RMQS and MIR inferred data.

previous findings (Viscarra Rossel et al., 2006; Vohland et al., 2014; Barthès et al., 2016; Barthès et al., 2020; Clairotte et al., 2016; Hutengs et al., 2019; Ng et al., 2019; Gomez et al., 2020; Gomez et al., 2022) that MIR had higher predictive ability than NIR and Vis-NIR using RMQS data. From this point of view, the MIR technique is preferred for the spectral measurement in the laboratory condition when the budget is enough, as the soil sample preparation (e.g. fine grinding to <0.2 mm or smaller) is more important for MIR than Vis-NIR (Viscarra Rossel et al., 2006; Soriano-Disla et al., 2014).

Compared to SOC and clay, the large difference of spectral prediction on pH using Vis-NIR spectroscopy for RMQS and LUCAS Soil may result from the much greater variation of pH observed in RMQS than LUCAS Soil (Fig. 2c). It seems that grid sampling of RMQS and stratified sampling of LUCAS Soil had a much stronger effect the distribution of pH than SOC and clay. It was also affected by the fact that LUCAS Soil primarily targeted agricultural lands so that it did not cover the regions with low pH (southwestern France, mountainous regions) during the first sampling campaign in 2009 (Fig. 1, Fig. 6a, Tóth et al., 2013). When making the best use of soil samples from different sources, combining data simply is not a good idea since the sampling or analytical protocols are different. In this case, the linear model of coregionalization can be a good tool for spatial modelling by elucidating the effects of different sample support (Lark et al., 2019).

4.2. *The value of spectroscopically inferred soil data in DSM*

Our results confirmed the added value of additional SI soil data in improving model performance using DSM at a national scale. This added value was observed for SOC, clay, and pH; therefore, it was not specific to a particular soil property, at least for these soil properties that can be

well predicted by spectroscopic techniques. We summarize several factors that might impact the magnitude of performance improvement of the DSM model when incorporating SI soil data.

(1) the proportion of SI data. The improvement was negligible (gain of R^2 at 0–4%) when their proportion was low (S1M vs. S1C). It implies that the added value of SI data is rather limited when having a large set of observation data. However, noticeable improvement (gain of R^2 at 3–19%) could be achieved for clay and pH when the proportion of observation data was getting smaller (S2M vs. S2C, and S3M vs. S3C), while the improvement was somewhat marginal for SOC.
(2) the accuracy of the spectral predictive model. When the spectral predictive model had a low accuracy (e.g. R^2 <0.72 for SOC spectral modelling using LUCAS Soil data, Fig. 4), the improvement of the DSM model was somewhat limited (gain of R^2 <3%). In contrast, a more substantial improvement (gain of R^2 at 5–15%) of the DSM model was observed for a spectral predictive model with higher accuracy (e.g. R^2 >0.9 for SOC spectral modelling using RMQS data, and R^2 of 0.87 for pH using LUCAS Soil data, Table 4). This confirmed the study of Paul et al. (2019) that coupling SI soil data from a good spectral model (R^2 of 0.88 for soil organic matter) can improve DSM model even under the condition of high proportion of SI data (Table 1).
(3) the accuracy of the DSM model. The added value of SI data with similar quality (R^2 of 0.92–0.93 in SOC and pH prediction for RMQS using MIR spectra, Table 4) would be higher for a DSM model with better performance (R^2 of 0.41 and 0.53, gain of R^2 at 1–5% and 2–11% for SOC and pH, respectively).

When comparing the DSM models integrating SI data (S1M, S2M, S3M, Table 4) with those using same number of soil observations (S0), we conclude that SI data can provide similar function as soil observations in the DSM model when the spectral predictive model has a good accuracy (R^2 >0.85 for RMQS MIR, Table 4). This means that a large part of soil laboratory observations could be replaced with cheaper SI data without losing too much prediction precision. It may result from the relatively low accuracy of current DSM models (R^2 <0.55 in this study) that the small prediction error of SI data will not impact that much on the DSM model. Therefore, it can be expected that a higher quality of SI data would be needed for DSM models with higher accuracy. We suggest that the opposite result (adding SI data did not improve DSM model) from Somarathna et al. (2018) might result from low spectral predictive model, and another reason may be linked to the high proportion (60–95%) of SI data (Table 1) in DSM models, especially for the subsoil.

We found that integrating more SI data into DSM modelling lowers the PICP and PIW, which may result in underestimation of prediction uncertainty as quantified with a given PI. In this study this underestimation becomes apparent when the SI data make up a larger proportion of the model calibration data, with PICP values dropping below the 90% for some of the RMQS Vis-NIR and LUCAS Soil S3M models. The PICP values of the RMQS MIR S3M models remain around 90%, indicating that the magnitude of the effect of including SI data on the PI is related the performance of the spectral predictive model. Therefore, caution should be taken when using the uncertainty maps created from a DSM model based on a large proportion of SI data because the pre-defined PI of these maps can be probably underestimated. Since statistical models (including machine learning) always overestimate low values and underestimate high values, it will result in an SI data with smaller variability. When merging such an SI data with observations, the overall variability of merged data will decrease, leading to narrower PIW subsequently.

4.3. Limitations and way forward

Our results clearly showed that the improvement of model performance was quite limited when the SI data had low accuracy. It is possible that integrating SI data with low accuracy would even decrease the performance of the DSM model. Therefore, further research is needed to investigate the effect of low-quality SI data in DSM modelling. It could be solved by determining the most likely accuracy threshold or criteria (e.g. R^2) for excluding the data from a low-quality spectral predictive model in DSM practices. These unknown samples beyond the validity domain (e.g. 95% confidence intervals in first two principal components of spectra) of a good spectral predictive model should also be avoided in DSM modelling (Chen et al., 2018b). The trade-off will be there to find the right way to incorporate more or less large uncertain datasets without reducing the prediction performance. These large datasets may provide the advantage of better covering the geographical space, because they can be acquired at a low cost for a large range of locations, but they should not reduce the performance of the DSM model. It should be also noted that most of current spectroscopic techniques are still conducted in the laboratory condition that means we still need to collect soil data from field work (i.e. getting sampling staff into the field, collecting soil, bringing it back to the lab, air-drying, grinding and sieving). As a result, the reduction of cost from laboratory observations to SI data can be quite small when compared to the total cost of the whole procedure. The in-situ and on-the-go spectroscopic techniques can greatly reduce the cost as it avoids sample transportation and pretreatment (Li et al., 2015; Viscarra Rossel et al., 2017; Nawar and Mouazen, 2019; Guerrero et al., 2022). Being greatly impacted by external factors (e.g., soil moisture, surface roughness), the accuracy of in-situ and on-the-go spectroscopic techniques are relatively lower than laboratory spectroscopic techniques. Therefore, how to improve the accuracy of in-situ and on-the-go spectroscopic techniques needs to be urgently explored in future study.

Apart from proximal sensing platforms, more and more multispectral or hyperspectral data are available in large quantities from remote sensing platforms, such as Landsat 8, Sentinel 2, GF-5. Previous studies have demonstrated the potential and efficiency of these data in predicting soil properties for large-scale studies while their accuracy was still lower than proximal sensing based predictions (Lagacherie et al., 2019; Dkhala et al., 2020). Therefore, it will become more and more urgent to find the suitable way to incorporate this massive but still rather inaccurate data into DSM modelling.

Finally, the results presented in this study can be a reference for broad-scale DSM practices while it remains to be validated in more local DSM studies where the pedo-climate conditions are much more homogeneous.

5. Conclusions

We have evaluated the value of additional SI data for mapping soil properties using DSM technique at a national scale. Based on two soil databases, namely, RMQS and LUCAS Soil, we summarized the main messages as below.

(1) Complementing laboratory observations with SI data could improve the performance of DSM model when the ratio of SI data was greater than 50%, otherwise the improvement was negligible;
(2) The magnitude of model improvement was greatly influenced by the proportion of SI data used in DSM model and by the accuracy of spectral predictive models;
(3) A large part of laboratory observations could be replaced with cheaper SI data for DSM modelling when the spectral predictive model does not lose too much prediction precision;
(4) Increasing the proportion of SI data in DSM model lowered the PICP and PIW, which may result in underestimation of prediction uncertainty for a given prediction interval, and thus caution was needed when using the uncertainty maps derived from such a DSM model for decision making;

(5) Further research is needed to better assess how much uncertainty of SI data is acceptable for DSM modelling.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study is funded by National Natural Science Foundation of China (No. 42201054). RMQS soil sampling and physico-chemical analyses were supported by the GIS Sol, which is a scientific group of interest on soils involving the French Ministry for ecology and sustainable development, the French Ministry of agriculture, the French National institute for geographical and forest information (IGN), the French government agency for environmental protection and energy management (ADEME), the Institut de recherche pour le développement (IRD, which is a French public research organization dedicated to southern countries) and the Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE, which is a French public research institute) dedicated to agriculture, food and environment). We thank all the people involved in sampling RMQS sites and in sample preparation. D.A. is coordinator and M.P.M is collaborator of the research consortium GLADSOILMAP supported by the Loire Valley Institute for Advanced Studies through its LE STUDIUM Research Consortium Programme. We are grateful to Gerard Heuvelink and Stephan van der Westhuizen for sharing the R scripts on running the measurement error-filtered quantile regression forest. The authors would like to thank the Joint Research Centre of the European Commission for sharing the LUCAS 2009 TOPSOIL data, and three anonymous reviewers for providing suggestive comments to greatly improve the quality of this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116467.

References

Afnor, 2003. Qualité du sol – Détermination de la distribution granulométrique des particules du sol – Méthode à la pipette, NF X31–107. AFNOR.
AFNOR, 1994. Qualité du sol – Détermination du pH, ISO 10390:1994. AFNOR.
Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Zhang, G.L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Advances in Agronomy 125, 93–134.
Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Regional 9, 1–4.
Barthès, B.G., Kouakoua, E., Moulin, P., Hmaidi, K., Gallali, T., Clairotte, M., Chevallier, T., 2016. Studying the physical protection of soil carbon with quantitative infrared spectroscopy. Journal of Near Infrared Spectroscopy 24 (3), 199–214.
Barthès, B.G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Saby, N.P., Chevallier, T., 2020. Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration–The case of soil inorganic carbon prediction by mid-infrared spectroscopy. Geoderma 369, 114272.
Bouma, J., Montanarella, L., Evanylo, G., 2019. The challenge for the soil science community to contribute to the implementation of the UN Sustainable Development Goals. Soil Use and Management 35, 538–546.
Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353.
Cerdan, O., Govers, G., Le Bissonnais, Y., Van Oost, K., Poesen, J., Saby, N., Dostal, T., 2010. Rates and spatial variations of soil erosion in Europe: A study based on erosion plot data. Geomorphology 122, 167–177.
Chatterjee, S., Hartemink, A.E., Triantafilis, J., Desai, A.R., Soldat, D., Zhu, J., Huang, J., 2021. Characterization of field-scale soil variation using a stepwise multi-sensor fusion approach and a cost-benefit analysis. Catena 201, 105190.
Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D., 2018a. Fine resolution map of top-and subsoil carbon sequestration potential in France. Science of the Total Environment 630, 389–400.
Chen, S., Richer-de-Forges, A.C., Saby, N.P., Martin, M.P., Walter, C., Arrouays, D., 2018b. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma 312, 52–63.
Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Walter, C., 2019. National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. Science of the Total Environment 666, 355–367.
Chen, S., Mulder, V.L., Heuvelink, G.B.M., Poggio, L., Caubet, M., Román Dobarco, M., Arrouays, D., 2020a. Model averaging for mapping topsoil organic carbon in France. Geoderma 366, 114237.
Chen, S., Xu, D., Li, S., Ji, W., Yang, M., Zhou, Y., Shi, Z., 2020b. Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron. Land Degradation & Development 31 (8), 1026–1038.
Chen, S., Richer-de-Forges, A.C., Mulder, V.L., Martelet, G., Loiseau, T., Lehmann, S., Arrouays, D., 2021a. Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France. Catena 198, 105062.
Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Shi, Z., 2021b. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. Geoderma 400, 115159.
Chen, S., Arrouays, D., Mulder, V.L., Poggio, L., Minasny, B., Roudier, P., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567.
Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P., Bernoux, M., Barthès, B.G., 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma 276, 41–52.
Demattê, J. A., Dotto, A. C., Paiva, A. F., Sato, M. V., Dalmolin, R. S., Maria do Socorro, B., ..., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. Geoderma, 354, 113793.
Dkhala, B., Mezned, N., Gomez, C., Abdeljaouad, S., 2020. Hyperspectral field spectroscopy and SENTINEL-2 Multispectral data for minerals with high pollution potential content estimation and mapping. Science of The Total Environment 740, 140160.
Feranec, J., Jaffrain, G., Soukup, T., Hazeu, G., 2010. Determining changes and flows in European landscapes 1990–2000 using Corine land cover data. Applied Geography 30, 19–35.
Filippi, P., Cattle, S.R., Pringle, M.J., Bishop, T.F., 2021. Space-time monitoring of soil organic carbon content across a semi-arid region of Australia. Geoderma Regional 24, e00367.
Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems 110 (1), 168–176.
Gogé, F., Gomez, C., Jolivet, C., Joffre, R., 2014. Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma 213, 1–9.
Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Barthès, B. G., 2020. Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. Geoderma 375, 114469.
Gomez, C., Chevallier, T., Moulin, P., Arrouays, D., Barthès, B.G., 2022. Using carbonate absorbance peak to select the most suitable regression model before predicting soil inorganic carbon concentration by mid-infrared reflectance spectroscopy. Geoderma 405, 115403.
Gray, J., Karunaratne, S., Bishop, T., Wilson, B., Veeragathipillai, M., 2019. Driving factors of soil organic carbon fractions over New South Wales, Australia. Geoderma 353, 213–226.
Grinand, C., Barthès, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). European Journal of Soil Science 63 (2), 141–151.
Guerrero, A., Javadi, S.H., Mouazen, A.M., 2022. Automatic detection of quality soil spectra in an online vis-NIR soil sensor. Computers and Electronics in Agriculture 196, 106857.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978.
Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., Vohland, M., 2019. In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. Geoderma 355, 113900.
IUSS Working Group WRB, 2006. World reference base for soil resources. World Soil Resources Report 103.
Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled srtm for the globe version 4. available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).
Jenny, H., 1941. Factors of Soil Formation. McGraw Hill, New YorkLondon.
Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2016. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. Soil and Tillage Research 155, 492–500.
Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N., 2006. Le Réseau de Mesures de la Qualité des Sols de France (RMQS). État d'avancement et premiers résultats. Étude et Gestion Sols 13, 149–164.

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environmental Modelling & Software 144, 105139.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Bardgett, R.D., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. SOIL 2, 111–128.
King, D., Jones, R., Thomasson, A., 1995. European land information systems for agro-environmental monitoring. European Commission 284.
Knadel, M., Thomsen, A., Schelde, K., Greve, M.H., 2015. Soil organic carbon and particle sizes mapping using vis–NIR, EC and temperature mobile sensor platform. Computers and Electronics in Agriculture 114, 134–144.
Kuhn, M., Quinlan, R., 2020. Cubist: Rule- And Instance-Based Regression Modeling. R package version (2), 3. https://CRAN.R-project.org/package=Cubist.
Kuhn, M., 2020. caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret.
Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N.P., 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., Ander, E.L., Broadley, M.R., 2019. Combining two national-scale datasets to map soil properties, the case of available magnesium in England and Wales. European journal of Soil Science 70 (2), 361–377.
NASA LD, 2001. NASA land processes distributed active archive center (lp daac) usgs/earth resources observation and science (eros) center.
Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W., Webster, R., 2015. In situ measurements of organic carbon in soil profiles using vis-NIR spectroscopy on the Qinghai-Tibet plateau. Environmental Science & Technology 49 (8), 4980–4987.
Liu, S., Shen, H., Chen, S., Zhao, X., Biswas, A., Jia, X., Fang, J., 2019. Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment. Geoderma 348, 37–44.
Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., Zhang, G.-L., 2022. Mapping high resolution National Soil Information Grids of China. Science Bulletin 67 (3), 328–340.
Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Arrouays, D., 2019. Satellite data integration for soil clay content modelling at a national scale. International Journal of Applied Earth Observation and Geoinformation 82, 101905.
Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794.
McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.
McBratney, A.B., Santos, M.d.L., Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1-2), 3-52.
Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7 (6), 983–999.
Minasny, B., McBratney, A.B., 2016. 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. Science of the Total Environment 573, 1352–1369.
Nauman, T.W., Duniway, M.C., 2019. Relative prediction intervals reveal larger uncertainty in 3D approaches to predictive digital soil mapping of soil properties with legacy data. Geoderma 347, 170–184.
Nawar, S., Mouazen, A.M., 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. Soil and Tillage Research 190, 120–127.
Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. Geoderma 352, 251–267.
Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology and Biochemistry 68, 337–347.
Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Wetterlind, J., 2015. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. Advances in Agronomy 132, 139–159.
Paul, S.S., Coops, N.C., Johnson, M.S., Krzic, M., Smukler, S.M., 2019. Evaluating sampling efforts of standard laboratory analysis and mid-infrared spectroscopy for cost effective digital soil mapping at field scale. Geoderma 356, 113925.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.
Priori, S., Fantappiè, M., Bianconi, N., Ferrigno, G., Pellegrini, S., Costantini, E.A., 2016. Field-Scale Mapping of Soil Carbon Stock with Limited Sampling by Coupling Gamma-Ray and Vis-NIR Spectroscopy. Soil Science Society of America Journal 80 (4), 954–964.
Quinlan, R., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference On Artificial Intelligence, 343–348.
R Core Team, 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.
Ramifehiarivo, N., Brossard, M., Grinand, C., Andriamananjara, A., Razafimbelo, T., Rasolohery, A., Razakamanarivo, H., 2017. Mapping soil organic carbon on a national scale: Towards an improved and updated map of Madagascar. Geoderma Regional 9, 29–38.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Zhang, G. L., 2009. Digital soil map of the world. Science 325 (5941), 680–681.
Sanderman, J., Baldock, J.A., Dangal, S.R., Ludwig, S., Potter, S., Rivard, C., Savage, K., 2021. Soil organic carbon fractions in the Great Plains of the United States: an application of mid-infrared spectroscopy. Biogeochemistry 156, 97–114.
Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., Viscarra Rossel, R.A., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Science China Earth Sciences 57 (7), 1671–1680.
Somarathna, P.D.S.N., Minasny, B., Malone, B.P., Stockmann, U., McBratney, A.B., 2018. Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions. Science of the Total Environment 631, 377–389.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Applied Spectroscopy Reviews 49 (2), 139–186.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Advances in Agronomy 107, 163–215.
Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PloS One 8 (6), e66409.
Takoutsing, B., Heuvelink, G.B., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma 428, 116192.
Takoutsing, B., Heuvelink, G.B., Stoorvogel, J.J., Shepherd, K.D., Aynekulu, E., 2022. Accounting for analytical and proximal soil sensing errors in digital soil mapping. European Journal of Soil Science 73 (2), e13226.
Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental Monitoring and Assessment 185 (9), 7409–7425.
van der Westhuizen, S., Heuvelink, G.B., Hofmeyr, D.P., Poggio, L., 2022. Measurement error-filtered machine learning in digital soil mapping. Spatial Statistics 47, 100572.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 59–75.
Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology 20 (9), 2953–2970.
Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Research 53 (8), 845–864.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Reviews 155, 198–230.
Viscarra Rossel, R.A., Lobsey, C.R., Sharman, C., Flick, P., McLachlan, G., 2017. Novel proximal sensing for monitoring soil organic C stocks and condition. Environmental Science & Technology 51 (10), 5630–5641.
Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B., 2014. Determination of soil properties with visible to near-and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma 223, 88–96.
Wadoux, A.M.C., Padarian, J., Minasny, B., 2019. Multi-source data integration for soil mapping using deep learning. SOIL 5 (1), 107–119.
Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77 (1), 1–17.
Xiao, Y., Xue, J., Zhang, X., Wang, N., Hong, Y., Chen, S., 2022. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. Geoderma 428, 116208.
Yang, H., Kuang, B., Mouazen, A.M., 2012. Quantitative analysis of soil nitrogen and carbon at a farm scale using visible and near infrared spectroscopy coupled with wavelength reduction. European Journal of Soil Science 63 (3), 410–420.
Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Shi, Z., 2023. Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. Geoderma 432, 116383.
Zhang, Y., Ji, W., Saurette, D.D., Easher, T.H., Li, H., Shi, Z., Biswas, A., 2020. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma 366, 114253.