HAL open science

Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France

Songchao Chen, Anne C Richer-De-Forges, Vera Leatitia Mulder, Guillaume Martelet, Thomas Loiseau, Sébastien Lehmann, Dominique Arrouays

▶ To cite this version:

Songchao Chen, Anne C Richer-De-Forges, Vera Leatitia Mulder, Guillaume Martelet, Thomas Loiseau, et al.. Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France. CATENA, 2021, 198, pp.105062. 10.1016/j.catena.2020.105062 . hal-03467117

HAL Id: hal-03467117

https://hal.inrae.fr/hal-03467117

Submitted on 3 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0341816220306123
Manuscript_6ff9f7e7bd117bd3190654fd4b36bd31

Title: Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France

Authors:

Songchao Chen[a]. songchao.chen@inrae.fr

Anne C. Richer-de-Forges[a]. anne.richer-de-forges@inrae.fr

Vera Leatitia Mulder[b]. titia.mulder@wur.nl

Guillaume Martelet[c]. g.martelet@brgm.fr

Thomas Loiseau[a]. thomas.loiseau@inrae.fr

Sébastien Lehmann[a]. sebastien.lehmann@inrae.fr

Dominique Arrouays[a]. dominique.arrouays@inrae.fr

Affiliations:

[a] INRAE, Unité InfoSol, 45075 Orléans, France

[b] Soil Geography and Landscape Group, Wageningen University, PO Box 47 6700 AA Wageningen, The Netherlands

[c] BRGM, UMR 7327, 45060 Orléans, France

Corresponding author:

Songchao Chen: songchao.chen@inrae.fr

Postal address: INRAE, Unité InfoSol, 2163 Avenue de la Pomme de Pin, CS 40001 Ardon, 45075 Orléans, France

© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

Abstract:

Soil thickness (ST) plays an important role in regulating soil processes, vegetation growth and land suitability. Therefore, it has been listed as one of twelve basic soil properties to be delivered in *GlobalSoilMap* project. However, ST prediction has been reported with poor performance in previous studies. Our case study is located in the intensive agriculture Beauce area, central France. In this region, the ST mainly depends on the thickness of loess (TOL) deposits over a calcareous bedrock. We attempted to test the TOL prediction by coupling a large soil dataset (10978 sampling sites) and 117 environmental covariates. After variable selection by recursive feature elimination, quantile regression forests (QRF) was employed for spatial modelling, as it was able to directly provide the 90% prediction intervals (PIs). Averaging a total of 50 models, generated by repeated stratified random sampling, showed a substantial model performance with mean R^2 of 0.33, RMSE of 30.48 cm and bias of -1.20 cm. The prediction interval coverage percentage showed that 86.70% of the validation samples fall within the predefined 90% PIs, which also indicated the prediction uncertainty produced by QRF was reasonable. The relative variable importance indicated the importance of airborne gamma-ray radiometric data and Sentinel 2 products in TOL prediction. The produced TOL map with 90% PIs makes sense from a soil science and physiographic point of view. The final product can guide evidence-based decision making for agricultural land management, especially for irrigation in our case study.

1. Introduction

Soil thickness (ST) is an important soil property due to its influence as a controlling factor of numerous surface and subsurface soil processes. Through its influence on the plant rootable depth (Leenaars *et al.*, 2018), ST is a major controlling factor of vegetation growth and land suitability, and is a key variable to estimate available water capacity (AWC). As a consequence, ST has been retained as a mandatory soil attribute to be delivered in *GlobalSoilMap* products (Arrouays *et al.,* 2014). Previously, some attempts have been made to map ST at national, continental or global levels (e.g., Grundy *et al.*, 2015; Lacoste *et al.*, 2016; Mulder *et al.*, 2016; Hengl *et al.*, 2017; Chen *et al.*, 2019). Grundy *et al.* (2015) mapped ST in Australia using about 300,000 points of observations and environmental covariates as inputs for a Cubist model. Lacoste *et al.* (2016) tested three digital soil mapping (DSM) approaches, based on regression tree modelling, gradient boosting modelling, and multi-resolution kriging for a dataset of ca 14,000 observations in France. Hengl *et al.* (2017) mapped ST at a global scale. Predictions were based on ca. 150,000 soil profiles used for training and a stack of 158 covariates which were used to fit an ensemble of machine learning methods—random forest and gradient boosting and/or multinomial logistic regression. Mulder *et al.* (2016) used Cubist predictions for mainland France using ca 16,000 observation points and a set of 20 spatially exhaustive covariates. Chen *et al.* (2019) further demonstrated how right-censored data can be accounted for in the ST modelling of mainland France. Using random survival forest, the probability of exceeding a given depth was modelled using freely available spatial data representing the main soil-forming factors. However, most of these results gave rather poor prediction performances compared to other soil properties such as soil organic carbon or clay content and pH (e.g., Mulder *et al.*, 2016; Hengl *et al.*, 2017). In many cases, ST prediction proved to be hampered either 1) by the lack of data measurements (Leenaars *et al.*, 2018) or 2) by the fact that the collected ST data is often right-censored data (i.e. the observed ST is less than true ST, Chen *et al.*, 2019), or 3) because ST has a high short-range spatial variability in specific pedological contexts (e.g., Bourennane *et al.*, 1996; Lacoste et *al.*, 2016). Moreover, most of the examples taken from the literature were produced using digital soil mapping (DSM, McBratney *et al.*, 2003) and in some cases, one may expect that no relevant covariate was available to improve the performance of the predictions.

For example, while topography-related covariates such as elevation or slope often explain large part of the variability of soils in a given areas, these covariates might not improve the performance of predictions in flat areas.

Mapping ST enables several applications among which are agronomy and agricultural practices (plant rootable depth, drainage, irrigation), crop growth modelling, geotechnical engineering, water balance modelling at catchment to global scale, and Quaternary science studies.

The Beauce area, located in central France is a limestone plateau irregularly covered with Quaternary loessic silt (Macaire, 1971; Lorain, 1973; Ménillet, 1974). Soil classification varies from Luvisols to Calcic and Calcaric Cambisols. The region is rather flat, and has a noticeable proportion of rather thin soils. Intensive agriculture in this region often utilizes irrigation and most of the surface area is occupied by cereal crops (mainly maize and wheat) and sugar beet. Also, the Beauce area is home of the largest aquifer of France in the underlying calcareous rock. Upper horizons were affected by peri-glacial winds that redistributed loess deposits (Macaire, 1971; Bourennane *et al*., 1996), resulting in a rather homogeneous particle-size distribution of the fine earth (i.e., silt, clay, and sand). Therefore, most of the soils consist now of silt, silt loam or silty clay loam layers derived from this aeolian deposit developed on a lacustrine limestone substrate. In general, the illuviation process occurred when the thickness of loess was the largest, resulting in less clayey topsoil textures.

The available water capacity (AWC) is the maximum amount of available water that can be stored for crop growth, therefore it is an important soil information for agricultural management. Therefore, the thickness of loess (TOL) deposit is also one of the primary factors influencing the calculation of the soil AWC. Although Tetegan *et al*. (2015) demonstrated that the percentage of rock fragments was also one of the controlling factors of AWC in this region. Overall, in this region, irrigation management is of upmost importance in order to maintain crop yields, while preserving the underlying water table and water quality. Knowing the TOL is essential for determining a water balance and for piloting irrigation management. In terms of agronomy and environment, the TOL is a determining factor (Nicoullaud *et al.,* 1995; Ould-Mohamed *et al.,* 1997). Therefore, the TOL should be known accurately and cheaply over the study area. Several traditional soil maps have been produced in this region, with scales ranging from 1:50,000 to 1:250,000, resulting in a various density of point scale soil information.

The objective of this study is to assess to which extent using this legacy data and environmental covariates (from existing geological maps, digital elevation model derivatives, airborne gamma-ray radiometry, and remote sensing data) in a DSM model allows to reach acceptable performances for TOL prediction. In this study, we decided to model the TOL up to a depth of 120 cm using Quantile Regression Forests (QRF) because the TOL was deemed useful for agricultural practices. Maize cropping is especially of interest because it is known for its high water requirement (Doorenbos *et al.*, 1978) and thus typically requires the largest amounts of irrigation. The average rooting depth of maize is equal to 120 cm (British Standards Institution, 1988; Tetegan *et al.*, 2015). Therefore, we only mapped the TOL up to a depth of 120 cm, as there is no difference of soil water management between soils with a TOL deeper than 120 cm and soils with a TOL of 120 cm.

2. Material and methods

2.1. Study area

This study was conducted in the Beauce area located at the middle Loire catchment, central France (Figure 1). It covers a total area of 4835 km^2, of which agriculture is the dominant land use (88.5%, Inglada *et al.*, 2017). It has a continental-oceanic climate with a mean annual temperature of 11.5°C and a mean annual rainfall of 700 mm (Paroissien *et al.*, 2014). Most of the soils in this study area are developed from periglacial loess deposits which covered a limestone bedrock. Cambisol (48.3%) and Luvisol (25.6%) are the major soil groups observed in this region (IUSS Working Group WRB, 2006). At the southern border of the Beauce region, some other soil groups (not developed from loess) are observed.

2.2. Soil data

We used available soil data from the French Soil Inventory Program (IGCS). The thickness of loess derived horizons (TOL) was determined by several criteria: 1) digging soil pits down to the calcareous material and 2) by auger borings. The presence of small rock fragments could in some case lead to an underestimation of TOL done by augering. Therefore, if a TOL of 120 cm was not reached, two other augerings were made randomly 0.5 m apart from the first one and the maximum TOL reached was recorded. TOL should have a texture of silt, silt loam or silty clay loam (Bertran *et al.*, 2016; Borderie *et al.*, 2017). The deeper the TOL the more the illuviation processes are pronounced and the lighter the topsoil texture.

In total, 10978 sites were used in this study to map the TOL up to a depth of 120 cm. The TOL for sites with a TOL deeper than 120 cm (n=14) was set to 120 cm before modelling to eliminate the effect of extreme values in modelling.

2.3. Environmental covariates

The environmental covariates used in this study and their data sources are listed in Table 1. These covariates provide information on the environmental factors assumingly controlling TOL, based on the Scorpan conceptual model (McBratney *et al.*, 2003). For illustrative purposes, several covariates are shown in Figure 2.

2.3.1. Relief

The Digital Elevation Model of mainland France was derived from BD TOPO 3 of the French National Geographical Institute (IGN, 2011), at 25 m resolution. SAGA GIS (Conrad *et al.*, 2015) was used to calculate its derivatives (relief factors), including channel network base level (CNBL), multiresolution index of valley bottom flatness (MrVBF), plan curvature (PlC), profile curvature (PrC), slope (Sl), slope position (SlP), slope length (SlL), terrain wetness index (TWI), valley depth (VD), and vertical distance to channel network (VDCN). As the relief factor at neighbouring locations is able to provide additional useful information in modelling soil patterns (McBratney *et al.*, 2003), some previous studies investigated the potential of incorporating local neighbourhood information into the training pixels, using convolution filtering operations (e.g., Grinand *et al.*, 2008; Loiseau *et al.*, 2019). Filtering can be achieved by passing a moving window over the variable to calculate a value of the processing cell (central pixel) using the values of its neighbouring cells. In this study, we used mean convolution circular windows to calculate the focal means for these relief factors with radius at 200, 500 and 1000 m (Grinand *et al.*, 2008), which resulted in three raster layers derived from each original relief factor (25 m).

2.3.2. Soil

The soil type information were extracted from the French national soil type map at 1:1 M scale (King *et al.*, 1994). The soil types in this study area were mainly Cambisols and Luvisols. However, some Podzols, Gleysols, Fluvisols, Arenosols and Vertisols were rarely present, mainly at the southern border of the region.

2.3.3. Parent material

The map of parent material was extracted from the French national parent material map (King *et al.*, 1994). Undifferentiated alluvial deposits, calcareous rocks, clayey materials, sandy materials and loamy materials are the main parent materials in the study area. Note that the loamy materials are nearly always located over underlying calcareous rocks.

The gamma radiometric data, including Potassium (K), Thorium (Th) and Uranium (U), and total count (TC), was derived from an airborne high-resolution magnetic and radiometric survey over the Région Centre, flown by Terraquest Ltd, Canada, under the supervision of BRGM between 2008 and 2009 (Martelet *et al.*, 2014). The line-spacing of the survey was 1 km and, along the flight lines the footprint of each gamma radiometric measurement was an ellipse of 150 × 250 m^2; accordingly the data were interpolated on 250 m grids using a standard minimum curvature interpolation.

2.3.4. Organisms

A land use map was extracted from the French land use map, which was produced from Sentinel 2 data at 10 m resolution, for year 2016 (Inglada *et al.*, 2017). This land use map was aggregated to 25 m resolution by majority sampling and the proportions (0~100%) of the nine main land-use classes within each 25×25 m pixel (which contained 6 10×10 pixels) were also included as covariates.

The monthly normalized difference vegetation index (NDVI) from the MODIS (MCD43A4 16-day Version 6) in 500 m resolution and the PROBA-V 10-day product level 2B TOC (Copernicus, 2016) in 300 m resolution were used in this study. These 24 monthly NDVI data in 2003 (extreme warm and dry year) and 2016 (normal year) were collected and reduced into the first three principal components by principal component analysis to eliminate their multicollinearity. For more details, we refer to Loiseau *et al.* (2019).

We also included 42 covariates related to Sentinel 2 bands (year of 2016 to 2017) and indices, which were produced in an earlier study from Loiseau *et al*. (2019) for mainland France at 90 m resolution. The Sentinel 2 data were processed to Level-2A (atmospheric and topographic corrections) by the French National Centre for Space Studies (Hagolle *et al*., 2015). These covariates included 10 Sentinel 2 bands (2, 3, 4, 5, 6, 7, 8, 8A, 11 and 12), 11 spectral indices (brightness index, saturation index, hue index, coloration index, redness Index, carbonate index, ferrous iron, clay index, normalized difference 1, normalized difference 2 and grain size index) and their focal means determined by a low-pass filter with an average within a 2×2 km window. For more details, we refer to Loiseau *et al*. (2019).

2.3.5. Position

The coordinates, i.e., latitude and longitude (extracted for each recorded sampling site), were used in modelling. In addition, 10 oblique geographic coordinates were calculated at angles of 15°, 30°, 45°, 60°, 75°, 105°, 120°, 135°, 150° and 165°. The oblique coordinate (OC) at an angle of θ can be calculated as below (Møller *et al*., 2019):

$$\mathrm{OC} = \sqrt{X^2 + Y^2} \times \cos(\theta - \tan^{-1}(\frac{Y}{X})) \quad (1)$$

where X and Y are the latitude and longitude.

Note that when θ is 0° or 90°, the oblique coordinate equals to latitude or longitude.

2.3.6. Harmonization of environmental covariates

The environmental covariates had different resolutions and scales, we therefore harmonized them at 25 m resolution using nearest neighbour interpolation for spatial predictive modelling and mapping at non-visited locations.

2.4. Variable selection using recursive feature elimination

Considering the large set of environmental covariates (*n*=117), variable selection was applied by recursive feature elimination (Kunn, 2020) prior to fitting the spatial predictive model. The recursive feature elimination (incorporating resampling) adopts a backwards selection, which includes several steps: (1) split data into training and test set by resampling (i.e., *k*-fold cross-validation); (2) train the model on the training set using all predictors, calculate the model performance on the test set, and rank predictors using their model importance; (3) for each predictor subset size S_i (i=1, 2, …, s), train the model on the training set using the S_i most important predictors, and calculate the model performance on the test set; (4) compare the model performance profile over the S_i on the test set, and determine the optimal number of predictors.

To select the important covariates and improve the mapping efficiency, the recursive feature elimination was performed on the whole data using *rfe* function in *caret* package (Kunn, 2020) in R (R Core Team, 2019). The model was set to Random Forest (default values with tree number of 500 and mtry of p/3 where p is the size of predictors) using 5-fold cross-validation. Seven predictor subset sizes (5, 10, 15, 20, 40, 60, 80 and 100) were tested and the model performance indicated that 80 variables (Table 2) were optimal and then used for later modelling.

2.5. Spatial predictive modelling and model performance evaluation

Quantile Regression Forest (QRF, Meinshausen, 2006) has been growingly used in DSM for delivering soil information as it is able to provide uncertainty estimates straightforwardly with a fair model performance (e.g., Vaysse and Lagacherie, 2017; Lombardo *et al*., 2018; Loiseau *et al*., 2019). Therefore, QRF was used for modelling TOL in this study.

Since QRF is an extension of Random Forest (RF, Breiman, 2001), we start with RF. Assume X and Y are the predictor variables and responses, for regression, RF generates a large number (b) of bootstrap trees by using m training samples (X_i, Y_i), i=1,…, m. Here, bootstrap refers to repeated (b times) selection of a random sample with replacement of the training samples. For each node in a bootstrap tree, a random subset of the predictor variables is used for split-point selection. The prediction of a bootstrap tree for a new sample D=X_d is the conditional mean estimate ($\hat{X}$) of Y, which can be represented by:

$$\hat{X} = \sum_{i=1}^{m} w_i Y_i \quad (2)$$

where w_i is the weight of the sample (X_i, Y_i) in the same leaf of the bootstrap tree.

The final prediction of the new sample D is approximated by the mean predictions of b bootstrap trees.

Apart from the conditional mean estimate in RF, QRF also uses the weighted samples to derive a conditional distribution. This distribution function is able to provide the probability of *Y* being lower than a given percentile and thus to calculate the prediction intervals. For more details about the constructions of the conditional distribution, we refer to Meinshausen (2006).

We used the *quantregForest* package (Meinshausen, 2017) in R (R Core Team, 2019) for implementing QRF to derive the median prediction and 90% prediction intervals (90% PIs, 5^{th} and 95^{th} quantiles). The default number of tree (*ntree*=500) and minimum size of terminal nodes (*nodesize*=5) were used for QRF, and the number of variables randomly sampled as candidates at each split (*mtry*) was optimized in the *caret* package (Kunn, 2020) by 5-fold cross-validation in R (R Core Team, 2019). The variable importance was determined by the increased mean square error (IncMSE, in %) between the model excluding and including a given variable, and this information was integrated in QRF model. In our case, the variable importance was calculated by the average of 50 repeated models.

Considering the highly varying soil sampling density (Brus *et al*., 2011), we divided the study area into 20 compact equal area geographical strata (Figure 3) using the *spcosa* package (Walvoort *et al*., 2020) in R (R Core Team, 2019), and performed stratified random sampling (5 sites for each strata) for selecting the validation set. It resulted in a set of 10878 sites for model calibration and 100 sites for model validation. To derive a robust result, we repeated this procedure 50 times and took the average as the final model performance.

Four indicators were used to evaluate the model performance in validation set: (1) modelling efficiency (R^2); (2) root mean square error (RMSE); (3) bias; (4) prediction interval coverage percentage (PICP), which describes the percentage of the observed TOL falls within the estimated upper and lower 90% PIs.

$$\mathrm{R}^2 = 1 - \frac{\sqrt{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}} \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)}{n}^2} \quad (4)$$

$$\mathrm{Bias} = \sum_{i=1}^{n} \frac{(\widehat{y_i} - y_i)}{n} \quad (5)$$

where *n* is the sample size of observations, y_i and $\widehat{y_i}$ are observed value and predicted value for sample *i*, and $\overline{y_i}$ is the average of observed values.

In addition, we also reported the model performance by the internal validation using out-of-bag data (around 34% of data that is not used for growing each tree) in QRF. The final TOL map and its 90% PIs were produced by QRF fitted using all the sampling sites.

3. Results and discussion

3.1. Summary of TOL in the Beauce area

Table 3 displays the statistics of the TOL in the Beauce area. Among 10978 sites, TOL ranged from 0 to 120 cm, with a mean and median TOL at 36.46 cm and 30 cm, respectively. A skewness of 0.66 (larger than 0.5) indicated the data were slightly positively skewed while a kurtosis less than 3 (2.85) showed that the data were light-tailed. Though the log transformation, i.e. log(TOL+a), is able to convert soil data to normal distribution (a skewness less than 0.5 and a kurtosis close to 3), it did not improve the model performance. Therefore, instead of data transformation, we used the original TOL data for spatial modelling in this study.

Figure 4 presents the TOL located in 20 compact equal area geographical strata. It showed a large difference of number of sampling sites among these 20 geographical strata, ranging from 22 to 1351. These geographical strata with high median TOL (>60 cm) had much less sampling sites (51 to 201), and this is the main reason for evaluating the model performance by random stratified sampling.

Figure 5 shows the Pearson correlations coefficients between TOL and top 30 environmental covariates. Elevation and its focal mean derivatives had the highest positive correlations (r > 0.3). Other positively correlated covariates were related to oblique coordinates, channel network base level (CNBL), graphic coordinates, gamma-ray radiometry and two Sentinel 2 indexes (grain size index and clay index). Negative correlations with TOL were found with valley depth, Sentinel 2 bands and its indexes. Overall, the correlations between TOL and covariates were found to be rather low ($|r| < 0.35$).

3.2. Model performance of Quantile Random Forest

The mean R^2, RMSE and bias from the internal validation using the out-of-bag data in QRF were 0.31, 26.88 cm and 0.23 cm (data not shown). Figure 6 indicates the model performance after we repeated 50 times the validation procedure using QRF. The mean R^2 and RMSE were 0.33 and 30.48 cm respectively. The mean bias of -1.20 cm indicated that the prediction was almost un-biased for 50 repeats. The mean PICP indicated that on average 86.7% of the validation samples fall within the defined 90% PIs, therefore the uncertainty estimates from the QRF model was valid for non-visited locations.

As shown in Table 4, the global soil thickness (ST) products (Hengl *et al*., 2017; Shangguan *et al*., 2017) had better model performance than those at national or regional scale (Guerrero *et al*., 2014; Kidd *et al*., 2015; Vaysse and Lagacherie, 2015; Lacoste *et al*., 2016; Mulder *et al*., 2016; Zhang *et al*., 2018). This may be attributed to the fact that global ST products include a substantial proportion of very thin soils (i.e., soils prone to severe erosion) and of very thick ones (i.e., Arenosols in desert dunes, Shangguan *et al*., 2017). There was no large difference of model performance between national and regional products, even if we used nearly 11000 sampling sites in this study. This is because the TOL is highly variable at short distances. By incorporating a large exhaustive set of environmental covariates, however, the map produced in this study performed slightly better than almost all the previous studies at regional and national scales.

The large range between upper and lower limits of 90% PIs for R^2 and RMSE indicated the randomness involved in data split brought a large amount of uncertainty in model evaluation. Therefore, instead of a single time data split, repeated random (stratified) sampling adopted in this study would provide more robust estimates for the model performance so as to avoid under- or over- optimistic decision making in management of soil resources.

3.3. Variable importance of environmental covariates

Figure 7 displays the top 30 environmental covariates in QRF model calculated as the average of 50 repeats. It indicated that the gamma radiometric data (U, Th and TC) and hue index (focal mean) calculated from Sentinel 2 images were the most important environmental covariates in modelling TOL in the study area. They were followed by longitude, NDVI PC1, slope position (with a radius of 1000 m), TWI (with a radius of 1000 m), normalized difference (focal mean) and grain size index (focal mean), representing position, organisms and relief factors in Scorpan conceptual model. For many relief (e.g., slope position, TWI, VDCN, curvature, valley depth, CNBL, elevation, slope) and organisms (e.g., hue index, normalized difference, grain size index, ferrous iron) factors, their derivatives calculated from neighbouring information performed better than original covariates. Spatial position, i.e., latitude and longitude, were identified important in Figure 7 while oblique coordinates were not listed among the top 30 covariates.

Interestingly, the variable importance in the QRF model was not in line with the correlations between TOL and covariates (see Figure 5). This may be due to the fact that the relations between TOL and covariates are not linear. If the relationships were linear then the most important covariates should have been those with highest $|r|$ which is not the case in this study. Another reason may be that the importance of covariates results also from interactions between them, that are not visible using Pearson correlations but that are taken into account in QRF model.

Our results indicate a high importance of airborne gamma radiometric data in TOL modelling as they can capture soil information relevant to soil texture and to the presence of the calcareous rock at low depth. Indeed, the substrate of part of the study area (composed carbonates) is completely different from the TOL and it has been shown that calcium mitigates surface gamma-spectrometric signatures because it has a poor gamma-spectrometric response (Martelet *et al.*, 2013). Therefore, it is not surprising that gamma radiometric data plays an important role, especially for predicting thin TOL. Also, the large plateaus with deep TOL in the northern part are depleted in K (see Figure 2). This is because soils with large TOL were prone to illuviation, resulting in lower clay content in topsoil. So in this case, it is an indirect relationship with TOL. Our results also confirm the contribution of neighbouring information (e.g, focal hue index, slope position 1000m, TWI 1000m, focal normalized difference) of relief and organism factors in spatial modelling of TOL, which implicates the multi-scale influence of covariates on soil properties. Concerning slope and TWI, the importance of this neighbouring information may be due to the gradient of TOL that shows that very large flat plateaus (mainly in the north) are characterized by a deeper TOL. These derivatives likely performed well because it is not the same geomorphological context if you have a flat location inside a very large flat plateau than if you have locally flat "pixels" in a region where the relief is more accentuated, such for instance in the southwest (Behrens et al., 2019). Other studies also have shown the potential of multi-scale covariates derivatives in improving model performance in DSM (Behrens *et al*., 2018b, 2019). Compared to simple convolution approach (focal mean), wavelet transforms, empirical mode decomposition, and the Gaussian scale space may even better represent the multi-scale information of environmental covariates (Behrens *et al*., 2010, 2018a, 2018b; Biswas *et al*., 2013a, 2013b; Zhou *et al*., 2016; Huang *et al*., 2017; Zhao *et al*., 2018; Liang *et al*., 2019) so as to improve model performance in DSM. The Sentinel 2 spectral bands may not always provide direct information related to soil, while a great potential has been shown from its derived indicators (e.g., NDVI, hue index, normalized difference, ferrous iron, and grain size index) in this study. Considering its high spatial and temporal resolution, Sentinel 2 has a great potential in delivering useful information of soil surface for DSM across scales (Gholizadeh *et al*., 2018; Castaldi *et al*., 2019; Loiseau *et al*., 2019; Vaudour *et al*., 2019). Some of the Sentinel 2 data we used come from a mosaic of images of bare soils built by Loiseau *et al.* (2019). Therefore, these Sentinel 2 data provide direct information on soil colour which may reflect thin TOL or absence of TOL through the presence of white calcareous rocks at the surface. They may indirectly reflect also texture through bright colours due to slaking that occurs mainly on very loamy topsoil soils which correspond to the deepest TOL where Luvisols have developed. The land use map produced by Sentinel 2 was not among the top 30 environmental covariates as it may be masked by the NDVI data due to their correlation or NDVI better explains the spatial variability than land use map. Therefore, relative importance of environmental covariates should be taken with caution as high contributing covariates can inadvertently bear part of the contribution of the less contributing covariates (Chen *et al*., 2018).

3.4. Maps of thickness of loess and its 90% prediction intervals

Figure 8 presents the spatial distribution of TOL and its lower and upper limits of 90% PIs. It displays the general increasing thickness of loess soils from south-west to north in the study area. Very shallow loess (<10 cm) was found in south-west of the study area, and very deep loess (>100 cm) was mainly found in the northern part. Highest TOL were mainly located in rather flat areas located on high elevation plateaus, while shallow loess was mainly located at lower elevations and in more dissected relief, especially in the vicinity of small valleys. Note that there is a border effect from the south-west to the west of the region. This border effect corresponds to the outcropping limit of the TOL, where sandy or clayey materials locally overlay the calcareous. The regions with thin soil (<10 cm with a lot of outcrops of the calcareous) correspond to the areas with black gamma-ray radiometry patterns matching on steep slopes around the drainage lines (mostly rivers).

The maps of lower and upper limits of 90% PIs clearly show different spatial structures. On the northern part with the highest elevations and high mean TOL, the 95^{th} percentile is equal or deeper than 1.2 m, which means that high TOL are largely dominant in these plateaus. On the contrary the extreme southern part of the region exhibits TOL that rarely exceed 0.6 m. Moreover, except for some very local areas having a high mean TOL, the 5^{th} percentile map suggests that the upper calcareous surface is undulating at very short distances and that local calcareous outcrops may be found in nearly all the southern part of the study area. The wind direction of loess deposits was from northwest (Bertran *et al.*, 2016; Borderie *et al.*, 2017). The Beauce area corresponds to the southern margin of the Paris basin loess deposits which show a clear gradient from North to South (Bertran *et al.*, 2016). The gradient of loess that we observe in the Beauce region from north to south may be due to this. In addition, the northern part is characterized by large flat plateaus where no erosion occurred, except along the main deep valleys, whereas the southwestern part is characterized by a local relief that may have induced erosion and redistribution processes (Macaire, 1971). All these observations were confirmed by the expert knowledge of the soil surveyors who did some traditional reconnaissance soil mapping in this region. Interestingly, when doing reconnaissance maps at 1:250,000 the soil surveyors delineated small natural regions in order to create the broadest geographical ensembles of the legend (Richer-de-Forges, 2008; Richer-de-Forges *et* *al.*, 2008). Figure 9 shows these small natural regions drawn by the soil surveyors on the study area. The comparison between Figure 8 and Figure 9 clearly shows that the map of the TOL makes sense both from soil and physiography point of views. One should keep in mind that 90% PI is a very large PI. Therefore, it is normal that such wide ranges are found. Another reason for the large PI comes from the fact that our map does not have very high model performance, and there is still a large room to improve it. Useful outputs for irrigation or drainage management, however, are maps of probability of exceeding a given depth for TOL in the study area (see an example in the next section).

3.5. Example of application

One example of application is to map the probability of the TOL to exceed a given depth or, on the contrary, to map the probability of the TOL to be less than a given value. Figure 10 displays an example of these practical applications, which extracts the probability of exceeding of 30 cm from the function between the TOL and prediction quantile (from 0 to 100% with an interval of 2%) within the QRF model. The soils that have a very low probability to exceed a 30 cm TOL are unsuitable for conventional tillage and have a very low AWC. Therefore, optimizing the irrigation on these soils should greatly save water.

4. Conclusion

In this study, we utilized a large soil dataset (10978 sampling sites) and 117 environmental covariates relevant to soil, organisms, relief, parent material and spatial position for mapping thickness of loess at a regional scale. The 50 repeated Quantile Random Forest had an average R^2 of 0.33, which was slightly better than those obtained in most previous studies at regional or national scale (R^2 of 0.11~0.41). A PICP of 86.70% showed that around 86.70% of the validation samples fall within the predefined 90% PIs, which indicated that the prediction uncertainty produced by Quantile Random Forest was reasonable and can be properly used in decision making of land management. The relative importance of environmental covariates indicated the importance of elevation and gamma radiometry in modelling thickness of loess and also proved the necessity of incorporating neighbour information in relief and organisms for spatial modelling. The produced map of thickness of loess and its 90% prediction intervals made sense from a soil science perspective. This map can be further used for efficient irrigation management as well as crop growth and yield modelling.

Acknowledgement

Dominique Arrouays is coordinator, Vera Leatitia Mulder is member, and Anne C. Richer-de-Forges, and Guillaume Martelet are collaborators of the Research Consortium GLADSOILMAP supported by “LE STUDIUM” Loire Valley Institute for Advances Research Studies. This work was funded by the French National Research Agency (Soilserv program, ANR 16 CE32 0005 01). Songchao Chen has received the support of China Scholarship Council for 3 years’ Ph.D. study in INRAE and Agrocampus Ouest (under grant agreement no. 201606320211). We also acknowledge Dr. Hélène Tissoux (BRGM, France) for sharing her knowledge on loess deposits in France and the Beauce region.

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., ..., Mendonca-Santos, M.d.L., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. In Advances in agronomy (Vol. 125, pp. 93-134). Academic Press.

Behrens, T., MacMillan, R. A., Viscarra Rossel, R.A., Schmidt, K., Lee, J., 2019. Teleconnections in spatial modelling. Geoderma, 354, 113854.

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018a. Multi-scale digital soil mapping with deep learning. Scientific reports, 8(1), 1-9.

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018b. Multiscale contextual spatial modelling with the Gaussian scale space. Geoderma, 310, 128-137.

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155(3-4), 175-185.

Bertran, P., Liard, M., Sitzia, L., Tissoux, H., 2016. A map of Pleistocene aeolian deposits in Western Europe, with special emphasis on France. Journal of Quaternary Science, 31(8), 844-856.

Biswas, A., Cresswell, H.P., Chau, H.W., Viscarra Rossel, R.A., Si, B.C., 2013a. Separating scale-specific soil spatial variability: A comparison of multi-resolution analysis and empirical mode decomposition. Geoderma, 209, 57-64.

Biswas, A., Cresswell, H.P., Viscarra Rossel, R.A., Si, B.C., 2013b. Characterizing scale‐and location‐specific variation in non‐linear soil systems using the wavelet transform. European Journal of Soil Science, 64(5), 706-715.

Borderie, Q., Chamaux, G., Roussaffa, H., Douard, M., Fencke, E., Rodot, M.-A., Perrichon, P., Selles, H., 2017. La couverture lœssique d'Eure-et-Loir (France): Potentiel pédo-sedimentaire et organisation spatiale. Quaternaire, 28(3), 389-400.

Bourennane, H., King, D., Chery, P., Bruand, A., 1996. Improving the kriging of a soil variable using slope gradient as external drift. European Journal of Soil Science, 47(4), 473-483.

Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.

British Standards Institution, 1988. BSI Standards Catalogue: 1988. The Institution.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394-407.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., ..., van Wesemael, B., 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. ISPRS Journal of Photogrammetry and Remote Sensing, 147, 267-282.

Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D., 2018. Fine resolution map of top-and subsoil carbon sequestration potential in France. Science of the Total Environment, 630, 389-400.

Chen, S., Mulder, V.L., Martin, M.P., Walter, C., Lacoste, M., Richer-de-Forges, A.C. C., Arrouays, D., 2019. Probability mapping of soil thickness by random survival forest at a national scale. Geoderma, 344, 184-194.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development, 8, 1991-2007.

Copernicus, 2016. Contains Modified Copernicus Service Information. https://land.copernicus.eu/global/products/ndvi.

Doorenbos, J., Kassam, A.H., Bentvelder, C., Uittenboogaard, G., 1978. Yield response to water. FAO Irrigation and Drainage Paper 33. FAO, Rome, Italy, p. 144.

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sensing of Environment, 218, 89-103.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma, 143(1-2), 180-190.

Grundy, M.J., Viscarra Rossel, R.A., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L. J., 2015. Soil and landscape grid of Australia. Soil Research, 53(8), 835-844.

Guerrero, E., Pérez, A., Arroyo, C., Equihua, J., Guevara, M., 2014. Building a national framework for pedometric mapping: soil depth as an example for Mexico. In: GlobalSoilMap: Basis of the global spatial soil information system, Taylor & Francis, CRC Press, London 103-108.

Hagolle, O., Huc, M., Villa Pascual, D., Dedieu, G., 2015. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and Sentinel-2 images. Remote Sensing, 7(3), 2668-2691.

Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Guevara, M.A., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

Huang, J., Wu, C., Minasny, B., Roudier, P., McBratney, A.B., 2017. Unravelling scale-and location-specific variations in soil properties using the 2-dimensional empirical mode decomposition. Geoderma, 307, 139-149.

IGN, 2011. BD ALTI® Version 2.0 - Descriptif de Contenu. Available online at: http://professionnels.ign.fr/sites/default/files/DC_BDALTI_2-0.pdf

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sensing, 9(1), 95.

IUSS Working Group, WRB, 2006. World reference base for soil resources. World Soil Resources Report, 103.

Kidd, D., Webb, M., Malone, B., Minasny, B., McBratney, A., 2015. Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. Soil Research, 53(8), 932-955.

King, D., Daroussin, J., Tavernier, R., 1994. Development of a soil geographic database from the Soil Map of the European Communities. CATENA, 21(1), 37-56.

Kuhn, M., 2020. caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. Evaluating large-extent spatial modeling approaches: A case study for soil depth for France. Geoderma Regional, 7(2), 137-152.

Leenaars, J.G., Claessens, L., Heuvelink, G.B., Hengl, T., González, M.R., van Bussel, L.G., ..., Cassman, K.G., 2018. Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. Geoderma, 324, 18-36.

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., Viscarra Rossel, R.A., 2019. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. Geoderma, 335, 47-56.

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.

Lombardo, L., Saia, S., Schillaci, C., Mai, P.M., Huser, R., 2018. Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. Geoderma, 318, 148-159.

Lorain, J.M., 1973. La géologie du calcaire de Beauce. Bulletin de Liaison des Laboratoires des Ponts et Chaussées. Special Issue, 78.

Macaire, J.M., 1971. Etude sédimentologique des formations superficielles sur le tracé de l'autoroute A10 entre Artenay et Meung sur Loire (PhD thesis). University of Orleans, France.

Martelet, G., Drufin, S., Tourlière, B., Saby, N. P. A., Perrin, J., DeParis, J., Prognon, J. F., Jolivet, C., Ratié, C., Arrouays, D., 2013. Regional regolith parameters prediction using the proxy of airborne gamma ray spectrometry. Vadose Zone Journal, 12(4), 1-14.

Martelet, G., Nehlig, P., Arrouays, D., Messner, F., Tourlière, B., Laroche, B., Deparis, J., Saby, N., Richer-de-Forges, A.C., Jolivet, C., Ratié, C., 2014. Airborne gamma-ray spectrometry: potential for regolith-soil mapping and characterization. In: GlobalSoilMap: Basis of the global spatial soil information system. Taylor & Francis, CRC Press, London 401-408.

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, 983-999.

Meinshausen, N., 2017. quantregForest: Quantile Regression Forests. R package version 1.3-7.

Ménillet, F., 1974. Etude pétrographique et sédimentologique des calcaires d'Etampes et de Beauce. Formations dulcaquicoles du Stampien supérieur à l'Aquitanien dans le Bassin de Paris. Thesis, University of Paris-Sud (Orsay).

Møller, A.B., Beucher, A.M., Pouladi, N., Greve, M.H., 2019. Oblique geographic coordinates as covariates for digital soil mapping. SOIL Discussions, 1-20.

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.

Nicoullaud, B., Darthout, R., Duval, O., 1995. Etude de l'enracinement du blé tendre d'hiver et du maïs dans les sols argilo-limoneux de Petite Beauce. Etude et Gestion des Sols, 2, 183-200.

Ould-Mohamed, S., Bruand, A., Bruckler, L., Bertuzzi, P., Guillet, B., Raison, L., 1997. Estimating long-term drainage at a regional scale using a deterministic model. Soil Science Society of America Journal, 61, 1473-1482.

Paroissien, J. B., Saby, N., de Forges, A. R., Arrouays, D., Louis, B., 2014. Populating soil maps with legacy data from a soil testing databases. In:

GlobalSoilMap: Basis of the global spatial soil information system. Taylor & Francis, CRC Press, London 319-324.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Richer-de-Forges, A.C., 2008. Notice explicative de la carte des pédopaysages du Loiret à 1/250 000 (Référentiel Régional Pédologique de la région Centre). INRA InfoSol. (in French)

Richer-de-Forges, A.C., Desbourdes, S., Lehmann, S., 2008. Référentiel Régional Pédologique de la région Centre : carte des pédopaysages du Loiret à 1/250 000. INRA InfoSol. (in French)

Shangguan, W., Hengl, T., de Jesus, J. M., Yuan, H., Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. Journal of Advances in Modeling Earth Systems, 9(1), 65-88.

Tetegan, M., Richer-de-Forges, A.C., Verbeque, B., Nicoullaud, B., Desbourdes, C., Bouthier, A., Arrouays, D., Cousin, I., 2015. The effect of soil stoniness on the estimation of water retention. Catena, 129, 96-102.

Vaudour, E., Gomez, C., Fouad, Y., Lagacherie, P., 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. Remote Sensing of Environment, 223, 21-33.

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

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

Walvoort, D., Brus, D., de Gruijter, J., 2020. Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata. R package version 0.3-9. https://CRAN.R-project.org/package=spcosa

Zhang, W., Hu, G., Sheng, J., Weindorf, D.C., Wu, H., Xuan, J., ..., Gu, Z., 2018. Estimating effective soil depth at regional scales: Legacy maps versus environmental covariates. Journal of Plant Nutrition and Soil Science, 181(2), 167-176.

Zhao, R., Biswas, A., Zhou, Y., Zhou, Y., Shi, Z., Li, H., 2018. Identifying localized and scale-specific multivariate controls of soil organic matter variations using multiple wavelet coherence. Science of The Total Environment, 643, 548-558.

Zhou, Y., Biswas, A., Ma, Z., Lu, Y., Chen, Q., Shi, Z., 2016. Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain. Geoderma, 271, 71-79.

Figures

Figure 1 Study area and soil sampling sites, the Beauce area that locates at the middle Loire catchment, central France

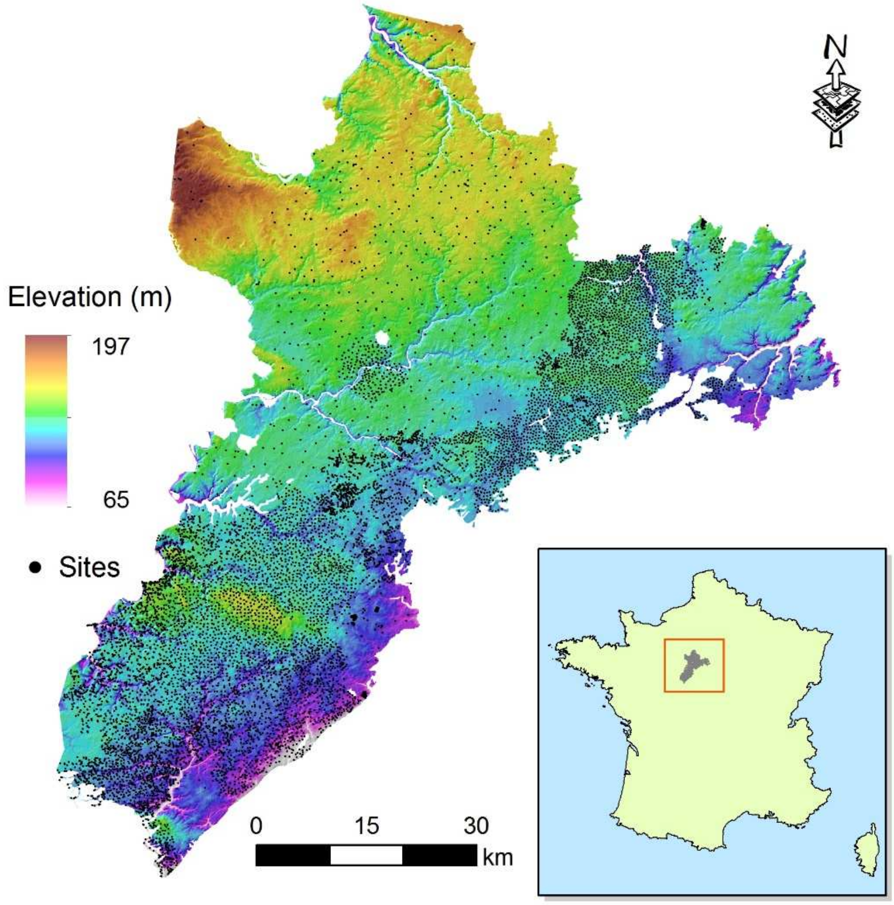


Figure 2 Examples of environmental covariates

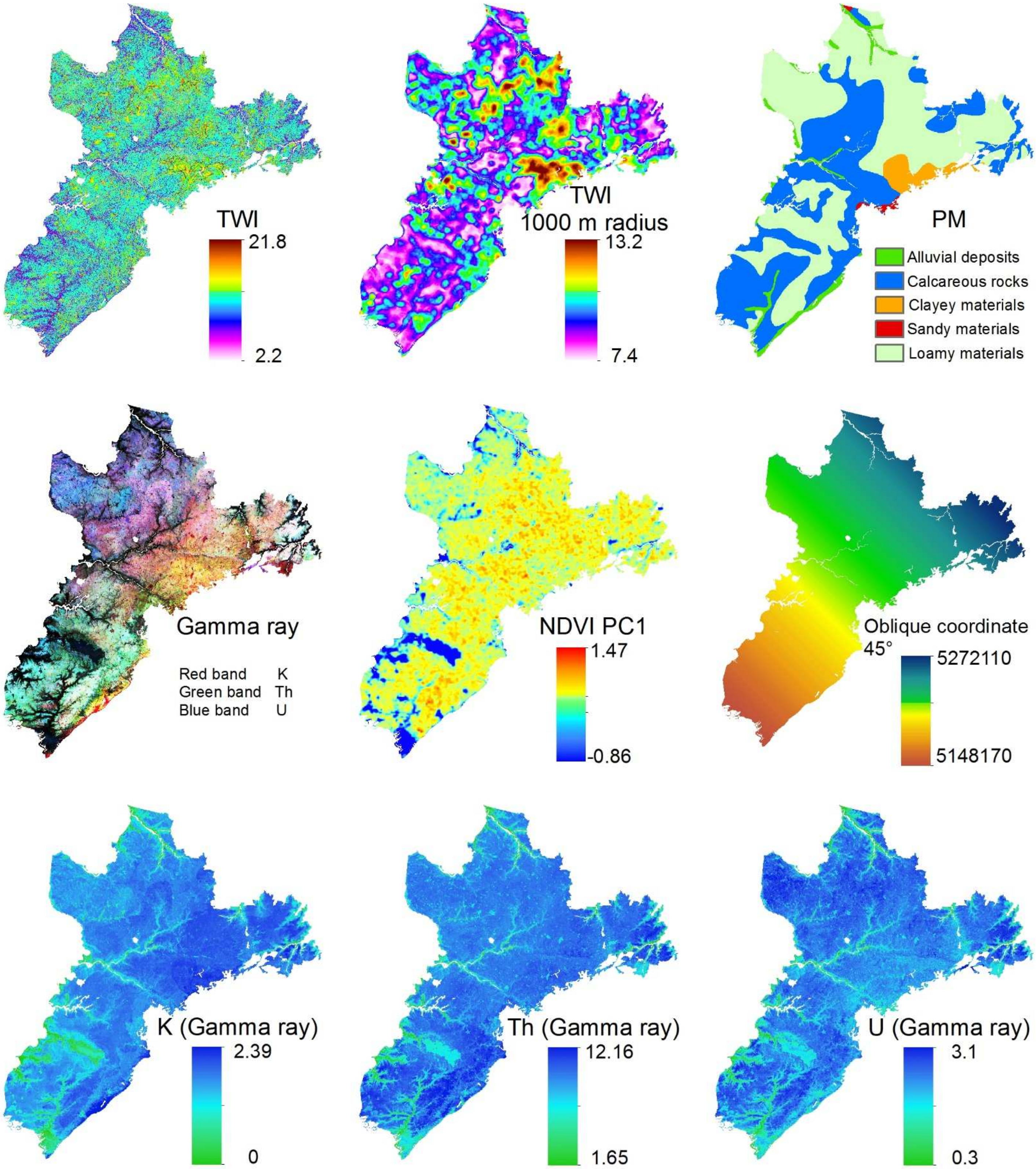


Figure 3 Compact equal area geographical strata

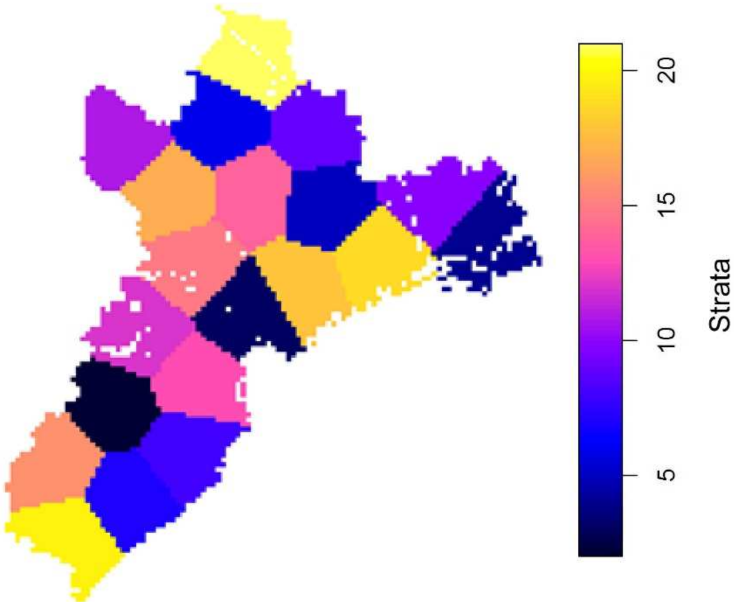


Figure 4 Boxplot of TOL for each compact equal area geographical strata. The number of sampling sites is indicated in blue.

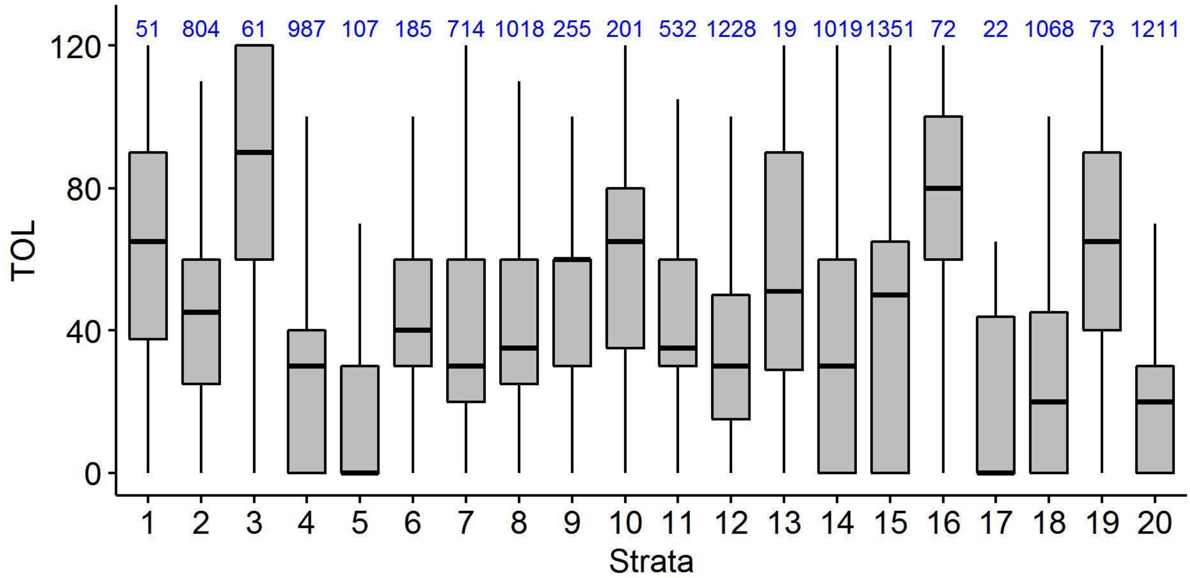


Figure 5 Pearson correlation coefficient (r) between TOL and top correlated environmental covariates (r > 0.15 or r < -0.15).

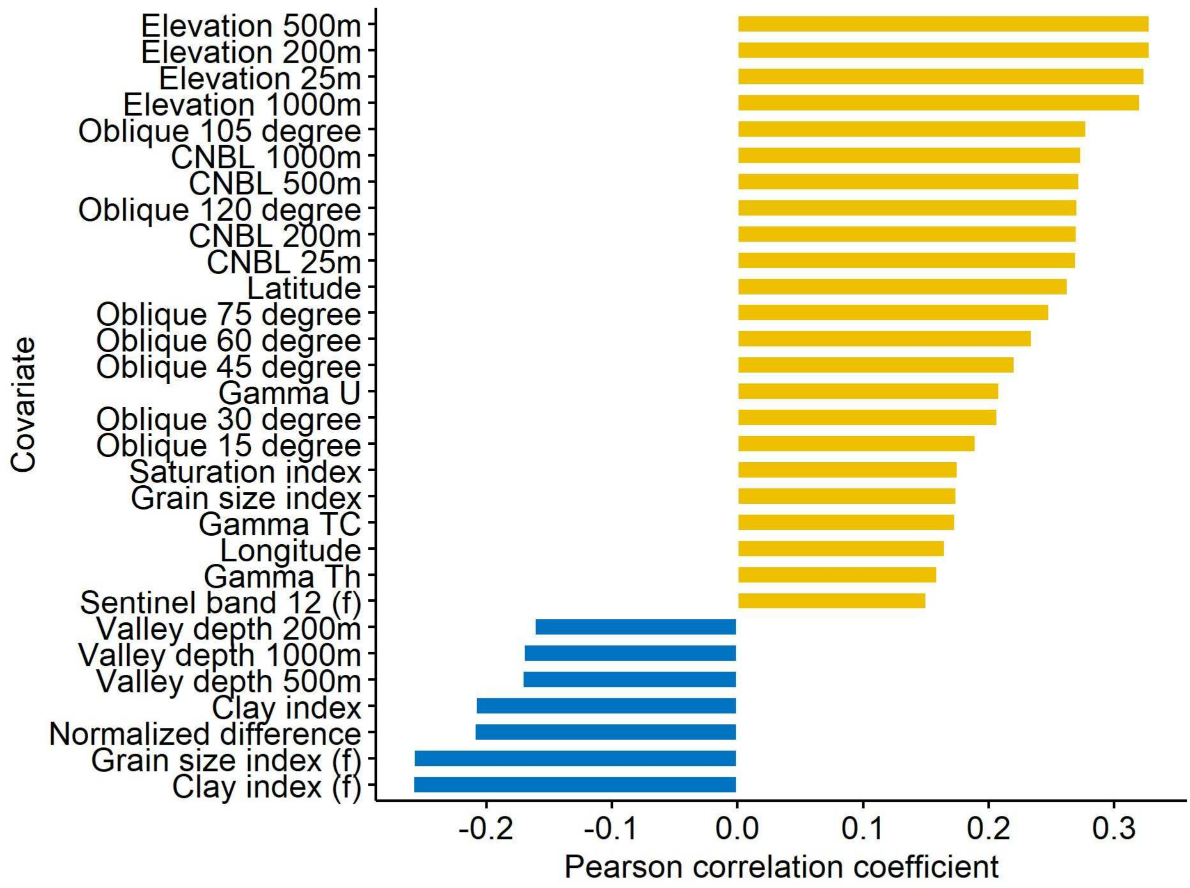


Figure 6 Model performance of 50 repeats evaluated by R^2 (a), RMSE (b), bias (c) and coverage of PICP (d) on validation set

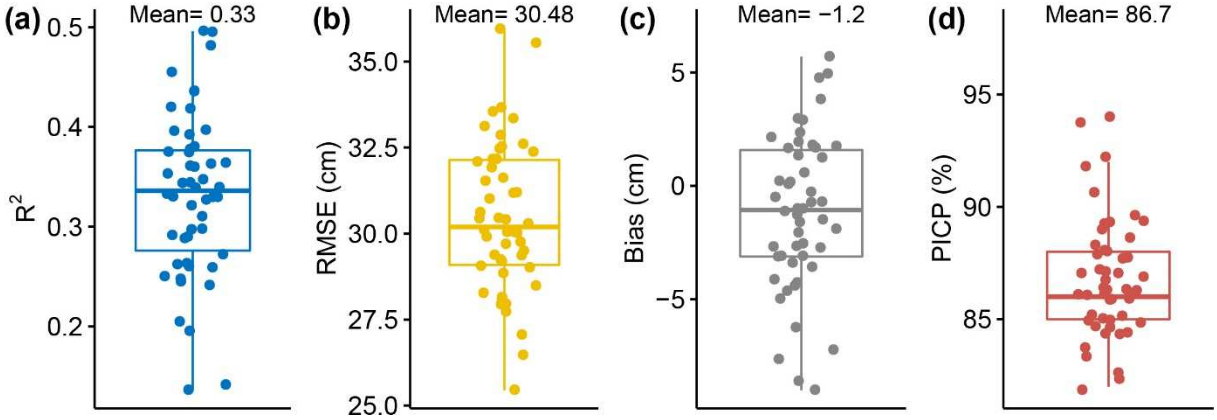


Figure 7 Relative importance of environmental covariates in Quantile Random Forest (average of 50 repeats). Only the top 30 variables are shown here.

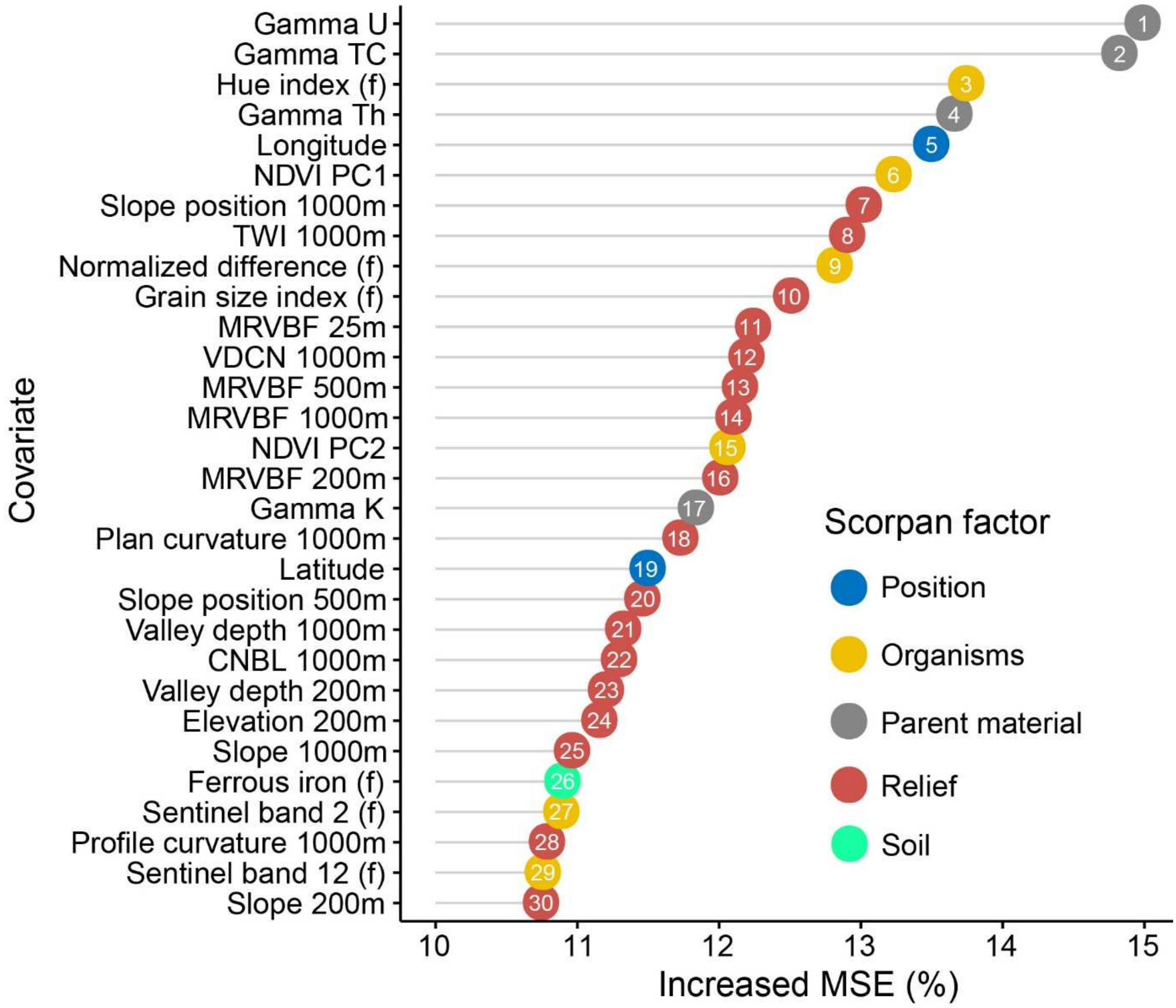


Figure 8 Spatial distribution of the thickness of loess and its 90% prediction intervals

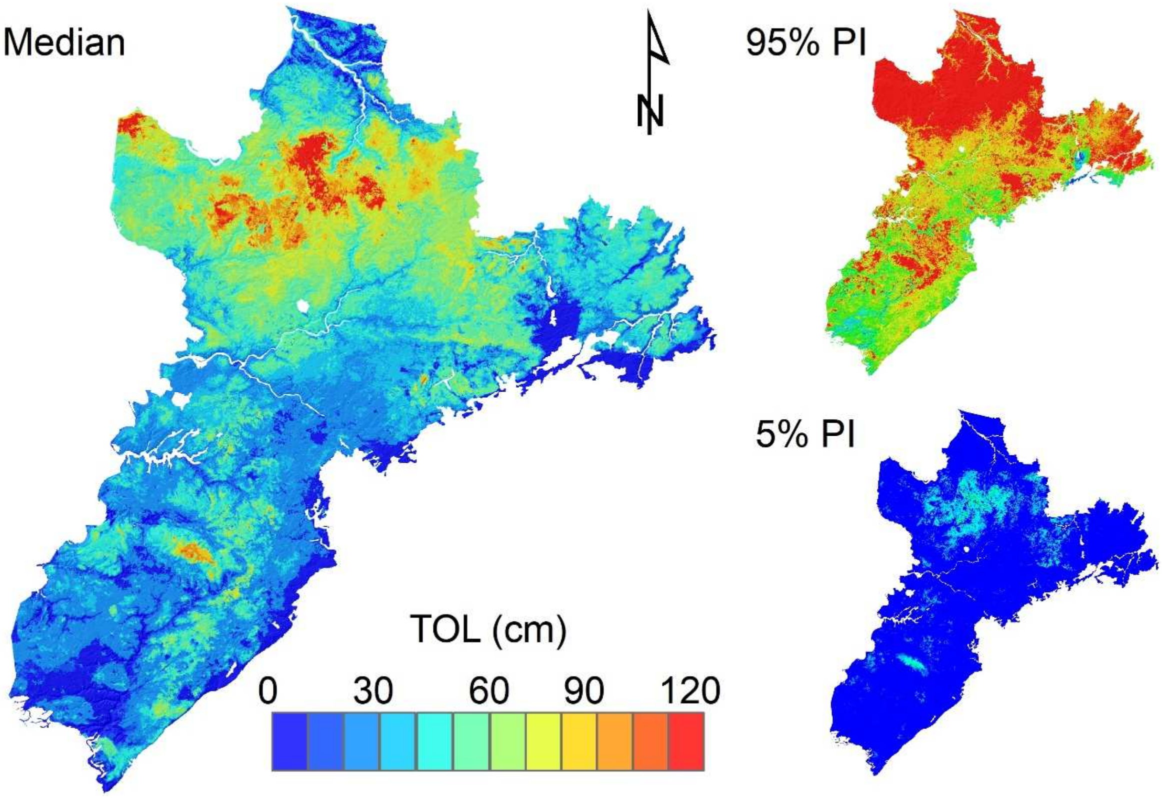


Figure 9 Natural regions delineated by soil surveyors (Richer-de-Forges, 2008; Richer-de-Forges *et al.*, 2008)

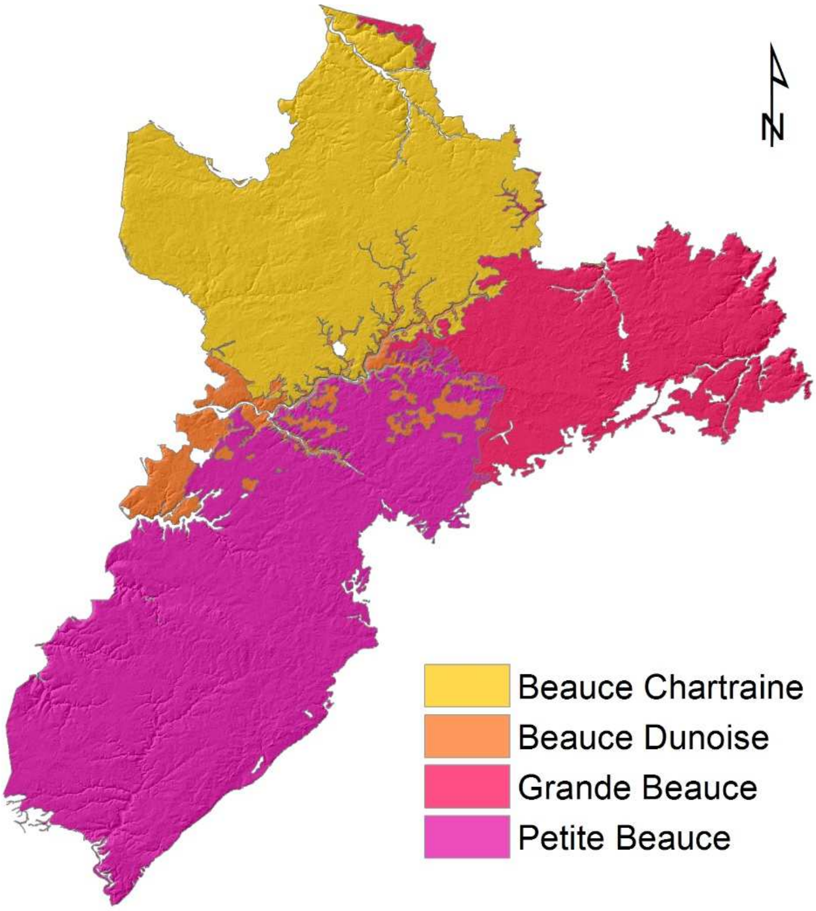
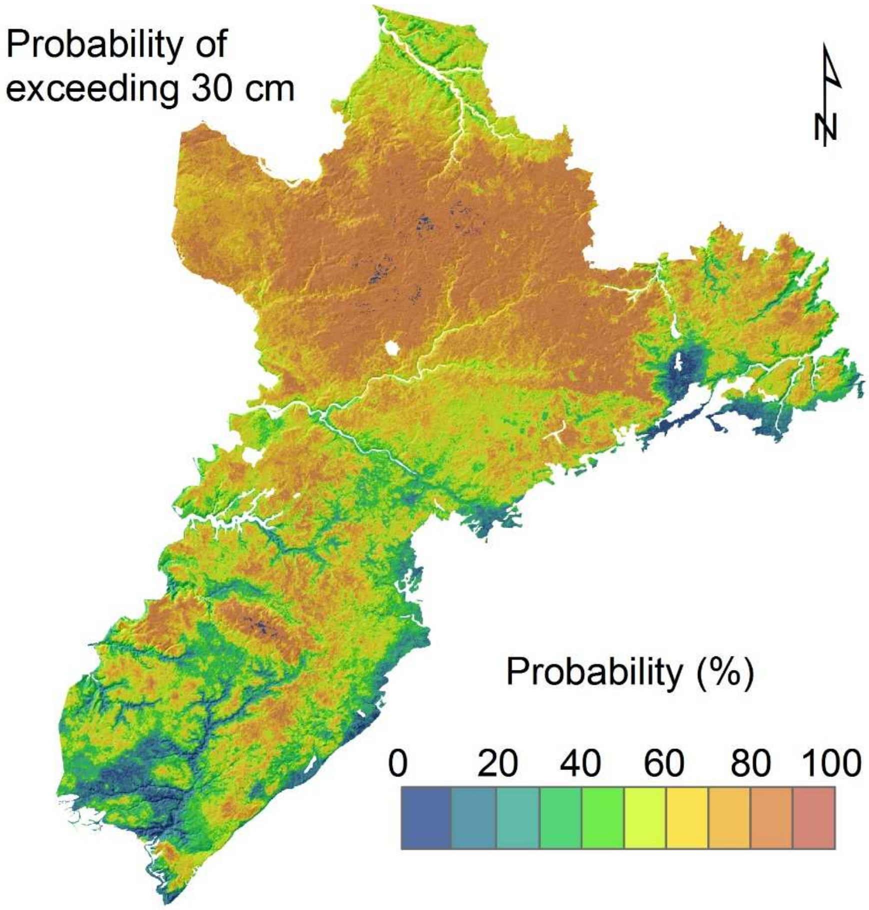


Figure 10 The probability of exceeding 30 cm for TOL in the study area. The probability at 30 cm is extracted from the probability distribution of Quantile Random Forest for each pixel.



Tables

Table 1 Environmental covariates used for digital soil mapping

Variable	Number	Resolution	Scorpan factor	Reference
Channel network base level	4	25 m	Relief[a]	IGN, 2011
Elevation	4	25 m	Relief	IGN, 2011
Multiresolution index of valley bottom flatness	4	25 m	Relief	IGN, 2011
Plan curvature	4	25 m	Relief	IGN, 2011
Profile curvature	4	25 m	Relief	IGN, 2011
Slope	4	25 m	Relief	IGN, 2011
Slope position	4	25 m	Relief	IGN, 2011
Slope length	4	25 m	Relief	IGN, 2011
Terrain wetness index	4	25 m	Relief	IGN, 2011
Valley depth	4	25 m	Relief	IGN, 2011
Vertical distance to channel network	4	25 m	Relief	IGN, 2011
Soil type	1	1:1000000	Soil	King *et al.* (1995)
Parent material	1	1:1000000	Parent material	King *et al*. (1995)
Gamma radiometric (K, U, Th, TC)	4	200 m	Parent material	Martelet *et al*. (2014)
Land cover and probability	10	10, 20, 60 m	Organisms	Inglada *et al.* (2017)
Sentinel 2 spectral bands and indices	42	90 m	Organisms	Loiseau *et al*. (2019)
First three PCs of monthly NDVI[b]	3	300, 500 m	Organisms	Loiseau *et al.* (2019)
Coordinates (Latitude, Longitude)	2	25 m	Position	IGN, 2011
Oblique coordinates[c]	10	25 m	Position	Møller *et al*. (2019)

[a] For all the covariates in relief factor, except for the original products, their local mean values with radius at 200, 500 and 1000 m are also calculated by convolution circular windows. [b] PCs, principal components; NDVI, normalized difference vegetation index. [c] Oblique coordinates at angles of 15°, 30°, 45°, 60°, 75, 105°, 120°, 135°, 150° and 165° are produced

Table 2 Random Forest model performance over the predictor subset size using recursive feature selection

Subset size	RMSE	R^2	Selected
5	28.70	0.1808	No
10	27.07	0.2712	No
15	26.72	0.2904	No
20	26.52	0.3007	No
40	26.50	0.3022	No
60	26.50	0.3019	No
80	**26.44**	**0.3052**	**Yes**
100	26.50	0.3025	No
117	26.48	0.3038	No

Table 3 Statistics of the thickness of loess (in cm)

Variable	Number	Minimum	Q1	Median	Mean	Q3	Maximum	Skewness	Kurtosis
TOL	10978	0	0	30	36.46	60	120	0.66	2.85

Q1, the first quartile; Q3, the third quartile.

Table 4 A summary of the model performance of soil thickness (or soil depth) mapping from regional to global scales

Reference	Location	R^2
Regional scale		
Kidd et al. (2015)	Tasmania, Australia	0.16
Vaysse and Lagacherie (2015)	Languedoc-Roussillon, France	0.23
Zhang et al. (2018)	Xinjiang, China	0.28
This study	Beauce, France	0.34
National scale		
Guerrero et al. (2014)	Mexico	0.41
Lacoste et al. (2016)	France	0.22
Mulder et al. (2016)	France	0.11
Global scale		
Hengl et al. (2017)	Globe	0.57
Shangguan et al. (2017)	Globe	0.59