Harvesting spatially dense legacy soil datasets for digital soil mapping of available water capacity in Southern France

Quentin Styc, François Gontard, Philippe Lagacherie

Harvesting spatially dense legacy soil datasets for digital soil mapping of available water capacity in Southern France

Styc Quentin[1,2]

François Gontard[2]

Lagacherie Philippe[1]

[1] LISAH, Univ Montpellier, INRAE, IRD, Institut Agro, Montpellier, France

[2] BRL Exploitation, Nîmes, France

Corresponding author: Lagacherie Philippe

Abstract

Although considerable work has been conducted in recent decades to build soil databases, the legacy data from a lot of former soil survey campaigns still remain unused. The objective of this study was to determine the interest in harvesting such legacy data for mapping the soil available water capacities (SAWCs) at different rooting depths (30 cm, 60 cm, 100 cm) and to the maximal observation depth, over the commune of Bouillargues (16 km^2, Occitanie region, southern France)

An increasing number of available auger hole observations with SAWC estimations – from 0 to 2781 observations – were added to the existing soil profiles to calibrate quantile regression forests (QRFs) using the Euclidean buffer distances from the sites as soil covariates. The SAWC was first mapped separately for different soil layers, and the mapping outputs were pooled to estimate the required SAWC. The uncertainty of the SAWC prediction was estimated from the estimated mapping uncertainties of the individual soil layers by an error propagation model using a first-order Taylor analysis.

The performances of the SAWC predictions and their uncertainties were evaluated with a 10-fold cross validation that was iterated 20 times. The results showed that the use of a quantile regression forest that was fed with auger hole observations and that used the Euclidean buffer distances as soil covariates considerably augmented the performances of the SAWC predictions (percentages of explained variance from 0.39 to 0.70) compared to the performance of a classical DSM approach, i.e., a QRF that solely used soil profiles and only environmental covariates (percentages of explained variance from 0.04 to 0.51). The analysis of the results revealed that the performances were also dependent on the spatial patterns of the different examined SAWCs and was limited by the observational uncertainties of the SAWCs determined from auger holes. The best performance tended to also provide the best view of the uncertainty patterns with an overestimation of uncertainty.

Despite these gains in performance, the cost-efficiency analysis showed that the augmentation of soil observations was not cost efficient because of the highly time-consuming manual data harvesting protocol. However, this result did not account for the observed gain in map details. Furthermore, the cost efficiency could be further improved by automation.

1. Introduction

Digital soil mapping (DSM) has been recognized as the appropriate solution to provide spatial soil information for land users, scientist communities and policy and decision makers in agriculture and the environment (McBratney et al., 2003; Sanchez et al., 2009). The principle of DSM is to predict a soil property or soil classes and the associated prediction uncertainty by determining the quantitative relationships between the soil information available over a limited set of locations and the spatial data reflecting the state factors of soil formation (envionmental covariates). DSM has now moved from a largely academic movement toward an operational activity (Minasny & McBratney, 2016, Arrouays et al, 2017).

However, the performances of DSM predictions of soil properties often exhibit more uncertainty than initially expected. For example, the percentages of explained variances of less than 0.5 were observed for 95%, 76%, 100% and 86% of the tested soil properties for DSM applications at the catchment scale (Nussbaum et al., 2018), at the regional scale (Vaysse and Lagacherie, 2015), at the national scale (Mulder et al., 2016), and at the global scale (Hengl et al., 2014), respectively.

These authors converged toward the conclusion that the density of soil observations used for calibrating the DSM models was the main factor that limited the DSM performances. Most of the soil information used as input in DSM applications has been either soil maps or the spatial sampling of sites with soil property measurements. The average densities used in most operational DSM applications have been low, e.g., 4-12 sites/km² (several study areas in Nussbaum et al., 2018), 0.07 sites/km² (Vaysse and Lagacherie, 2015), 0.03 sites/km² (Mulder et al., 2016), and 0.001 sites/km² (Hengl et al., 2014), which limits the performances of soil prediction, especially when the pattern of variation in the soil property is largely below the spacing of soil profiles (Vaysse and Lagacherie, 2015; Gomez and Coulouma, 2018). In addition, further experiments that consisted of varying the spatial density of soil input confirmed this analysis (Somarathna et al. 2017, Wadoux et al. 2019, Lagacherie et al, 2020). Consequently, it is of paramount importance to increase the density of soil inputs to improve the performance of DSM models in predicting soil properties (Voltz et al., 2020).

The most straightforward way to increase the density of DSM model soil inputs involves harvesting the legacy soil data that have not yet been stored in the existing soil databases. Arrouays et al. (2017) showed that during the period 2009-2015, the numbers of legacy soil profiles stored in global and national soil databases increased by 1,046% and 45%, respectively. However, they estimated that a large amount of soil legacy data can still be harvested. This is even more true in some areas across the world where soil surveying has been particularly active in the past.

For example, in southern France, the BRL irrigation company conducted detailed soil surveys over its irrigation perimeter between 1957 and 1992, which resulted in detailed soil maps, 25,000 soil profiles (5/km²) and 203,000 auger hole observations (31/km²). At this stage, such soil data have not yet been harvested and therefore cannot be used as input for DSM applications. However, this data has great potential for improving DSM performance and should be thoroughly examined.

In this paper, a spatially dense set of soil observations harvested from soil survey documents was tested for improving the performances of DSM models in mapping soil available water capacities for different rooting depths (0-30 cm, 0-60 cm, 0-100 cm) and at maximum observation depth, and the associated uncertainties. Our aim was to evaluate the cost-efficiency ratio of using such soil observations and to evaluate the added value of using euclidian buffer distances as additional inputs of DSM models as proposed by Hengl et al (2018). The study is conducted in the commune of Bouillargues, which is one of the communes included in the BRL irrigation perimeter.

2. The case study

2.1. The study area

This study took place in the administrative commune of Bouillargues in the Occitanie administrative French region (Figure 1). Located in southern France, Bouillargues covers 16 km² and is mainly devoted to vineyards, agricultural lands, forests, and scrublands.

Bouillargues has a Mediterranean climate characterized by a moderate average annual rainfall (600 mm) and dry and hot summers.

The study area is topographically split into two subregions with the large flat valley of the Vistrenque in the northern part and old fluviatile alluvium terraces belonging to the Nîmes “Costière” in the southern part. The two subregions have contrasting parent materials with i) loess and loamy clay deposition in the Vistrenque valley and ii) old alluvium in the Nimes Costière part, covered by some loess deposits. The contrast in parent materials induces variations in soils with i) fluvisols and calcisols developed in loess and loamy clay deposition, characterized by an absence of coarse fragments and a loamy texture, and ii) chromic luvisols developed in old alluvium terraces characterized by important coarse fragment contents and compacted clay accumulations (Figure 1).

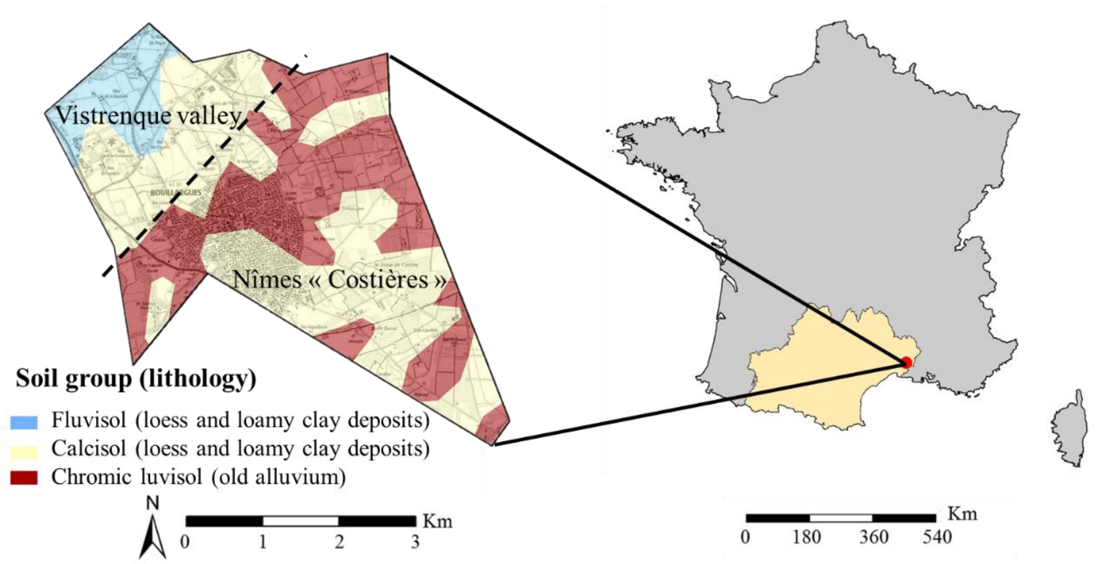


Figure 1. Location of the study area and distribution of soil samples

2.2. Soil data

2.2.1. History and content of the BRL soil database

The soil data of this study are a part of the soil survey led by the "Compagnie Nationale d'Aménagement de la Région du Bas-Rhône et du Languedoc" (CNARBRL) between 1957 and 1992 over the irrigated perimeter of this irrigation company, which covers 6,636 km². The objectives of this survey were to provide suitable soil information for i) improving the development master plan of the irrigation perimeter and estimating the surface area of arable and potentially irrigable lands and ii) supporting the cultural intensification made possible by irrigation, assessing the irrigation supply, and setting technical assistance for landholders to start irrigation and crop conversion.

The compilation of those studies resulted in a database of 228,000 soil observations with 25,000 soil profile descriptions and laboratory analyses (Figure 2) and 203,000 auger holes (Figure 3), which correspond to average spacings of 515 m and 181 m for the soil profiles and auger holes, respectively.

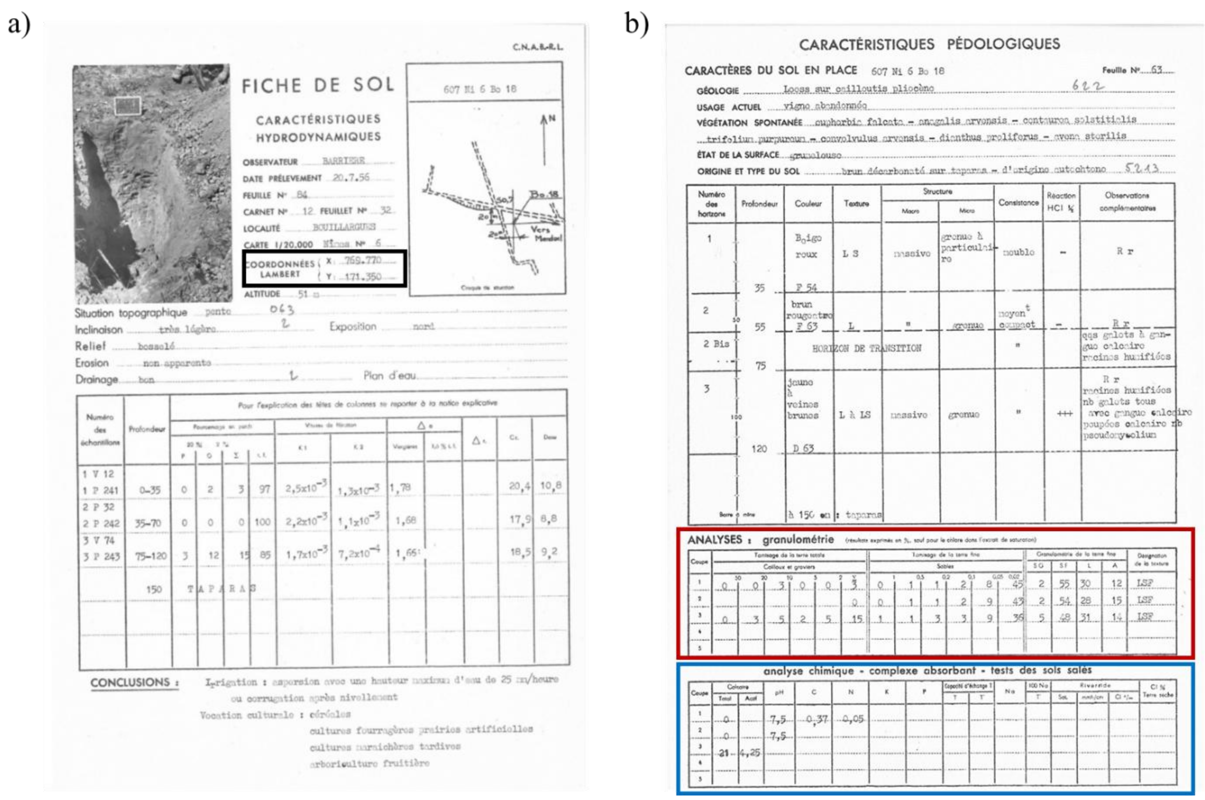
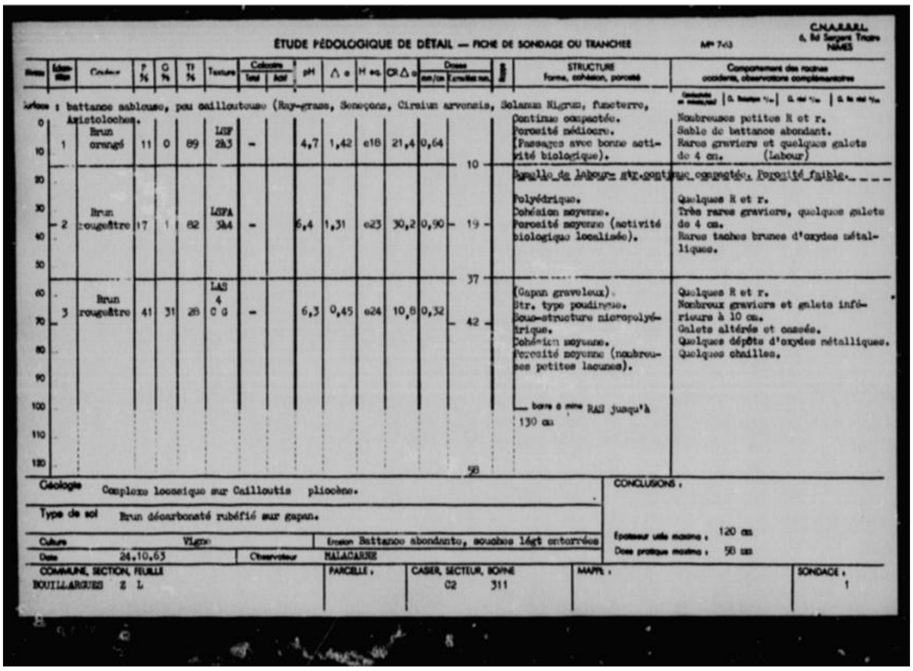


Figure 2. Soil profile a) horizon descriptions with geographical coordinates (black box) and b) laboratory analysis results, physical analysis (red box) and chemical analysis (blue box)



ÉTUDE PÉDOLOGIQUE DE DÉTAIL — FICHE DE SONDAGE OU TRANCHEE

C.N.A.R.B.R.L.
6, Bd Sergent Triaire
NIMES

Surface : battance sableuse, peu caillouteuse (Ray-grass, Seneçons, Cirsium arvensis, Solanum Nigrum, fumeterre, Aristoloches.

Horizon	Couleur	F %	G %	Tf %	Texture	Calcaire Total	Calcaire Actif	pH	Λ e	H eq.	CR Δ e	Doses		STRUCTURE Forme, cohésion, porosité	Comportement des racines accidents, observations complémentaires
1	Brun orangé	11	0	89	LSF 2h3	-		4,7	1,42	e18	21,4	0,64	10	Continue compactée. Porosité médiocre. (Passages avec bonne activité biologique).	Nombreuses petites R et r. Sable de battance abondant. Rares graviers et quelques galets de 4 cm. (Labour)
														Semelle de labour: str. continue compactée. Porosité faible.	
2	Brun rougeâtre	17	1	82	LSFA 3h4	-		6,4	1,31	e23	30,2	0,90	19	Polyédrique. Cohésion moyenne. Porosité moyenne (activité biologique localisée).	Quelques R et r. Très rares graviers, quelques galets de 4 cm. Rares taches brunes d'oxydes métalliques.
3	Brun rougeâtre	41	31	28	LAS 4 C G	-		6,3	0,45	e24	10,8	0,32	42	(Gapan graveleux). Str. type poudingue. Sous-structure micropolyédrique. Cohésion moyenne. Porosité moyenne (nombreuses petites lacunes).	Quelques R et r. Nombreux graviers et galets inférieurs à 10 cm. Galets altérés et cassés. Quelques dépôts d'oxydes métalliques. Quelques chailles.

barre à mine RAS jusqu'à 130 cm

Géologie : Complexe loessique sur Cailloutis pliocène.

Type de sol : Brun décarbonaté rubéfié sur gapan.

Culture : Vigne — Erosion : Battance abondante, souches légt enterrées

Date : 24.10.63 — Observateur : MALACARNE

CONCLUSIONS : Epaisseur utile maxima : 120 cm — Dose pratique maxima : 58 mm

COMMUNE, SECTION, FEUILLE : BOUILLARGUES Z L — PARCELLE : — CASIER, SECTEUR, BORNE : C2 311 — MAPPE : — SONDAGE : 1

Figure 3. Auger hole descriptions

2.2.2. Spatial sampling and georeferencing in the study area

Focusing on the commune of Bouillargues, the harvested dataset is composed of 2850 sites with soil observations that include 2781 auger holes and 69 soil profiles, which correspond to average spacings of 76 m and 500 m, respectively (Figure 4). Both the soil profiles and the auger hole observations were fairly evenly distributed over the study area; however, some gaps corresponded to urbanized areas or lands that were not expected to have any agricultural potential.

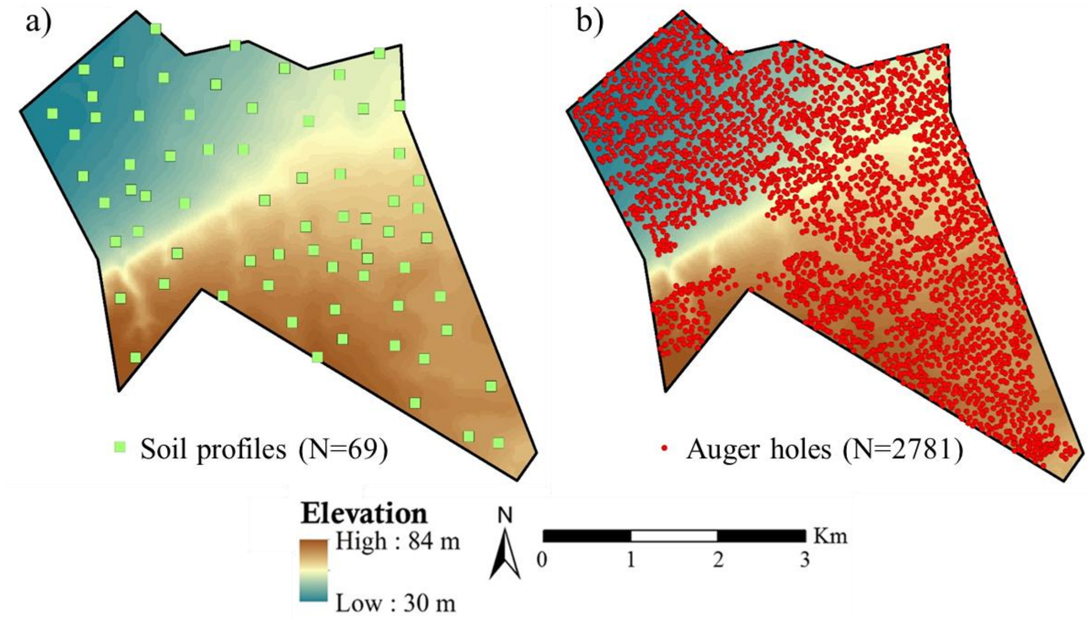
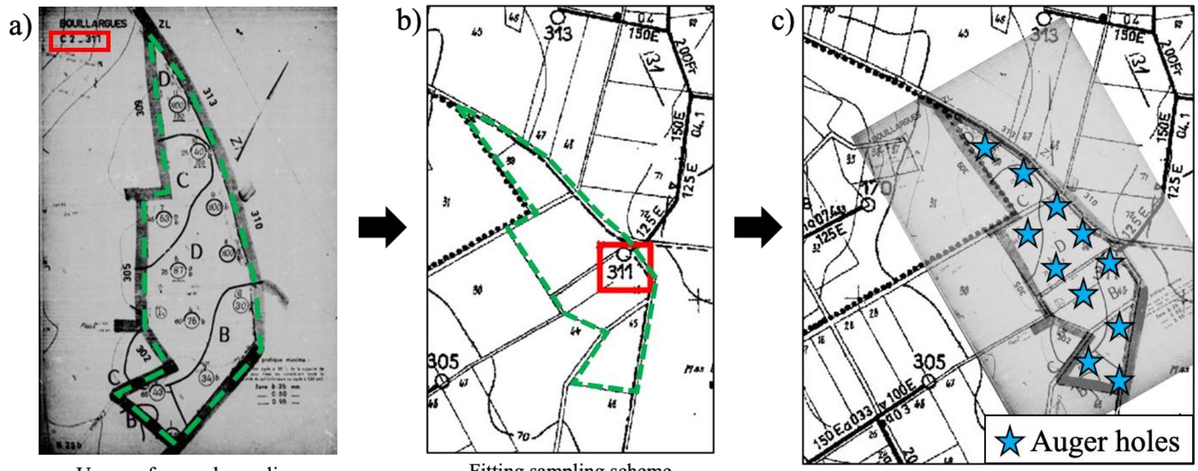


Figure 4. Spatial distribution of a) soil profiles and b) auger holes over the commune of Bouillargues

The soil profile data records included geographical coordinates (Lambert III, black box in Figure 2a), whereas manual preprocessing was necessary for georeferencing the auger holes.

The auger holes were initially located through a non-georeferenced map representing the local sampling scheme (Figure 5a). Each sampling scheme corresponded to an area of water distribution supplied by an irrigation water access point of the BRL irrigation network. This access point was georeferenced and could be positioned onto a georeferenced former cadastre (red box in Figure 5b). To acquire the coordinates of the auger holes, the sampling scheme was first located in the georeferenced cadastre using the coordinates of the irrigation water access point. Its boundaries were then positioned (green dashed perimeter on Figure 5b) using the geometry of the parcels and communication paths. Finally, each auger hole was manually positioned onto the georeferenced cadastre (blue stars on Figure 5b) using the sampling scheme (Figure 5a), and the coordinates of the auger holes were obtained using the coordinates acquisition tool of BRL’s web-GIS (Figure 5c).



Ungeoreferenced sampling scheme

Fitting sampling scheme boundaries on georeferenced former cadastre

Auger holes adjustment and coordinates acquistion

Figure 5. Fitting the non-georeferenced sampling scheme of auger holes in the georeferenced former cadastre

2.2.3. Soil available water capacity determinations at sites with soil observations

This study took the mapping of soil available water capacity (SAWC) as an example of applying DSM. SAWC refers to the capacity of the soil to store water for plant growth (Veihmayer and Hendrickson, 1927). This functional property plays a key role in many ecosystem services, such as food production, soil drought or climate and gas regulation. Consequently, it is a crucial parameter used in land evaluations and recently in ecosystem services assessments (Dominati et al., 2014). Information about the SAWC distribution in space is essential for planning and management in agriculture and for ecological modeling. In the present example, SAWC was required for fulfilling the irrigation objectives evoked above (section 2.2.1). Currently, SAWC is computed in the literature as follows (Cousin et al., 2003):

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \quad (1)$$

where SAWC is the soil available water capacity (cm), dh_i = the thickness of the ith horizon (cm), bd_i = the bulk density (g/cm^3) of the ith horizon, st_i = the coarse fragment content of the ith horizon (% volumetric), and θr_i and θw_i are the gravimetric soil water contents at field capacity (i.e., the soil water content that remains in the soil after water has drained due to gravitational force) and the permanent wilting point (i.e., the soil water retained so strongly that it is no longer available for plant roots, so plants wither and cannot recover their turgidity) of the ith horizon (cm^3.cm^{-3}), respectively.

Historically, the CNARBRL had a different approach for expressing the water retention term of the fine earth, i.e., $(\theta r_{\ i} - \theta w_i)$, which leads to the following equation:

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - \ st_i}{100}\right) * (b_i * EqW_i) \qquad (2)$$

The equivalent water content (EqWi) corresponds to $\theta r_{\ i}$ of Eq. 1, and the textural coefficient b_i is an expression of the water content at the permanent wilting point that weights EqW_i to account for the water content that is not available for the plant (i.e., beyond the wilting point, defined as θw_i in Eq. 1).

The values of bd_i and EqW_i were measured at each soil profile; bd_i was determined in the field following the Vergières protocol (Bourrier, 1965) but was estimated as 1.6 times the mass fraction of the fine earth from the ensemble coarse fragment and fine earth, when the coarse fragment phase of the soil sample was too important to perform the Vergières protocol (Legros, 1996).

The EqW_i of sieved samples was determined in the laboratory using a centrifuge apparatus set at 100 kPa (pF = 3.0), a reference pressure that was considered, at the time of the CNARBRL soil survey, as yielding the best approximation of the water content at the field capacity (see section 2.2.1) (Baize and Jabiol, 1995). The EqW_i values were estimated on auger hole observations by local pedotransfer functions using the field estimated textural classes.

The b_i coefficient was determined both on soil profiles and on auger hole observations by a local pedotransfer function using the textural classes determined from granulometric analyses and field estimation, respectively, for soil profile and auger hole observations.

The coarse fragment content and the horizon thicknesses of Eq. 2 were retrieved from the descriptions of the physical analyses and descriptions of the soil profiles and of the auger hole observations, respectively (Figures 1 and 2). Different total soil thicknesses (i.e., $\sum_{i=1}^{n} dh_i$) were considered to determine the different rooting depths related to the different possible crops of the study area (from market gardening to vineyard passing by annual crops). In addition to the maximum soil thicknesses given by the soil observations that were considered for calculating the maximum soil available water capacity (SAWCmax), restricted thicknesses of 30 cm, 60 cm and 100 cm were then considered, leading to different restricted SAWCs, denoted further as SAWC30, SAWC60, and SAWC100.

It must be noted that both the profiles and auger holes had limited observation depths of 140 and 120 centimeters, respectively, which may cause underestimations of SAWCmax.

2.3. Environmental covariates

The DSM approach, as formalized by the scorpan model (McBratney et al., 2003), considers quantitative relationships between a target soil property and environmental variables, which are also known as “covariates”.

The selection of environmental covariates depends on two criteria: i) they could be derived from geodatasets freely available at least at the French national level, and ii) they have a logical and process-based relationship with soil properties according to the literature.

Following these criteria, we derived covariates related to the scorpan model component, i.e., topography, organisms, and parent material, that regroups the major landscape types across the study area. Climate data were not considered in this study since we did not find any climate data at a spatial resolution fine enough to represent the climate variations over such a small area. The relief component was described by a set of geomorphometric indicators currently considered in DSM studies: elevation, slope, aspect, multiresolution valley bottom flatness (MRVBF), multiresolution ridge top flatness (MRRTF), topographic wetness index (TWI), topographic position index, plan curvature and profile curvature. These indicators were derived from the French altimetry database (BD ALTI, 25 m resolution) digital elevation model (DEM). They were computed using the SAGA GIS software (Böhner et al., 2006) and his Terrain Analysis procedures.

Organisms and parent materials were derived from the Landsat 7 imagery and geological map, respectively, and were both resampled at the native resolution of the DEM (i.e., 25 m). Additionally, parent material covariates were developed by Vaysse and Lagacherie (2015) from the geological map (1:50,000) qualitative descriptions to quantitative indicators describing the hardness, mineralogy and texture of alteration materials.

Table 1. Exhaustive categorical and continuous covariates

Variables	**Abbreviation**	**Resolution/Scale**	**Source**	**Soil-forming factor**[1]	**Type**[2]
Topography					
Elevation	ELEV	25 m	BD ALTI	r	Q
Multiresolution Valley Bottom Flatness	MRVBF	25 m	BD ALTI	r	Q
Slope	SLOPE	25 m	BD ALTI	r	Q
Topographic Wetness	TWI	25 m	BD ALTI	r	Q

Index					
Plan Curvature	PLANCURV	25 m	BD ALTI	r	Q
Profile Curvature	PROCURV	25 m	BD ALTI	r	Q
Multiresolution Ridge Top Flatness	MRRTF	25 m	BD ALTI	r	Q
Topographic Position Index	TPI	25 m	BD ALTI	r	Q
Geology					
Hardness	HARDNESS	25 m	Geological map/soil profile	p	C
Texture	TEXTURE	25 m	Geological map/soil profile	p	C
Mineralogy	MINERALOGY	25 m	Geological map/soil profile	p	C
Organisms					
Land use	LANDUSE	25 m	Landsat 7	o	C

[1]: SCORPAN factors (o = organisms, r = relief, p=parent material)

[2]: Q = quantitative, C = categorical

2.4. Acquisition process and cost assessment

In section 2.2., we presented the main difference in using soil profiles and auger holes in a DSM application, i.e., the accessibility of the data. While soil profile acquisition is quite straightforward, i.e., recording soil data and locations, auger hole acquisition is more complicated as the locations are not directly available and manual georeferencing is required, thus, the acquisition process is longer. In Table 2, we provide the main information about the acquisition process for soil profiles and auger holes. As the number of auger hole observations is substantially larger than the number of soil profiles and take longer to record, we provided an assessment of the cost of soil data acquisition.

Table 2. Information to assess the cost of the acquisition process

	Auger holes	Soil profiles
Recorded time of soil properties* (min/observation)	0.8	0.8
Recorded time of geo-localizations* (min/observation)	2.2	0.2
Number of observations	2721	69

*Computed from timed sessions of harvesting

To compute the cost of the acquisition process, we applied the following formula using the information in Table 2:

$$Cost = \left(\frac{N * rec\_time}{Daytime}\right) * Sal. \qquad (5)$$

With N the number of harvested soil observations, rec_time the recorded times of harvesting a given soil observation in mn (see table 2), Daytime is 1440 (number of mn in a day) and Sal is the daily salary of the harvester.

3. Methods

3.1. DSM models for soil profiles

In this study, we used several mapping models derived from the random forest algorithm. Hereafter, we provide a general description of random forest and its derivatives used in this study.

3.1.1. Random forest

Random forest models (RF) (Breiman, 2001) are an ensemble learning method for both classification and regression. A forest, i.e., an ensemble of randomized decision trees, is built and trained based on a bootstrap approach. Individual trees are built using the principle of recursive partitioning. "*The feature space is recursively split into regions containing observations with similar response value*" (Strobl et al., 2009). The predictions of the individual trees are finally averaged to obtain a single prediction.

3.1.2. Quantile regression forest

The quantile regression forest algorithm (QRF) (Meinshausen, 2006) is an extension of random forests that has become one of the most commonly used algorithms in DSM studies (Hengl et al., 2015; Ugbaje and Reuter, 2013; Vaysse and Lagacherie, 2017). As a RF, QRF provides an ensemble prediction based on *n* regression trees. However, while RF provides solely the conditional mean, QRF supplies the whole conditional distribution of the target variable by keeping all observations at the terminal nodes. This allows us to infer estimates for the conditional quantiles (Meinshausen, 2006). More details on QRF can be found in Meinshausen (2006).

QRF was performed with the ranger package, which is a fast implementation of Breiman's random forest and Meinshausen's quantile regression forest for big data (Wright and Ziegler, 2017). QRF was run with the default parameters given by ranger.

3.2. Mapping models for dense spatial sampling

The usual applications of RF and its derivative to DSM only exploit the relationships between the soil properties to be predicted with landscape elements characterized by a set of covariates derived from the available spatial data. However, they do not consider the spatial relationships between sites or spatial autocorrelation, which allows the spatial interpolations of a given soil property between sites. This can lead to suboptimal predictions and possibly systematic over- and underestimation of predictions, especially if the target variable is spatially autocorrelated and if point patterns show clear sampling bias (Hengl et al., 2018). In the case of dense sampling, such spatial interpolation can be of great interest to overcome the limitations of landscape covariates for predicting soil properties (Lagacherie et al, 2020).

To correct the non-spatial approach of RF and its derivative, Hengl et al. (2018) proposed adding new covariates that consider the locations of the sites. These covariates are defined as the Euclidean buffer distances from the observation sites. To limit the number of covariates and the computing time in the case of a large dataset (> 1,000 sites), these distances to the nearest points were not calculated for each individual observation site but for *n* equal classes (from low to high AWC values). As RF is sensitive to the number of classes (Hengl et al., 2018), we performed a trial and error process, which was conducted to choose different classes according to the maximal soil thickness considered and to the density scenario (number of classes varying between 6 and 15). For each targeted SAWC, a map was generated. In this DSM model, we considered soil profile and auger hole observations indifferently as soil inputs, omitting their possible differences of uncertainty on the SAWC determinations. This model will be denoted further QRF_{dist}. Euclidean buffer distance mapping was performed using the *GSIF* package (Hengl, 2019).

3.3. Inference trajectories

Since we aimed to map SAWC, which is a soil indicator involving several soil properties and several soil depths, it could be estimated following various possible inferences following the order with which “combining primary soil properties”, “aggregating soil layers across depths” and “mapping” were performed to provide the SAWC (Styc and Lagacherie, 2019). Styc and Lagacherie (2019) experienced a total of 18 inference trajectories throughout Languedoc-Roussillon that were performed to obtain the most appropriate SAWC map. From this study, we considered the best-performing inference trajectory, i.e., we mapped the first AWC of four separate layers (0-30, 30-60, 60-100 and 100-200 cm) and then aggregated the maps of the four soil layers to obtain the final SAWC map.

3.4. Uncertainty analysis using error propagation

In this section, we provide the main details of uncertainty assessment using propagation error. More details of the procedure can be found in (Román Dobarco et al., 2019, Styc and Lagacherie, submitted).

The selected inference trajectory, i.e., SAWC estimated as the aggregation of AWC predicted at four depth soil layers, required an error propagation to estimate the variance in SAWC, considered as a proxy of the uncertainty prediction of the target variable (Heuvelink et al., 1989). In this study, we used a first-order Taylor expansion to calculate the error variance of SAWC that results from the error variances of its components (here, the different mapped AWC for the four considered soil layers). This calculation involved i) the error variances of AWC for each soil layer obtained from the conditional distributions provided by QRF for each predicted location (Meinshausen, 2006) and ii) the correlation coefficients between the errors at each soil layer provided by the mapping residuals. Then, the estimate of the SAWC variances was translated into a 90% prediction interval, assuming a normal distribution, by:

$$CIL_i = \hat{y}_i \pm 1.645\ \sigma_{\hat{y}_i} \tag{6}$$

where CIL_i is the interval limits of the prediction, $\hat{y}_i$ is the mean of the distribution, $\sigma_{\hat{y}}$ is the standard deviation and 1.645 is the Student's coefficient for a 90% confidence interval estimation.

Error propagation was performed using the *propagate* R package (Spiess, 2018).

3.5. The experiment

The goal of the experiment was two-fold: i) to evaluate the efficiency of the DSM model proposed for dealing with dense spatial sampling of auger holes (QRF_{dist}) and ii) to evaluate the cost-efficiency ratio of using auger hole observations with increasing densities.

For that, QRF_{dist} was applied to different soil input scenarios with increasing numbers of auger holes. The performances of the QRF_{dist} were compared with those of a baseline QRF application that did not consider any spatial relation between the sites, as practiced in most DSM applications. The four SAWCs presented in section 2.2.3 were considered. In the following, we provide some details about the sampling strategy for selecting auger holes, the evaluation protocol and the cost-benefit analysis.

3.5.1. The sampling procedure of auger holes

Different data scenarios were considered, all of which included all the available soil profiles as inputs. An increasing number of auger holes were sampled from the available set and added to the soil profiles in the soil input datasets (from 10% to 100% of the auger hole observations each 10%, e.g., average spacing of 278 m, 556 m, 834 m, 1112 m, 1391 m, 1669 m, 1947 m, 2225 m, 2503 m and 2781 m).

At each step, the auger holes were selected using a stratified random sampling technique using compact geographical strata (Walvoort et al., 2010), as recommended by (Brus et al., 2011). Thirty-three geographical strata of 0.5 km^2 were considered. Spatial stratification sampling was performed using the *spcosa* R package (Walvoort et al., 2018).

3.5.2. Evaluation protocol

The performance of the SAWC DSM models was evaluated by k-fold cross validation. This evaluation procedure consisted of randomly dividing the data into k subsets. Then, the holdout method was repeated k times such that one of the k subsets was used as the validation set in each repetition, while the other k-1 subsets were combined to form the calibration set. Following this procedure, every data point was included in a calibration set k-1 times. In this study, we selected k = 10 and to increase the robustness of the evaluation, the 10-fold cross validation was iterated 20 times. The k-fold cross validation was performed using *cvTools* (Alfons, 2012).

To avoid uncertain estimations of the model performances due to the inherent uncertainty of SAWC estimations from the auger hole observations, the evaluation protocol presented hereafter was solely applied to the soil profiles.

To evaluate the prediction performances, we used classic performance indicators, e.g., the mean square error skill score (Nussbaum et al., 2018), which has the same interpretation as the percentage of variance explained by the model, the root mean square error (RMSE) and the bias.

Furthermore, we evaluated the estimation of the prediction uncertainty using the prediction interval coverage probability (PICP; Shrestha and Solomatine, 2006) and error-predicted uncertainty plots. The PICP was computed as follows:

$$PICP = \frac{count(LPL_i \leq y_i \leq UPL_i)}{n} \times 100 \tag{7}$$

where n is the total number of observations in the validation set, and the numerator counts if the observation y_i fits within the prediction limits prior to estimation by the error propagation method. For a 90% confidence level, which is usually chosen in DSM studies (Arrouays et al., 2014b), the uncertainty is optimally predicted when the PICP value is close to 90%.

The PICP provides an assessment of the overall uncertainty prediction bias (underestimation or overestimation) but does not tell anything about the ability to map differences in uncertainty across the study area. The PICP was therefore completed by error-predicted-uncertainty estimations that materialized the evolution of the cross validation RMSE with the widths of the predicted confidence intervals. To remove noise, the RMSEs were averaged per quartile of prediction interval widths denoted “low/fairly low/fairly high/high predicted uncertainty”. It was expected that the RMSE would increase from low to high predicted uncertainty.

3.5.3. The cost efficiency of SAWC Digital Soil Mapping

Soil data need to be recorded, but this process can be time consuming and therefore costly. To answer the question, “Is all the data necessary to reach quality predictions?”, we set two indicators to assess i) the cost of a unit of gained RMSE and ii) the relative cost efficiency, which were both calculated for each percentage of auger holes added to the soil profiles. The cost of a unit of RMSE was evaluated using the following equation (Eq. 8):

$$Err_{cost} = \frac{cost_i}{RMSE_i} \quad (8)$$

where Err_{cost} is the cost of a unit of RMSE (in €/cm) and $RMSE_i$ is the root mean square error of the combination of i% of auger hole and soil profile datasets.

The relative cost efficiency was assessed following the recommendation of Kish (1965) used by (Viscarra Rossel and Brus, 2018, Eq. 9):

$$CE_r = \frac{cost_{ref} * RMSE_{ref}}{cost_i * RMSE_i} \quad (9)$$

where CE_r is the relative cost-efficiency ratio, $cost_{ref}$ and $RMSE_{ref}$ are the cost and the error of a reference design, respectively, here using solely soil profiles in the SAWC DSM, and $cost_i$ and $RMSE_i$ are the cost and the error, respectively, of the combination of i% of auger hole observation and soil profiles. A CE_r larger than one reveals more efficient sampling than the reference (Viscarra Rossel and Brus, 2018).

4. Results

4.1. Preliminary results

In Figure 5, we present the distributions of SAWC30, SAWC60, SAW100 and SAWCmax for the soil profiles (left panel of Figure 5) and auger holes (right panel of Figure 5). We first observed that the distributions of SAWC regardless of the considered soil depth were bimodal for both the soil profiles and auger holes, with i) a higher peak for higher values of SAWC30 and SAWC60 and with ii) a higher peak for lower values of SAWC100 and SAWCmax. Additionally, it is worth noting that both the SAWC ranges and the means of the auger holes were systematically greater than those of the soil profiles. This could be explained by i) possible underestimations of coarse fragments by visual determinations on very small volumes using auger holes compared to real measurements of coarse fragments on larger volumes using soil profiles and ii) possible biases of the field determination of textural class on auger holes compared with laboratory analyses performed on soil profiles.

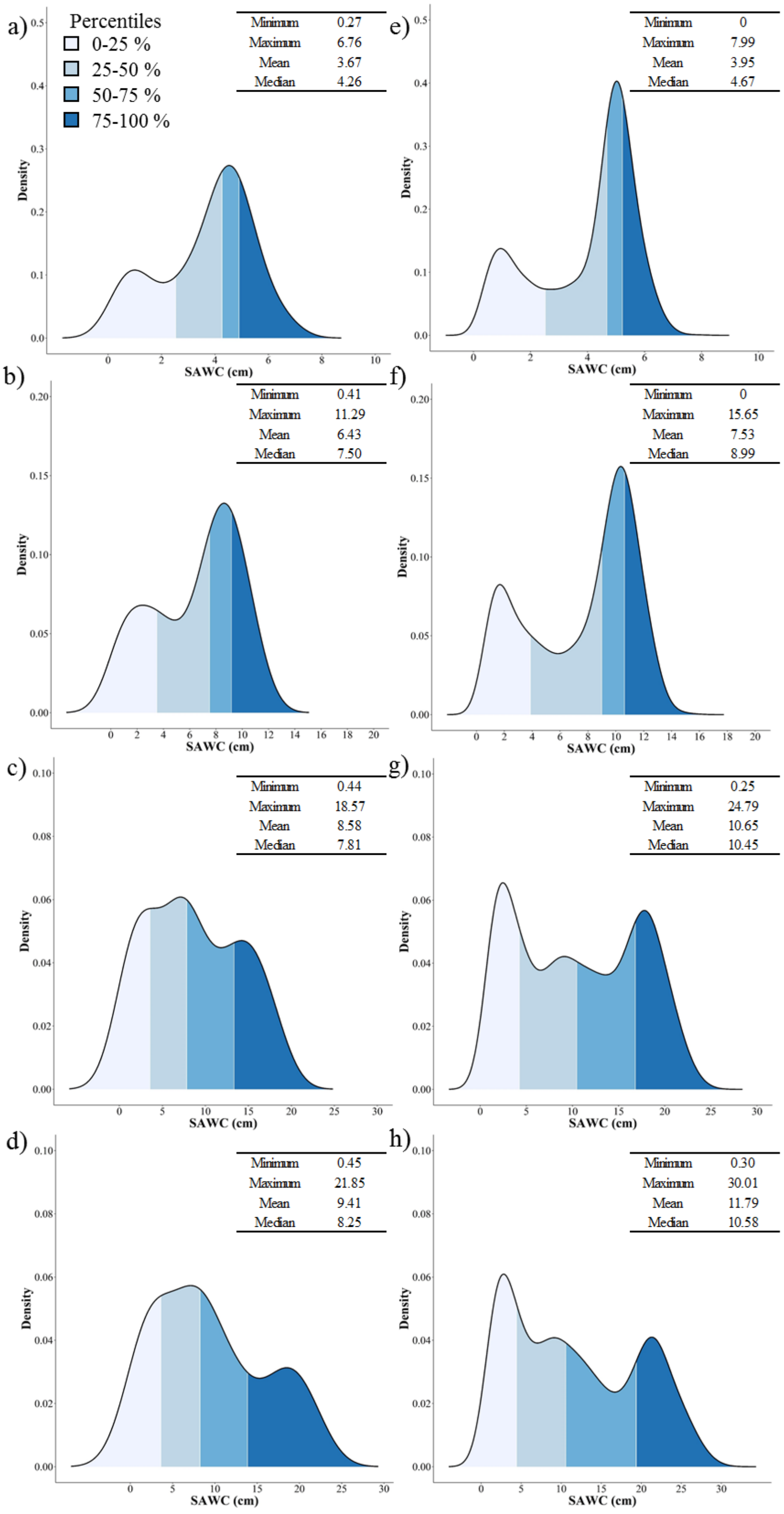


Figure 5. Distributions of the soil available water capacity of soil profiles at a) 0-30 cm, b) 0-60 cm, c) 0-100 cm and d) 0-depth$_{max}$ and of auger holes at e) 0-30 cm, f) 0-60 cm, g) 0-100 cm and h) 0-depth$_{max}$

In addition, empirical variograms and their fitted models were computed using the *gstat* package (Pebesma, 2004) both from the soil profile data (Figure 6, left panel) and from the auger hole data (Figure 6, right panel), and for the different considered SAWC (lines of Figure 6). The Spatially structured variance ratio (SSVR, Eq. 10), which estimated the portion of the variance that was spatially structured, was computed from the variograms as follows:

$$SSVR = 1 - \left({}^{nugget}/_{variance} \right) x\,100 \quad (10)$$

First, we noted that the variogram of the SAWC determined from auger hole observations exhibited clear spatial structures regardless of the maximal depth (SSVR ranging from 66% to 76%). The variograms showed a mix of short-range spatial structures (fitted ranges between 332 and 341 m) and large-range structures (fitted ranges exceeding 30 km). Conversely, the variograms of SAWC30 and SAWC60 determined from the soil profile empirical variogram exhibited less clear spatial structures (SSVR of 25% and 33%), whereas a clear structure appeared for SAWC100 and SAWCmax (SSVR of 82% and 89%). Because of their larger spacing, the soil profiles did not allow us to see the short-range spatial structures revealed by the auger hole observations. Additionally, significant decreases in nuggets were observed from the variograms of SAWC30 and SAWC60 processed from profiles to those processed from auger holes. This decrease can be interpreted as the result of increasing sampling densities that better captured the short-range spatially structured variance that was otherwise included in the profile variogram nuggets. It is interesting to note that the converse occurred for SAWC100 and SAWCmax. The probable increase in the uncertainty of observations with depth due to the difficulties in observing deep horizons from auger holes yielded a nugget increase that largely counterbalanced the effect of the sampling density evoked previously.

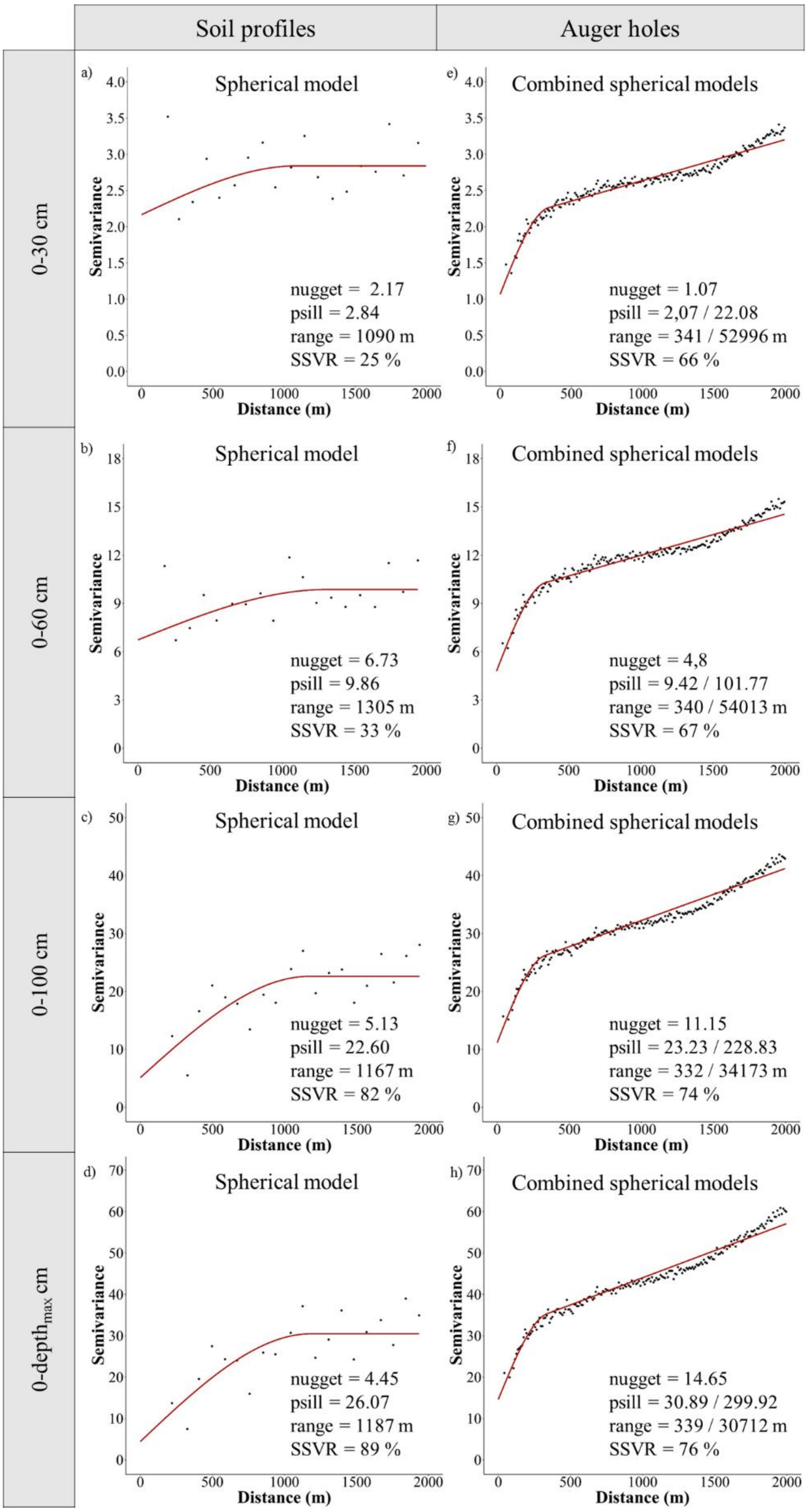


Figure 6. Empirical variograms computed for SAWC using 69 soil profiles at a) 30 cm, b) 60 cm, c) 100 cm and d) 200 cm and using 2781 auger hole observations at e) 30 cm, f) 60 cm, g) 100 cm and h) 200 cm, and their theoretical variograms.

4.2. Comparing DSM model prediction and uncertainty prediction performances

Table 3 shows the prediction and the uncertainty prediction performances of the two considered DSM models in predicting the SAWCs at four different depths. Only the extreme data scenario, i.e., no auger hole vs. the whole set of auger holes, is shown.

First, better performances of SAWC predictions were generally obtained by adding the auger hole observations, with the noticeable exceptions of the predictions of SAWC60, SAWC100 and SAWCmax using a classical (nonspatial) QRF. When using QRF_{dist}, the performance increases by adding auger hole observations tended to decrease as the maximum considered depth increased.

Additionally, using QRF_{dist} that included geographical information led to better prediction performances regardless of the SAWC only when the auger hole observations were added to the soil profiles. Otherwise, (i.e., when only the soil profiles were used for calibrating the model), using QRF yielded equal or slightly better prediction performances.

Concerning the ability of the models to provide unbiased estimates of prediction uncertainty, as measured by the PICP, larger PICP values were obtained with QRF_{dist} than with QRF, except for the PICP for SAWC100 with only soil profiles. Furthermore, the effects of including auger holes in QRF calibration were different according to the selected model: the PICP decreased when QRF was selected, whereas the PICP increased when the QRF_{dist} model was selected. As far as the closeness to the nominal value of 90% is concerned, better results were generally obtained when the auger hole observations were not used, with the noticeable exception of the SAWC30 predictions using QRF. Furthermore, QRF_{dist} had more PICP values close to the 90% nominal value (< 2%) than did QRF (4 out of 8 vs. 1 out 8).

Table 3. Prediction and uncertainty prediction performances of SAWC using multiple DSM models

DSM models		QRF				QRF_{dist}			
SAWC	Auger holes portion (%)	SS_{MSE}	RMSE (cm)	Bias (cm)	PICP (%)	SS_{MSE}	RMSE (cm)	Bias (cm)	PICP (%)
SAWC30	0	0.04	1.66	0.17	86	-0.02	1.71	0.32	85
	100	0.38	1.34	0.49	86	0.49	1.22	0.37	90
SAWC60	0	0.33	2.74	1.08	87	0.3	2.79	0.35	89
	100	0.32	2.76	1.28	83	0.54	2.26	0.82	93
SAWC100	0	0.55	3.73	-0.47	92	0.46	3.97	0.22	90
	100	0.43	4.06	1.82	85	0.63	3.27	1.09	95
SAWCmax	0	0.61	4.01	-0.68	90	0.53	4.41	-0.56	91
	100	0.54	4.37	1.88	85	0.7	3.54	0.18	96

As expected, the averaged RMSE tended to increase with the widths of the confidence intervals predicted by QRF_{dist} (Table 4), which demonstrated the overall validity of the uncertainty predictions. However, non-monotonous increases were observed for the SAWC predictions at small depths that also exhibited the weakest performances (Table 3). This non-monotonousness was clearer when the auger hole observations were added. Similar trends were observed for the confidence interval widths predicted by QRF (results not shown).

Rooting depth (cm)	Uncertainty	RMSE (cm)	
		Soil profiles	Soil profiles and auger holes
30	Low	1.09	1.31
	Fairly low	1.25	0.79
	Fairly high	2.75	1.10
	High	1.9	1.59
60	Low	2.31	1.25
	Fairly low	2.24	2.02
	Fairly high	2.81	3.25
	High	3.46	2.08
100	Low	2.81	1.52
	Fairly low	2.82	2.81
	Fairly high	3.49	3.69
	High	5.71	4.32
Maximum observation depth	Low	3.07	2.24
	Fairly low	2.88	2.82
	Fairly high	4.55	4.20
	High	6.09	4.37

Table 4. Error-predicted uncertainty results of QRF_{dist} using only soil profiles and using soil profiles and auger hole observations for predicting SAWC at multiple depths

4.3. Spatial distribution of the SAWC and its associated uncertainty

All the predicted maps of SAWC (Figure 7) exhibited spatial patterns of variation that were globally in accordance with the lithological variations shown in Figure 1. The highest values of SAWC were predicted in the northeastern section of the study area with fluvisols developed on loess. The smallest values corresponded to chromic luvisols developed on the old stony alluvial deposits.

The spatial pattern became increasingly clear and contrasted as the considered soil depth for calculating the SAWC increased (from the top to the bottom of Figure 7). The incorporation of auger holes (from the left to the right column in Figure 7) led to i) an increase in the predicted variabilities of the SAWC, leading to more contrasted patterns regardless of the predicted SAWC; ii) an increase in the spatial resolution of the SAWC pattern delineations, showing very fine details of variation; iii) the removal of some obvious artifacts of the map of SAWC100 obtained from the soil profiles (Figure 7c); and iv) the addition of some artifacts (isolated pixels) in the SAWC30 and SAWC60 maps (Figure 7e and 7f).

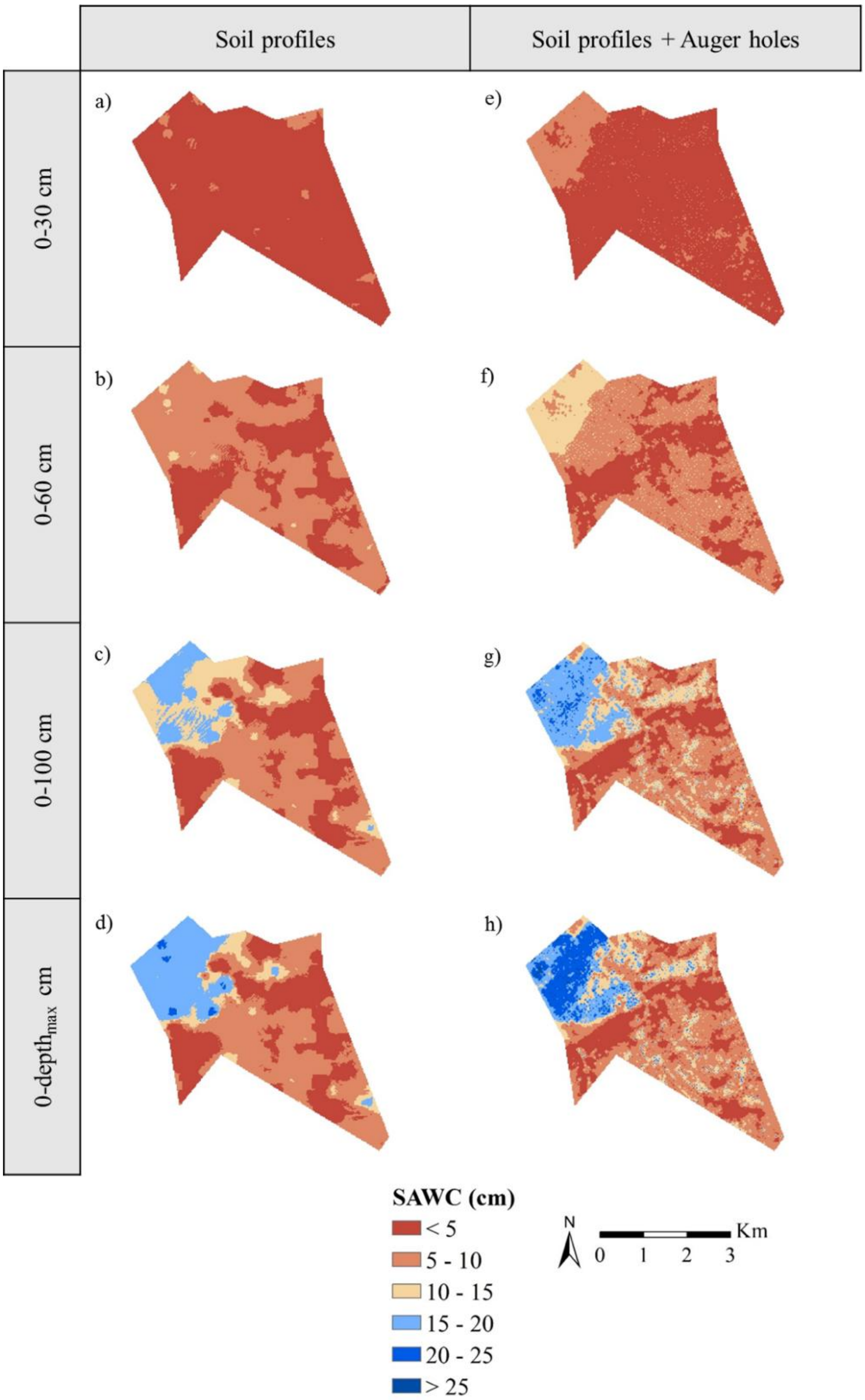


Figure 7. Predicted maps of SAWC over Bouillargues using QRF_{dist} with soil profiles for predicting a) SAWC30, b) SAWC60, c) SAWC100, and d) SAWCmax and using QRF_{dist} with soil profiles and auger hole observations for predicting e) SAWC30, f) SAWC 60, g) SAWC100, and f) SAWCmax

The uncertainty maps of SAWC predictions (Figure 8) obtained from the QRF_{dist} model exhibited spatial patterns that were both complex and very contrasted across the predicted SAWCs and soil inputs. When examining the variations between quartiles of predicted uncertainty that looked significant according to the error-predicted uncertainty results (Table

4), some of the maps revealed strong spatial pattern similarities with those of some uncertainty drivers, i.e., the SAWC30 uncertainty map using soil profiles (Figure 8a) with the lithology map (Figure 1), SAWC100 map using soil profiles (Figure 8c) with the spatial density of soil profiles that is observable on the map of soil profiles (Figure 2a), SAWC30 uncertainty map using auger hole observations (Figure 8e) with the spatial density of auger hole observations that is observable on the map of auger hole observations (Figure 2b), SAWCmax uncertainty map using auger hole observations (Figure 8h) with the predicted map of SAWCmax. The other uncertainty maps (Figure 8b, 8d, 8f) showed less interpretable patterns, with probably mixed impacts of the above evoked drivers.

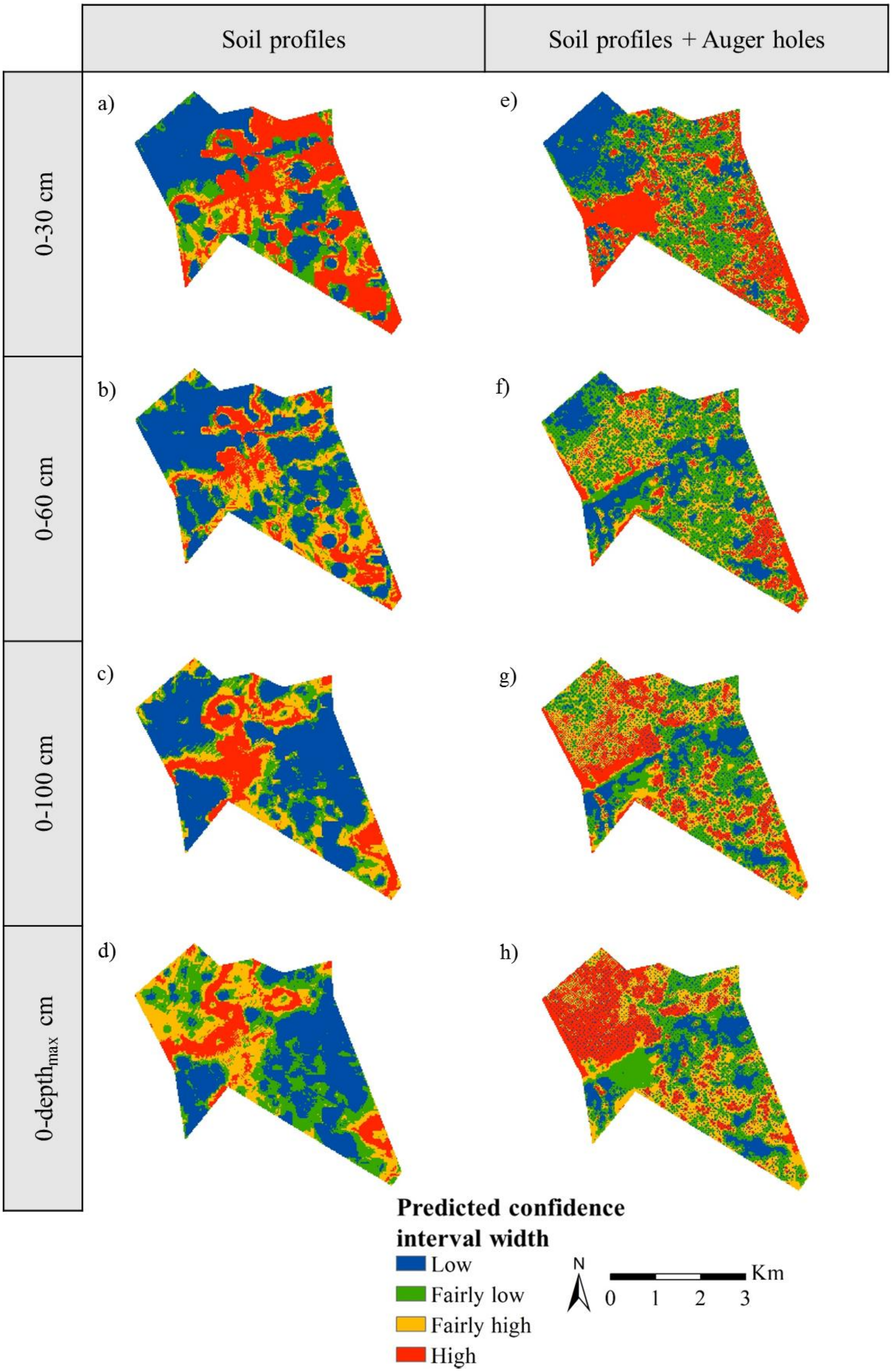

Soil profiles
Soil profiles + Auger holes
0-30 cm
0-60 cm
0-100 cm
0-depth$_{max}$ cm
a)
b)
c)
d)
e)
f)
g)
h)
Predicted confidence interval width
Low
Fairly low
Fairly high
High
N
0 1 2 3 Km

Figure 8. Predicted uncertainty maps of SAWC prediction over Bouillargues presented by the classes estimated from the quartiles of the validation distribution using QRF_{dist} with soil profiles for predicting SAWC at a) 30 cm, b) 60 cm and c) 100 cm; QRF_{dist} with soil profiles and the whole set of auger hole observations in covariates set for predicting SAWC d) 30 cm, e) 60 cm and f) 100 cm

4.4. Comparing the spatial densities of auger hole observations

In Figure 9, we present the evolution of the SS_{MSE} with the increasing number of auger hole observations in the calibration process. The density in the number of observations/km^2 is also expressed as the average spacing between observation sites, which means that the density increases as the average spacing decreases. The average spacing between observation sites was estimated as follows:

$$Average\ spacing = \sqrt{\frac{total\ area}{size}} \tag{11}$$

As already observed from Table 3, the general trend was an increase in performance as the number of auger hole observations increased regardless of the maximal depth at which the SAWCs were calculated. However, some local decreases in performance were observed, e.g., on SAWC60 and 100 predictions when adding 10% auger holes or on SAWC100 and SAWCmax predictions when passing from 20 to 30% auger holes. Conversely, the addition of 10% to 20% auger holes and 60% to 70% auger holes seemed beneficial for all predictions of the SAWC.

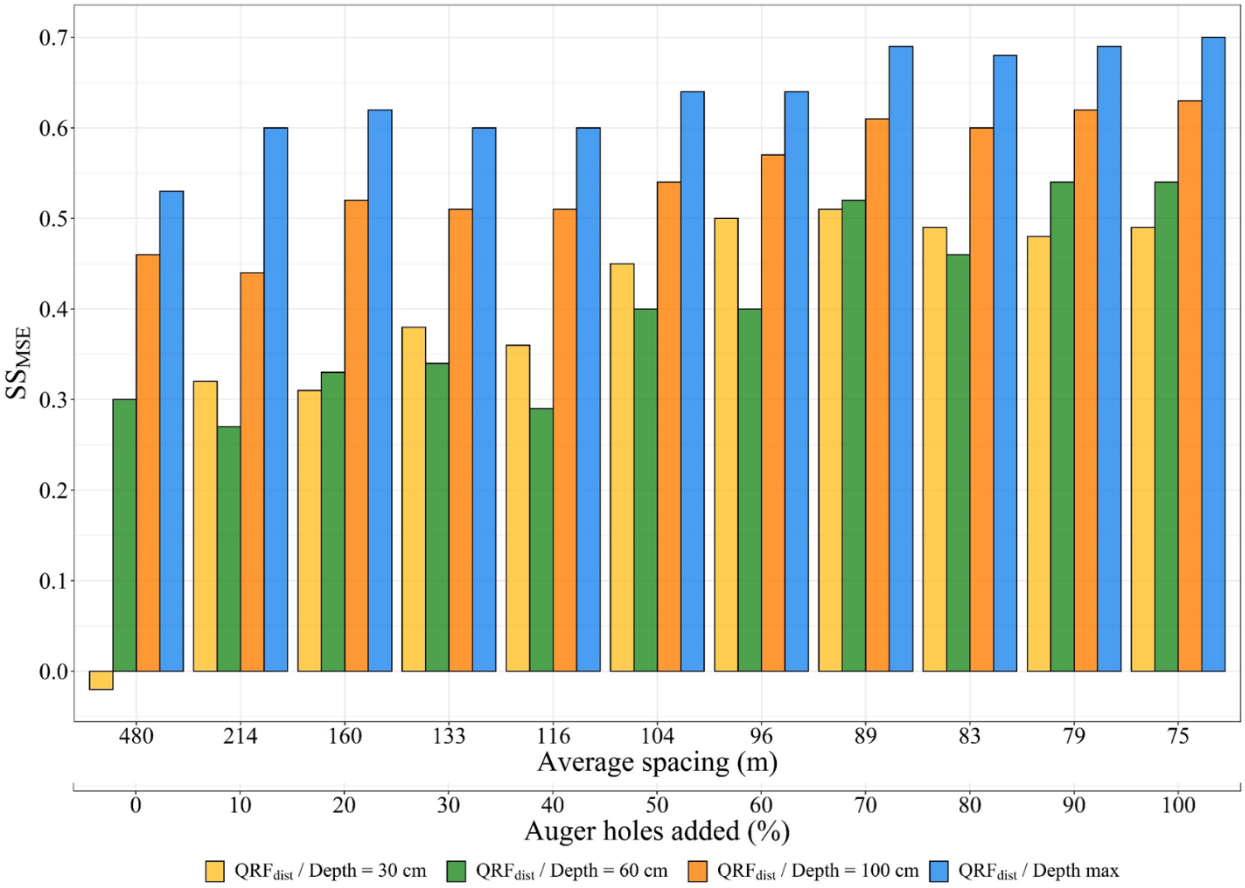


Figure 9. Evolutionary SS_{MSE} according to the number of auger hole observations added to the inputs for the four SAWCs

When considering the costs of adding new auger hole observations according to the two cost-efficiency indicators described in section 3.5., it appeared that the cost of gaining one unit of RMSE (the error cost, Err_{cost}) was important until the first addition of the auger hole and further linearly increased as new auger holes were added (Figure 10). This is translated by the relative cost-efficiency ratio (CE_r) by a dramatic decrease under the 1:1 ratio when adding the first auger hole observations and then a slow decrease for further additions.

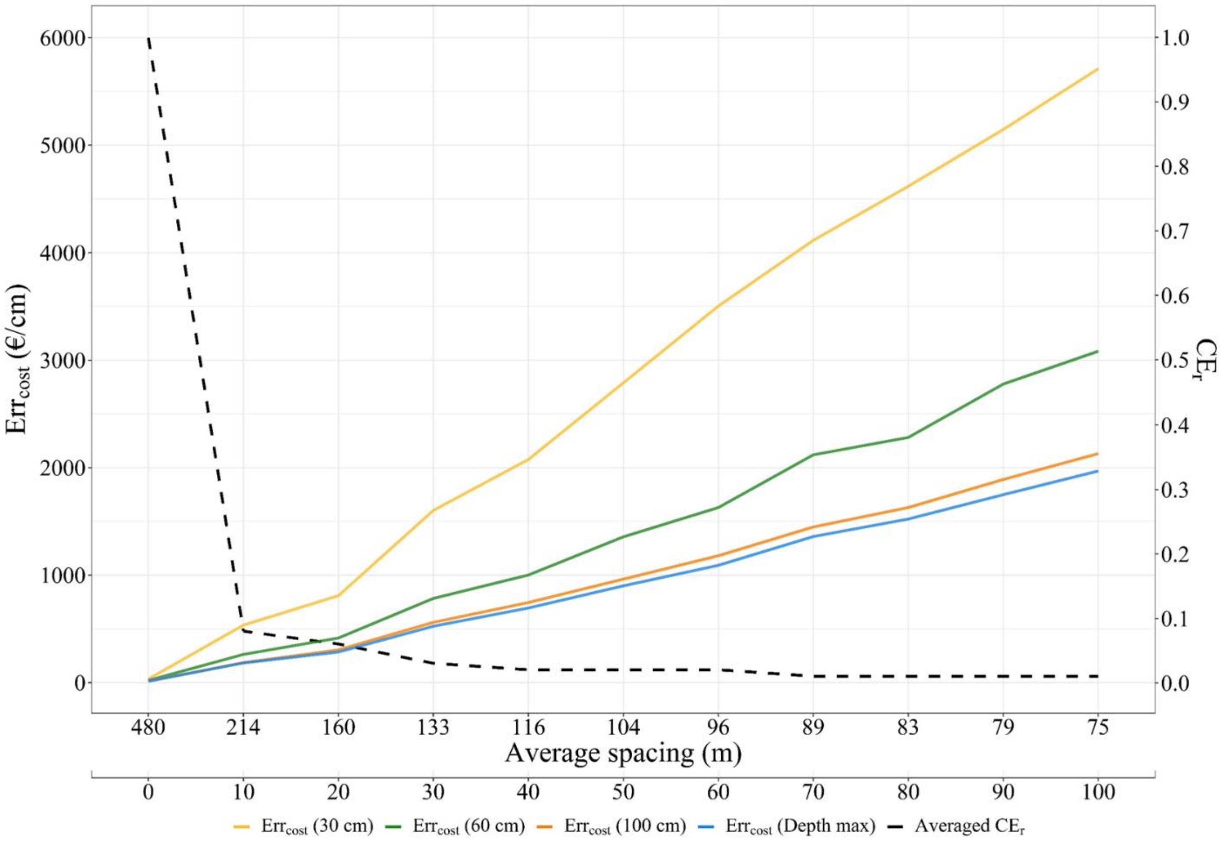


Figure 10. Cost-efficiency ratios according to the average spacing related to the number of auger holes

5. Discussion

5.1. Soil Available Water Capacity

The selected case study considered the soil available water capacity, which is among the most highly demanded properties of end users, as the targeted soil property (Richer de Forges et al, 2019). This paper completes the small set of papers that were devoted to the digital mapping of SAWC (Hong et al., 2013; Malone et al., 2009; Padarian et al., 2014; Poggio et al., 2010; Román Dobarco et al., 2019; Ugbaje and Reuter, 2013, Amirian-Chackan et al., 2019) and the even smaller set of papers that addressed all the SAWC components as defined by the original definition reported by Cousin et al. (2003) (Eq. 1) (Leenaars, 2018; Romàn Dobarco et al., 2019; Styc and Lagacherie, 2019, submitted).

However, as in many DSM applications, the SAWC was determined at local sites without the full measurements of its components. Visual estimations of the coarse fragment content and of the soil depth generated observational uncertainties and, for the latter, right-censored estimations due to the limitation in observation depths. Furthermore, the water retention capacity of each horizon was not fully measured, although it is worth noting that some components of this retention that are usually not measured (bulk density, field capacity) were measured here on the soil profiles. To overcome the measurement limitations, pedotransfer functions were used (see section 2.2.3). It is worth noting that these pedotransfer functions were highly case specific both regarding their input (textural classes + field capacity measurements) and their target (the b coefficient). The addition of all these peculiar uncertainties should result in a significant overall uncertainty of the soil inputs that is well reported by the nuggets of the variograms of the densest datasets (Figure 6, right panel). This uncertainty may greatly explain the limitation of performances that was observed, even for the densest datasets.

5.2. The interest of “spatial RFs”

Our results showed that the SAWC prediction performances were nearly systematically increased by adding some geographical information, i.e., the n of "scorpan" in McBratney et al.'s (2003) formula, to the set of candidate covariates used in a random forest. This confirmed the results obtained by Hengl et al. (2018) from various case studies. This, however, enriched these results by showing that the gains in performances provided by the addition of geographical covariates depend on the density of the sampling. Indeed, these gains were only effective when the dense sampling of auger hole observations was used (76 m spacing), whereas the low density of soil profiles did not provide clear improvements (Table 3). At high density levels, the classical landscape covariates were not sufficient to account for the variability shown in the dataset of soil inputs as represented by the variograms of Figure 6 (right panel), whereas the proximity effects brought by the geographical covariates allowed us to overcome this limitation.

In digital soil mapping, proximity effects have been traditionally addressed by using regression kriging (Hengl et al., 2004; Malone et al., 2009; Vaysse and Lagacherie, 2015). However, spatial QRF was demonstrated to have similar performances (Hengl et al, 2018) while having some decisive advantages in the context of our case study. Spatial QRF does not require any rigid statistical assumptions about the distribution and the stationarity of the target variable, which allows us to handle the bimodal distributions of SAWCs (Figure 5). It also does not require any geostatistical expertise for the manual fitting of variograms, which opens the possibility to fully automate the procedure so that non pedometrician, such as BRL staff, could use it for the other communes of the irrigation perimeter.

5.3. The interest of adding auger hole observations

The addition of dense spatial sets of auger hole observations in the modeling process significantly increased the level of performance when considering the best model (QRF_{dist}), which is in accordance with several previous experiments studying the impact of soil sampling densities (Somarathna et al. 2017, Wadoux et al. 2019 and Lagacherie et al., 2020). The performances observed in this case study were better than those in most of the published DSM applications dealing with SAWC (Ugbaje and Reunter, 2013; Styc and Lagacherie, 2019, submitted), which was the result of a much greater spatial density of the soil inputs (from 6/km^2 to 26/km^2) than in these previous applications (from 0.01/km^2 to 0.05/km^2)).

However, strong limitations in the SAWC prediction performances were still observed, even when using the most dense set of auger hole observations. These limitations increased as the maximum depth at which the SAWC was calculated decreased (Table 3). This means that significant proportions of the SAWC variabilities were not mapped despite the large densities of the auger hole observations used as input. To explain this fact, it is first interesting to note that for both the soil profiles and the soil profiles plus auger hole inputs, the performances and the spatially structured variance ratios of the input soil datasets were ranked similarly across SAWCs and spatial densities (Figure 6), which was already observed in the same region for different soil properties and study extents by Vaysse and Lagacherie (2015). Concerning the results using solely the profiles, this revealed that a part of the short-range variability shown by the variograms built from auger holes (Figure 6, left panel) was not captured by the soil dataset because of a limitation in spacing. However, this limitation decreased as the considered depth of the SAWC calculation increased, which explained the observed increase in performance from SAWC30 to SAWCmax. Concerning the results using the auger hole observations, a similar trend was observed since the local uncertainty as revealed by the variogram nuggets (Figure 6, right panel) remained important due to observational uncertainty (see section 5.1.), which may induce noise that may perturb the calibration of the QRF model.

Finally, it should be recalled that these performances were calculated for predictions of the SAWC at precise locations, whereas SAWC is required for field or in-field management zones for most of the decision making. It could be expected that these performances would increase when the SAWC prediction will be spatially aggregated (Vaysse et al, 2017).

5.4. Uncertainty predictions

Since SAWC is a soil functional property composed of several primary soil properties, uncertainty predictions were provided by a specific error model previously proposed by (Román Dobarco et al., 2019) and further refined by Styc and Lagacherie (submitted). The uncertainty predictions were classically evaluated with regard to their unbiasness (PICP, Table 3). They were also evaluated for their ability to identify contrasted uncertainty areas (comparisons between residuals and predicted uncertainty, Table 4), which, to our knowledge, has never been done in the DSM literature before Styc and Lagacherie (submitted). The results were highly variable across models and spatial densities. However, the more accurate models tended to also provide the best pictures of the uncertainty patterns (Figure 6) with an overestimation of uncertainty (QRF_{dist} on Table 3). This overestimation was already observed by Lagacherie et al. (2020) and was assumed to be due to the inclusion of outliers as the average spacing decreased, which probably disturbs the limit estimations of the confidence interval. On the other hand, a part of the inaccuracy of uncertainty predictions may come from the differences (Figure 5) between the distributions of SAWC values calculated from auger holes (used as calibration data only) and from soil profiles (used as evaluation data). More attention must be paid in the future to uncertainty predictions in view of identifying the possible causes of these uncertainty mispredictions.

It is interesting to note that some of the produced uncertainty maps showed strong similarities with possible drivers (see comments of Figure 8), which can be interpreted from our common sense pedological knowledge. The largest uncertainties were estimated i) in chromic Luvisols (Figure 6a) because of the large rates of coarse fragment content that are known to be difficult to quantify in the field, ii) in areas of lower densities of soil observations (Figure 6c and 6e) because of difficulties of model calibration at these locations and iii) for the largest predicted values of SAWCs with the best models (Figure 6h) because the estimates of relative uncertainty reached an unsurmountable floor that is likely related to the observational uncertainty. All these observations reinforce the credibility of the presented uncertainty maps.

5.5. The level of performance obtained and cost.

The use of auger hole observations as complementary soil input to soil profiles led to a substantial increase in performance, but the harvesting process was very time consuming, which resulted in high costs (see section 4.5). Figure 9 curves show that the performance gains were obtained by increasing costs as the density of the auger holes increased. A compromise should then be found, which can be formulated as “the number of auger hole observations that reach an acceptable level of performance while keeping an acceptable cost level”. The cost indicator curves of Figure 10 did not reveal a clear compromise. However, such curves could be used with a prior definition of what performance and costs are acceptable. Furthermore, such cost curves could be improved if either more sophisticated sampling is used (e.g., van Groningen et al, 1998) or if the harvesting costs could be reduced by a partial automation of digitizing procedures (Yang and Yang, 2017).

Finally, it should be stressed that the quantitative evaluation of prediction performance that served as a basis for building the curve costs should be completed by a qualitative examination of the maps. As revealed by the spatial patterns of the predicted SAWC maps, considerable gains in spatial resolution were obtained by adding auger holes, which may enable field-level decision making. This may constitute a more decisive added value than the moderate gain in precision quantitatively evaluated by the cost indicators.

6. Conclusion

In this study, the main lessons were as follows:

- A QRF approach using euclidian buffer distances outperformed a classical QRF approach in predicting SAWC with a dense set of profiles and auger holes
- The addition of a dense spatial sampling of auger hole observations dramatically increased the performance in predicting SAWCs and increased the spatial resolutions of the SAWC pattern delineations, but there were limitations due to the uncertainty of the auger hole observations.
- The performances in predicting SAWC values varied following some drivers that were expected - average spacing of sites, and type of observations (profiles vs. auger holes) - and following other drivers that were revealed by the uncertainty maps – pedological context, local density of sites, SAWC predicted values – (see section 5.4.).
- The cost-efficiency analysis did not reveal a clear compromise in terms of limiting the costly harvesting of auger hole data. Rather, the compromise should be user specific and should be updated as soon as partial automation is possible (see section 5.5)

7. Acknowledgments

This research was granted by the French National Research Institute for Agriculture, food and environment (INRAE), the French Research and Technology Agency (ANRT) and BRL Exploitation. Philippe Lagacherie is a collaborator of the GLADSOILMAP Consortium supported by the LE STUDIUM Loire Valley Institute for Advanced Research Studies

8. References

Alfons, A. (2012) cvTools: Cross-Validation Tools for Regression Models. R Package. Available online: https://cran.rproject.org/web/packages/cvTools/cvTools.pdf. (accessed on 11 May 2012).

Amirian-Chakan, Alireza, Budiman Minasny, Ruhollah Taghizadeh-Mehrjardi, Rokhsar Akbarifazli, Zahra Darvishpasand, and Saheb Khordehbin. (2019). Some Practical Aspects of Predicting Texture Data in Digital Soil Mapping. *Soil and Tillage Research* 194:104289.

Arrouays, D., Lagacherie, P., & Hartemink, A. E. (2017). Digital soil mapping across the globe. *Geoderma Regional*. https://doi.org/10.1016/j.geodrs.2017.03.002

Arrouays, D., Leenaars, J. G. B., Richer-de-Forges, A. C., Adhikari, K., Ballabio, C., Greve, M., … Rodriguez, D. (2017). Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ*. https://doi.org/10.1016/j.grj.2017.06.001

Baize, D., & Jabiol, B. (1995). *Guide des sols, Quae*.

Böhner, J., McCloy, K.R., Strobl, J., 2006. SAGA — analysis and modelling applications Göttinger Geogr. Abh., 115, p. 130

Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

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. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Cousin, I.,Nicoullaud, B., Coutadeur, C. (2003). Influence of rock fragments on the water retention and water percolation in a calcareous soil. Catena, 53, pp. 97-114

Dominati, E., Mackay, A., Green, S., & Patterson, M. (2014). A soil change-based methodology for the quantification and valuation of ecosystem services from agroecosystems: A case study of pastoral agriculture in New Zealand. *Ecological Economics*. https://doi.org/10.1016/j.ecolecon.2014.02.008

Gomez, C., & Coulouma, G. (2018). Importance of the spatial extent for using soil properties estimated by laboratory VNIR/SWIR spectroscopy: Examples of the clay and calcium carbonate content. *Geoderma*. https://doi.org/10.1016/j.geoderma.2018.06.006

Hengl, T. (2019). GSIF : Global Soil Information Facilities. R package version 0.5-5. https://CRAN.R-project.org/package=GSIF.

Hengl, T., De Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., … Gonzalez, M. R. (2014). SoilGrids1km - Global soil information based on automated mapping. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0105992

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Shepherd, K., Sila, A., … Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions S1 Regression-kriging in R using the Meuse data set. *PlosOne*, *10*(6), e0125814. https://doi.org/https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5518

Heuvelink, G. B. M., Burrough, P. A., & Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, *3*(4), 303–322. https://doi.org/10.1080/02693798908941518

Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., & Lee, K. (2013). Predicting and mapping soil available water capacity in Korea. *PeerJ*, *1*, e71. https://doi.org/10.7717/peerj.71

Kish, L. (1965). Survey sampling. New York, London: John Wiley

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., & Nkuba-Kasanda, L. (n.d.). Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: a numerical experiment using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. *Geoderma*, 375, 114503

Leenaars, J. G. B., Claessens, L., Heuvelink, G. B. M., Hengl, T., Ruiperez González, M., van Bussel, L. G. J., … Cassman, K. G. (2018). Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. *Geoderma*. https://doi.org/10.1016/j.geoderma.2018.02.046

Legros, J. P. (1996). *Cartographies des sols: de l'analyse spatiale à la gestion des territoires*. In *Collection Gérer l'environnement*. Retrieved from https://books.google.fr/books?id=MiONzDc-jnQC

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1–2), 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*. https://doi.org/10.1016/S0016-7061(03)00223-4

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.

Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

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. https://doi.org/10.1016/j.scitotenv.2016.07.066

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., … Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, *4*(1), 1–22. https://doi.org/10.5194/soil-4-1-2018

Padarian, J., Minasny, B., McBratney, A. B., & Dalgliesh, N. (2014). Predicting and mapping the soil available water capacity of Australian wheatbelt. *Geoderma Regional*, *2–3*(C), 110–118. https://doi.org/10.1016/j.geodrs.2014.09.005

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, *30*(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Richer-de-forges, A.C., Arrouays, D., Bardy, M., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Sauter, J., Voltz, M., (2019). Mapping of Soils and Land-Related Environmental Attributes in France : Analysis of End-Users ' Needs. Sustainability 11, 1–15.

Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N. P. A., Cousin, I., & Martin, M. P. (2019). Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. *Geoderma*, *344*(February), 14–30. https://doi.org/10.1016/j.geoderma.2019.02.036

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. https://doi.org/10.1126/science.1175084

Somarathna, P. D. S. N., Minasny, B., & Malone, B. P. (2017). More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. *Soil Science*

Society of America Journal, *81*(6), 1413–1426. https://doi.org/10.2136/sssaj2016.11.0376

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

Spiess, A-N., (2018). Propagate: Propagation of Uncertainty. R package version 1.0-6. Available on : https://CRAN.R-project.org/package-propagate

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Styc, Q., & Lagacherie, P. Uncertainty assessment of soil available water capacity using error propagation : a test in Languedoc Roussillon. Submitted on Geoderma since 24/01/2020.

Styc, Q., & Lagacherie, P. (2019). What is the Best Inference Trajectory for Mapping Soil Functions : An Example of Mapping Soil Available Water Capacity over Languedoc Roussillon (France). *Soil Systems*, *3*(34), 17. https://doi.org/10.3390/soilsystems3020034

Ugbaje, S. U., & Reuter, H. I. (2013). Functional Digital Soil Mapping for the Prediction of Available Water Capacity in Nigeria using Legacy Data. *Vadose Zone Journal*, *12*(4), 0. https://doi.org/10.2136/vzj2013.07.0140

Van Groenigen, J. W., Stein, A., (1998). Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27 (5), 1078–1086.

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. https://doi.org/10.1016/j.geodrs.2014.11.003

Vaysse, Kévin, & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017

Veihmeyer, F.J., Hendrickson, A.H. 1927. The relation of soil moisture to cultivation and plant growth. Soil Sci. 3, 498–513.

Viscarra Rossel, R. A., & Brus, D. J. (2018). The cost-efficiency and reliability of two methods for soil organic C accounting. *Land Degradation and Development*, *29*(3), 506–520. https://doi.org/10.1002/ldr.2887

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

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

Walvoort, D. J. J., Brus, D. J., & de Gruijter, J. J. (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*, *36*(10), 1261–1267. https://doi.org/10.1016/j.cageo.2010.04.005

Walvoort, D. J. J., Brus, D. J., & de Gruijter, J. J. (2018). Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata. R package version 0.3-8. https://CRAN.R-project.org/package=spcosa

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. https://doi.org/10.18637/jss.v077.i01

Yang, C. S., & Yang, Y. H. (2017). Improved local binary pattern for real scene optical character recognition. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2017.08.005