High resolution soil property maps over mainland France : application to soil organic carbon and its additional sequestration and storage potentials

Songchao Chen

UNIVERSITE ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

AGRO CAMPUS OUEST

INRA SCIENCE & IMPACT

THESE DE DOCTORAT DE

AGROCAMPUS OUEST
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : « *Sciences de l'Environnement* »

Par
Songchao Chen

Cartographie à haute résolution de propriétés du sol à l'échelle de la France métropolitaine: application au carbone organique du sol et à ses potentiels additionnels de séquestration et de stockage

Thèse présentée et soutenue à Rennes, le 09 décembre 2019
Unité de recherche : UMR Sols, Agrohydrosystèmes, Spatialisation INRA-Agrocampus Ouest et US InfoSol INRA Orléans
Thèse N° : 2019-22

Rapporteurs avant soutenance :

Mogens Humlekrog Greve	Aahrus University, Rapporteur
Cécile Gomez	Institut de Recherche pour le Développement, Rapporteure

Composition du Jury :

Vincent Chaplot	Institut de Recherche pour le Développement, Examinateur
Christophe Cudennec	Agrocampus Ouest, Examinateur
Laurence Hubert-Moy	Université de Rennes 2, Examinatrice
Jacqueline Hannam	Cranfield University, Examinatrice
Christian Walter	Agrocampus Ouest, Directeur de thèse
Dominique Arrouays	Inra, Co-directeur de thèse

High resolution soil property maps over mainland France: application to soil organic carbon and its additional sequestration and storage potentials

Cartographie à haute résolution de propriétés du sol à l'échelle de la France métropolitaine : application au carbone organique du sol et à ses potentiels additionnels de séquestration et de stockage

Songchao Chen

Thesis

submitted in fulfilment of the requirement for the degree of doctor

at Agrocampus Ouest

in presence of the

Thesis Jury Committee

to be defended in public

on Monday 09 December 2019

at Agrocampus Ouest, Rennes, France

Index

Chapter 1

General introduction

Soils are the biologically active and porous mediums that have developed in the uppermost layer of the Earth's crust. Soils consist of minerals, soil organic matter, organisms, gases and liquids that together support the life on Earth. The quality of the soil usually determines the nature of plant ecosystems and the capacity of land to support the life of animals and society (Weil and Brady, 2016). In the 21st century, soils also play a central role in many global issues. From food security, water pollution and climate change to sustainable energy, human health and biodiversity loss, the world's ecosystems are impacted by various biogeochemical processes carried out in soils (Koch et al., 2012; Weil and Brady, 2016). These global issues arise the soil security concept, which refers to the maintenance and improvement of the world's soil resources to deal with these challenges (Koch et al, 2013; McBratney et al., 2014). Soils are also central for reaching the United Nations Sustainable Development Goals (SDGs) (Table 1.1), such as goals 2 (Zero Hunger), 3 (Good Health and Well-being), 6 (Clean Water and Sanitation), 7 (Affordable and Clean Energy), 12 (Responsible Consumption and Production), 13 (Climate Action), 14 (Life below Water), and 15 (Life on Land) (Bouma et al., 2014; Keesstra et al., 2016).

Although soils are central to global issues, their management requires local actions and knowledge. Therefore, there is an emerging demand for soil information both at global and local scales, which is the main reason of the development of a fine-resolution global grid of soil properties (Sanchez et al., 2009; Arrouays et al., 2014a). Globally, around two thirds of the countries have conventional soil maps at a 1:1 million scale or finer, but more than two thirds of the total land area have not been mapped even at a 1:1 million scale (Hartemink et al., 2013). Conventional soil maps are often produced by obsolete data and their production is laborious, time-consuming, and expensive (Grunwald et al., 2011). Besides, they often do not provide uncertainty, and heavily rely on expert knowledge which makes them hard to be reproduced and updated. These limitations of conventional soil maps motivated the rise and development of a sub discipline of soil science, digital soil mapping (DSM), following the advancement of geo-information technology and computation power (Minasny and McBratney, 2016; Zhang et al., 2017).

DSM has been defined as: *the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observations and knowledge and from related environmental*

variables (Lagacherie and McBratney, 2006). This concept developed from the theory of

Table 1.1 The 17 Sustainable Development Goals

ID	Goal	Detailed topic
1	No Poverty	End poverty in all its forms everywhere
2	Zero Hunger	End hunger, achieve food security and improved nutrition and promote sustainable agriculture
3	Good Health and Well-being	Ensure healthy lives and promote well-being for all at all ages
4	Quality Education	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
5	Gender Equality	Achieve gender equality and empower all women and girls
6	Clean Water and Sanitation	Ensure availability and sustainable management of water and sanitation for all
7	Affordable and Clean Energy	Ensure access to affordable, reliable, sustainable and modern energy for all
8	Decent Work and Economic Growth	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
9	Industry, Innovation and Infrastructure	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
10	Reduced Inequality	Reduce inequality within and among countries
11	Sustainable Cities and Communities	Make cities and human settlements inclusive, safe, resilient and sustainable
12	Responsible Consumption and Production	Ensure sustainable consumption and production patterns
13	Climate Action	Take urgent action to combat climate change and its impacts
14	Life Below Water	Conserve and sustainably use the oceans, seas and marine resources for sustainable development
15	Life on Land	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
16	Peace and Justice Strong Institutions	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
17	Partnerships to achieve the Goal	Strengthen the means of implementation and revitalize the global partnership for sustainable development

soil forming factors by Dokuchaev (1883) and Jenny (1941) (*clorpt* model), and later elaborated on the *scorpan* model, $S=f(s, c, o, r, p, a, n)$ proposed by McBratney et al. (2003). It is worthy to note that many attempts of DSM took place before 2003, but

scorpan model was the first conceptualization of DSM for quantitative spatial prediction. In this model, soil attributes or soil classes can be predicted by their relationships with seven factors, including other soil information (*s*), climate (*c*), organisms (*o*), relief (*r*), parent material (*p*), age (*a*) and position (*n*). This empirical model and both soil class or soil property and seven factors are spatially and temporally explicit.

Table 1.2 Twelve soil properties recommended in *GlobalSoilMap*

Soil property	Unit
Total profile depth	cm
Plant exploitable (effective) soil doeth	cm
Organic carbon	g kg^{-1}
pH	*10
Sand	g kg^{-1}
Silt	g kg^{-1}
Clay	g kg^{-1}
Gravel	m^3 m^{-3}
ECEC	$cmol_c$ kg^{-1}
Bulk density of the fine earth (< 2 mm) fraction	Mg m^{-3}
Bulk density of the whole soil in situ	Mg m^{-3}
Available water capacity	mm

From 2003 to the early 2010s, DSM remained an academic and research activity. Then DSM became more operational in delivering soil information to both scientific community and decision and policy makers (Minasny and McBratney, 2016, Arrouays et al., 2017). One of the examples to make DSM becoming operational is the *GlobalSoilMap* initiative (Sanchez et al., 2009; Arrouays et al., 2014a). This initiative aims at delivering 12 major soil properties (Table 1.2) under several specifications (90 m or 3 arc-second resolution, uncertainty quantification, six fixed depth intervals) all over the world using a bottom-up approach (from country to globe). During the same period, top-down approaches were developed (from global to country). SoilGrids is one of the best examples using top-down approach. The SoilGrids1km product (10 soil properties and 2 soil classes) was produced in 2014 using 110,000 soil profiles all over the world and 75 global environmental covariates (Hengl et al., 2014). In 2017, the SoilGrids250m product was updated using 150,000 soil profiles and 280 environmental covariates. These products are free available online (https://soilgrids.org, v0.5.3). The present version of SoilGrids250m did not provide uncertainty estimates. Another big event happened in 2017: the Global Soil Partnership (host by UN-FAO) produced a Global Soil Organic

Carbon map (GSOCmap, http://54.229.242.119/GSOCmap/), which is a signal that DSM is now recognized by policy makers at the highest level. For GSOCmap production, about two third of the countries provided bottom-up products which shows that DSM is now operational at country level thanks to capacity building, and the rest of the world was covered by top-down approaches (mainly from LUCAS for some E.U. countries and SoilGrids for the rest of the world). Note that the current GSOCmap had a resolution of 1 km, and was only focused on topsoil (0-30 cm) without uncertainty estimates.

Actually, countries having delivered almost complete *GlobalSoilMap* products are rather few (USA and Australia) whereas many attempts to map some of the twelve basic soil properties have been done in some countries (Brazil, Chile, China, Denmark, France, Hungary, Nigeria, Scotland and South Korea) and a very large number of countries (e.g., Croatia, Estonia, Kenya, Madagascar and Sri Lanka) have produced national maps of some soil properties, although most of them did not follow the *GlobalSoilMap* specifications.

This thesis starts with a review of the state-of-the-art in **Chapter 2 Digital mapping of soil information at a broad-scale: A review**. This review included 160 articles relevant to broad-scale DSM published between 2003 and middle of 2019, and I identified some key issues and main challenges for the broad-scale DSM studies. Most of the DSM works were concentrated on SOC and soil particle size fractions (clay, silt and sand). Among these studies on SOC, most of them concentrated on SOC in topsoil and rather few studies addressed the SOC in deep soil and the potential of soil to store additional SOC.

The content from **Chapter 3** to **Chapter 7** mainly focused on SOC due to two reasons: (1) soil C pool is largest terrestrial C pool, which is more than the sum of C stored in the vegetation and atmosphere, making it crucial in global C cycle; (2) SOC plays a crucial role in ecosystem services, including food production, water regulation, erosion control, biodiversity and climate regulation (Sanchez et al., 2009; Adhikari and Hartemink, 2016; Rumpel et al., 2018). Being the agency for SGD indicator 15.3.1, United Nations Convention to Combat Desertification has recognized SOC stock map as an indicator to detecting and monitoring land degradation (IUCN, 2015). In addition, at the COP21, the initiative "4 per 1000 carbon sequestration in soils for food security and the climate" (4 per 1000, http://4p1000.org/understand) was launched with an expectation to increase

global SOC stocks by 0.4% y^{-1} as a compensation for global GHG emissions as well as to combat soil degradation, increases food security and enhances agriculture adaptation to climate change (Minasny et al., 2017; Soussana et al., 2015, 2019). Due to the significant importance of SOC, there is growing interest to spatially estimate SOC and the potential of soils to sequester additional SOC at fine resolution over broad-scales, and this is one of the main aims of my thesis. The internal links from **Chapter 3** to **Chapter 7** are shown in the Figure 1.1: SOC content (**Chapter 3**), bulk density (**Chapter 4**) and soil depth (**Chapter 5**) are all necessary inputs for calculating SOC stocks, which can be further used to determine SOC sequestration potential (**Chapter 6**) and SOC storage potential (**Chapter 7**) using DSM and statistical models. The general objectives of this thesis are (1) improving the prediction and reducing the uncertainty of soil properties highly relevant to the calculation of SOC stocks, which includes improving national SOC content maps using model averaging, predicting soil bulk density using pedotransfer functions and their validity domain, and dealing with right censored data in probability mapping of soil thickness using random survival forest; (2) moving from DSM to Digital Soil Assessment (DSA) by mapping soil functional properties such as SOC sequestration potential and SOC storage potential.

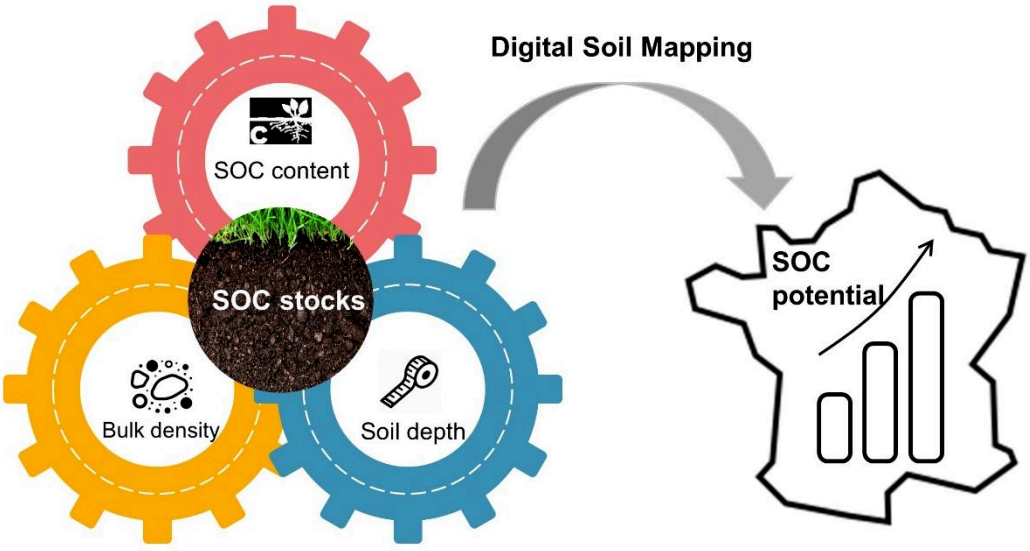


Figure 1.1 Internal links from Chapter 3 to Chapter 7

As a result of DSM development at multiple scales, there are often multiple SOC maps available in a given area and produced using various soil databases, environmental

covariates, and DSM methods. Users may have multiple maps of SOC with different predictions and different map accuracy which may lead to confusion regarding which map should be used or whether the maps could or should be combined. Conversely, some countries do not have enough point data to produce a country-based map using bottom-up approach. I dealt with these issues and proposed possible solutions in **Chapter 3 Model averaging for mapping topsoil organic carbon in France**.

Soil bulk density (BD) is one of the necessary parameters for weight-to-volume conversion for estimating SOC stocks. However, it is usually lacking in soil database worldwide mainly owing to the fact that determination of BD is usually time consuming and labor intensive. Therefore, pedotransfer functions (PTFs) are often used to derive missing BD before mapping SOC stocks using DSM. However, these studies either used very general PTFs (e.g., linear model, logarithmic model or exponential model) or did not check validity domain of the applied PTFs. Therefore, in **Chapter 4 Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over larger area**, I built machine learning based PTFs for BD, and also determined the validity domain for the PTFs to avoid invalid extrapolation.

Large percentage of current SOC maps focus on topsoil. However, SOC stocks in deep layers (>30 cm) are estimated to represent 53% of the SOC stocks in the upper 100 cm (about 1,500 Pg) and 71% of the SOC stocks in the upper 200 cm (about 2,400 Pg) (Batjes, 1996). Due to the poor understanding of deep SOC in soil, more and more recent studies suggested to investigate SOC deeper into soils as (1) carbon-climate feedback is sensitive to deep soil C decomposability (Koven et al., 2015); (2) deep soil may have more potential to sequester SOC (Lal, 2018). For a better understanding of spatial distribution of SOC stocks in deep soil, estimates of soil thickness (ST) are of crucial importance. However, point data on ST are very often censored (i.e. the observed ST is lower than actual ST) which makes DSM predictions more difficult than for properties having continuous measurements over their complete feature space. These challenges are addressed in the **Chapter 5 Probability mapping of soil thickness by random survival forest at a national scale**.

The 4 per 1000 initiative proposed the aspirational target to increase SOC at a rate of 0.4% y^{-1} in the first 30-40 cm of soil. Indeed, it is generally accepted that there is an upper limit of soil stable C storage, which is referred to as SOC saturation (Hassink, 1997; Six et

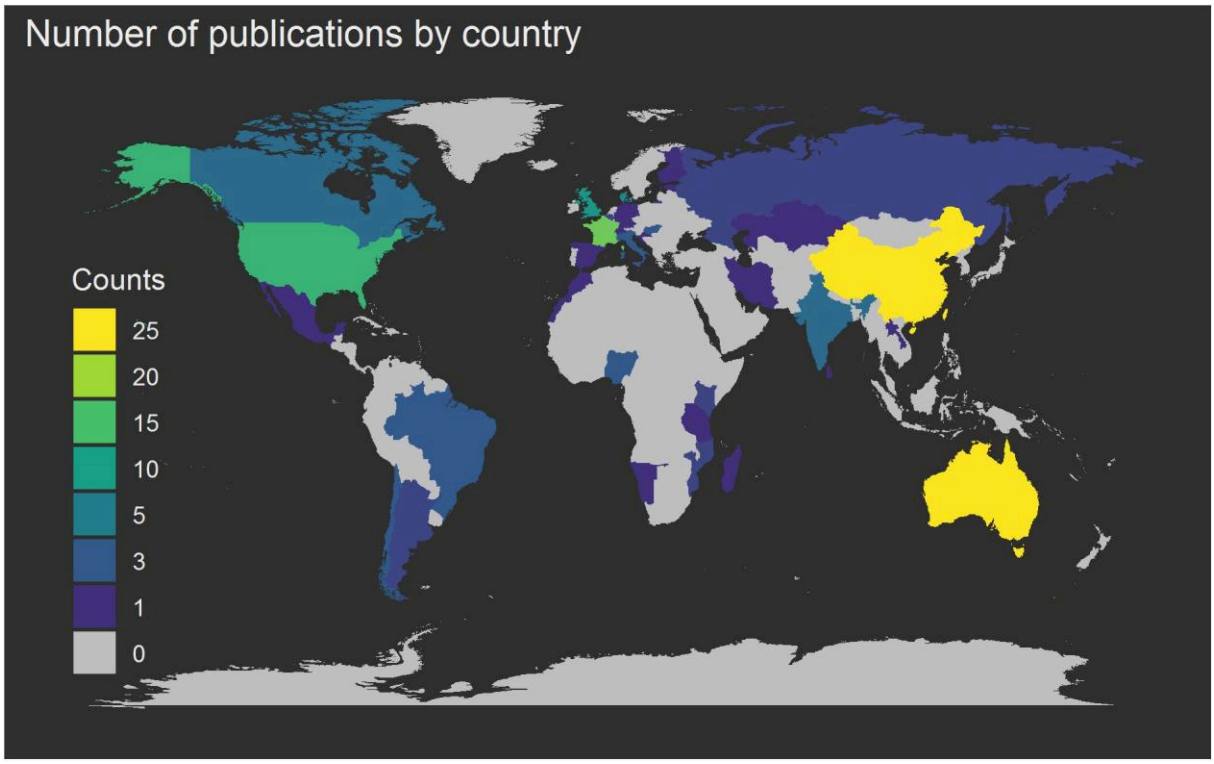
al., 2002; Angers et al., 2011; Chen et al., 2019b). Then SOC sequestration potential can be calculated by the difference between SOC saturation and current stable SOC content, which can be used to assess whether 4 per 1000 target can be theoretically achieved. A map of SOC sequestration potential will also help decision makers to put more efforts on the areas having higher potentials. In the **Chapter 6 Fine resolution map of top- and subsoil carbon sequestration potential in France**, I demonstrated how to map SOC sequestration potential using an empirical equation (based on fine fraction content) proposed by Hassink (1997) and DSM for both topsoil and subsoil.

In the context of the 4 per 1000 initiative, the target of increasing SOC stocks relates to the total (whole-soil) SOC stocks. Therefore, determining SOC storage potential using the maximum SOC associated with the fine fraction (SOC saturation) is not appropriate because the SOC stored in the coarse fraction can represent a significant percentage of the total SOC stocks. Another limitation about using the SOC saturation concept is that in many cases it might not be reached under given agro-pedo-climatic contexts. In the **Chapter 7 National estimation of soil organic carbon storage potential for arable soils: a data-driven approach coupled with carbon-landscape zones**, I present how to determine SOC storage potential in arable soil for both topsoil and subsoil using data-driven approach under different percentile setting and carbon landscape zones, and also evaluated the theoretical potential to meet 4 per 1000 target.

This thesis closes with the **Chapter 8 Conclusions and perspectives**, in which I summarized all aforementioned works and discussed the ways forward for DSM and *GlobalSoilMap* which I benefited a lot from the last Joint Workshop for Digital Soil Mapping and GlobalSoilMap organized in Santiago, Chile in March 2019 (Arrouays et al., 2019, submitted).

Chapter 2

Digital mapping of soil information at a broad-scale: A review



2.1 Introduction

In 21st century, the world is facing a number of challenges, such as population explosion, food security, environmental degradation, water scarcity, threatened biodiversity, climate change, and sustainable development (FAO, 2011). These challenges are more or less related to soil functions that are relevant to food production, climate regulation and adaption, carbon sequestration and water purification (McBratney et al., 2014; Adhikari and Hartemink, 2016). Moreover, soils are directly linked to some of the United Nation Sustainable Development Goals (i.e., goals 2, 3, 6, 7, 12-15) (Bouma et al., 2014; Keesstra et al., 2016). To address these global and regional issues, the demand for relevant and up-to-date soil information is increasing (Sanchez et al., 2009). The conventional way to produce soil maps (in polygon format) by soil survey are laborious, time-consuming, expensive and heavily based on experts' knowledge (Zhang et al., 2017). Moreover, these soil polygon maps are usually out-dated and their spatial resolution are rather low to support decision making in land management (Sanchez et al., 2009). In the last two decades, the concept of digital soil mapping (DSM) has been melded under the integration of soil survey data, Geographic Position System (GIS), Geostatistics, terrain analysis, machine learning, remote sensing, and high computing system (Arrouays et al., 2017).

As summarized by Lagacherie and McBratney (2006), DSM is the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge from related environmental variables. This concept was rooted from Jenny's five soil forming factor (climate, organisms, relief, parent materials, and time), and later developed into scropan-SSPFe (soil spatial prediction function with spatially auto-correlated errors) framework (Eq. (2.1)) by McBratney et al. (2003) for quantitative spatial prediction.

$$S_a[x,y,\sim t] \text{ or } S_c[x,y,\sim t]= f(s[x,y,\sim t],c[x,y,\sim t],o[x,y,\sim t],r[x,y,\sim t],p[x,y,\sim t],a[x,y,\sim t],n)+e \quad (2.1)$$

where S_a and S_c represent soil attributes and soil classes. The s refers to soil information, c refers to climate factor, o refers to organisms, vegetation or fauna or human activity, r refers to relief, p refers to parent material, a refers to age and time factor, n refers to spatial

or geographic position, and *e* is spatially correlated residuals.

Digital soil mapping has grown fast since the 1st Global Workshop on Digital Soil Mapping organized in Montpellier, France in 2004. Later on, the Global Workshop on Digital Soil Mapping was organized biannually from 2006 to 2016 (Table 2.1). To deal with soil related global issues motioned above, the idea of delivering a global grid of soil functional properties emerged at 2nd Global Workshop on Digital Soil Mapping held in Rio de Janeiro, Brazil in 2006. This culminated with the establishment and development of *GlobalSoilMap* Project, and then the 1st *GlobalSoilMap* was held in Orléans, France in 2013 (Sanchez et al., 2009; Arrouays et al., 2014a). Digital soil mapping has switched from a heavy academic focus to an operational purpose for delivering soil information to the scientific community, and decision and policy makers through the *GlobalSoilMap* project and the pillar 4 of the Global Soil Partnership initiative (Arrouays et al., 2017). Therefore, a Joint Workshop for Digital Soil Mapping and *GlobalSoilMap* was organized in Santiago, Chile in 2019 to benefit both IUSS Working Groups.

Table 2.1 Global workshops on digital soil mapping and *GlobalSoilMap*

No.	Year	Location	Event
1	2004	Montpellier, France	1st Global Workshop on Digital Soil Mapping
2	2006	Rio de Janeiro, Brazil	2nd Global Workshop on Digital Soil Mapping
3	2008	Logan, USA	3rd Global Workshop on Digital Soil Mapping
4	2010	Rome, Italy	4th Global Workshop on Digital Soil Mapping
5	2012	Sydney, Australia	5th Global Workshop on Digital Soil Mapping
6	2013	Orléans, France	1st *GlobalSoilMap* Conference
7	2014	Nanjing, China	6th Global Workshop on Digital Soil Mapping
8	2016	Aarhus, Denmark	7th Global Workshop on Digital Soil Mapping
9	2017	Moscow, Russia	2nd *GlobalSoilMap* Conference
10	2019	Santiago, Chile	Joint Workshop for Digital Soil Mapping and *GlobalSoilMap*

There were several reviews on DSM in the last decade. Grunwald (2009) characterized some recent progress on digital soil mapping and modelling. Grunwald et al. (2011) summarized some work on digital soil mapping and modelling at continental scale. Minasny et al. (2011) reviewed and discussed the recent advances in digital mapping

of soil. Minasny and McBratney (2016) illustrated a brief history and some lessons on digital soil mapping. Zhang et al. (2017) reviewed recent progress and future prospect of digital soil mapping. Different from previous reviews, the objective of this review was to summarize the recent progress, challenges and perspectives in broad-scale DSM studies with a spatial extent greater than 10,000 km².

Table 2.2 List of variables extracted in literature review

No	Variable	No	Variable	No	Variable
1	Year of publication	2	Journal	3	Open access[a]
4	Scale[b]	5	Continent	6	Country
7	Spatial extent (km²)	8	Soil sampling year	9	No of soil samples
10	Soil sampling density	11	Soil sampling strategy[c]	12	Validation strategy[d]
13	How to split calibration and validation sets	14	No of samples for calibration	15	No of samples for validation
16	Soil sampling elementary volume	17	Spatial resolution of produced map	18	Spatial predictive model
19	Soil (*scorpan*)	20	Climate (*scorpan*)	21	Organisms (*scorpan*)
22	Relief (*scorpan*)	23	Parent material (*scorpan*)	24	Age (*scorpan*)
25	Spatial position (*scorpan*)	26	Indirect satellite or airborne data	27	No of total covariates
28	Target soil property	29	Maximum soil depth	30	Depth interval of interest
31	Depth standardization	32	R^2 (Indicator)	33	R^2_{adj} (Indicator)
34	RMSE (Indicator)	35	ME (Indicator)	36	MAE (Indicator)
37	CCC (Indicator)	38	RPD (Indicator)	39	PICP (Indicator)
40	Accuracy (Indicator)	41	Kappa (Indicator)	42	AUC (Indicator)
43	Uncertainty estimate	44	Performance decrease with depth	45	*GlobalSoilMap* like product

[a] Yes or No; [b] regional, national, continental or global scale; [c] probability sampling, purposive sampling, mixture or NA; [d] internal validation, cross-validation, external validation, internal and cross-validation or NA;

2.2 Materials and methods

To assess the current progress in broad-scale digital mapping of soil information, we undertook a literature search related to DSM published after 2003, on which the *scorpan* concept was proposed. On 10th May, 2019, Web of Science was queried using several expressions applied to the topic of the articles. The search expressions were listed below:

"digital soil mapping" OR "globalsoilmap" OR "soilgrids" OR "soil-landscape modelling" OR "soil predictive modelling".

We kept all the relevant articles that were published in English recorded in Web of Science. Besides, as this literature review focus on broad-scale DSM, we only kept the articles that had a spatial extent larger than 10,000 km^2.

After manually filtering all relevant articles, a list of variables was extracted in order to derive systematic plots for the results section. Table 2.2 shows a total of 45 variables that were recorded for this review.

2.3 Results

In total, 160 articles were found to meet our requirements in Web of Science after manually selection, and their relevant information (45 variables) were extracted. The detailed information of these variables used in this review is shared online (https://drive.google.com/file/d/1KFnRjlwzkNyJ3APkYXq-mQRo6frtZ8YI/view?usp=sharing). Hereafter we show the most important results, which we present in 12 figures.

Figure 2.1 shows the annual number of articles that are relevant to broad-scale DSM from 2003 to 2019. Only a few articles addressed broad-scale DSM before 2010 (less than four per year), and a great increase in the number of publications was observed after 2010, with highest number of publications (28) in 2016 and 2017. In 2018, the number of publications decreased slightly to 24. As we only accounted the articles published before 10th May 2019, only 19 articles were observed in the results. However, the year of 2019 has high possibility to have more published articles than the year of 2018.

Figure 2.2 presents the geographic distribution of 145 articles specified by country, from which the continental or global studies were excluded (four in Africa, one in North and South America, five in Europe and five for the globe). It showed that DSM has been used in delivering soil information all over the world. Among these countries, China and Australia were the most active with the largest publications (25). France, United States, United Kingdom and Denmark ranked from third to sixth, with 17, 13, 9 and 7 publications, respectively.

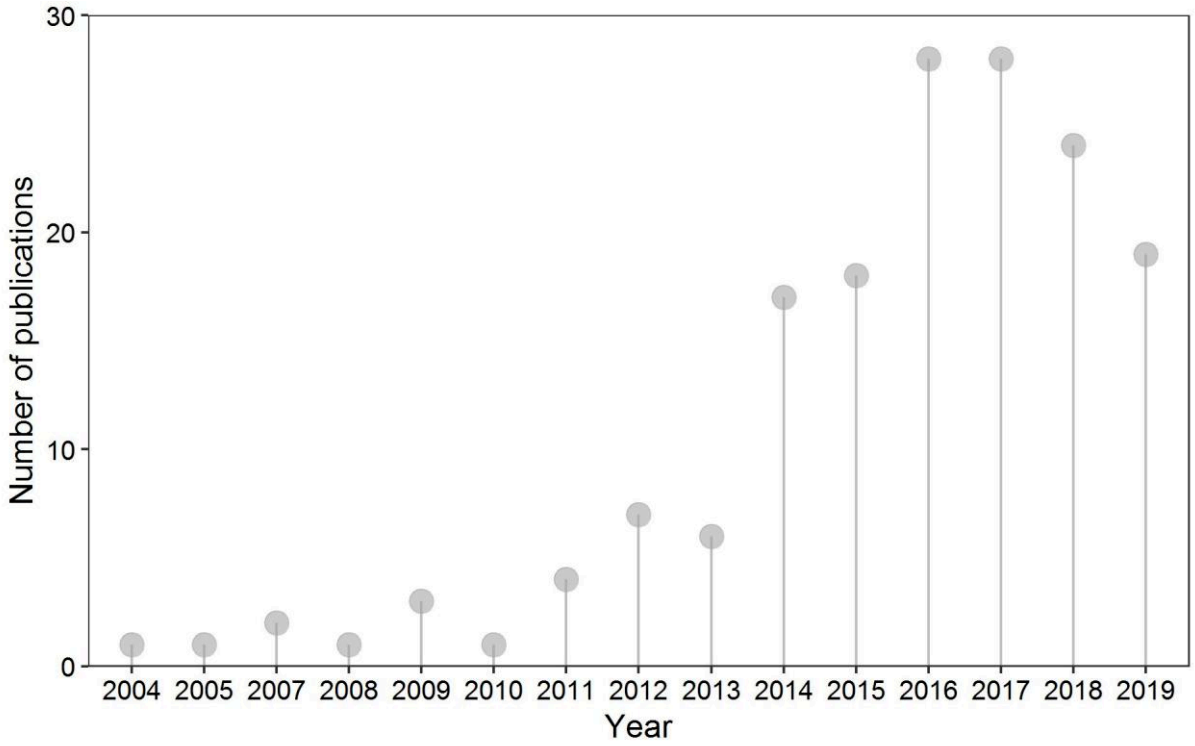


Figure 2.1 Number of publications by year

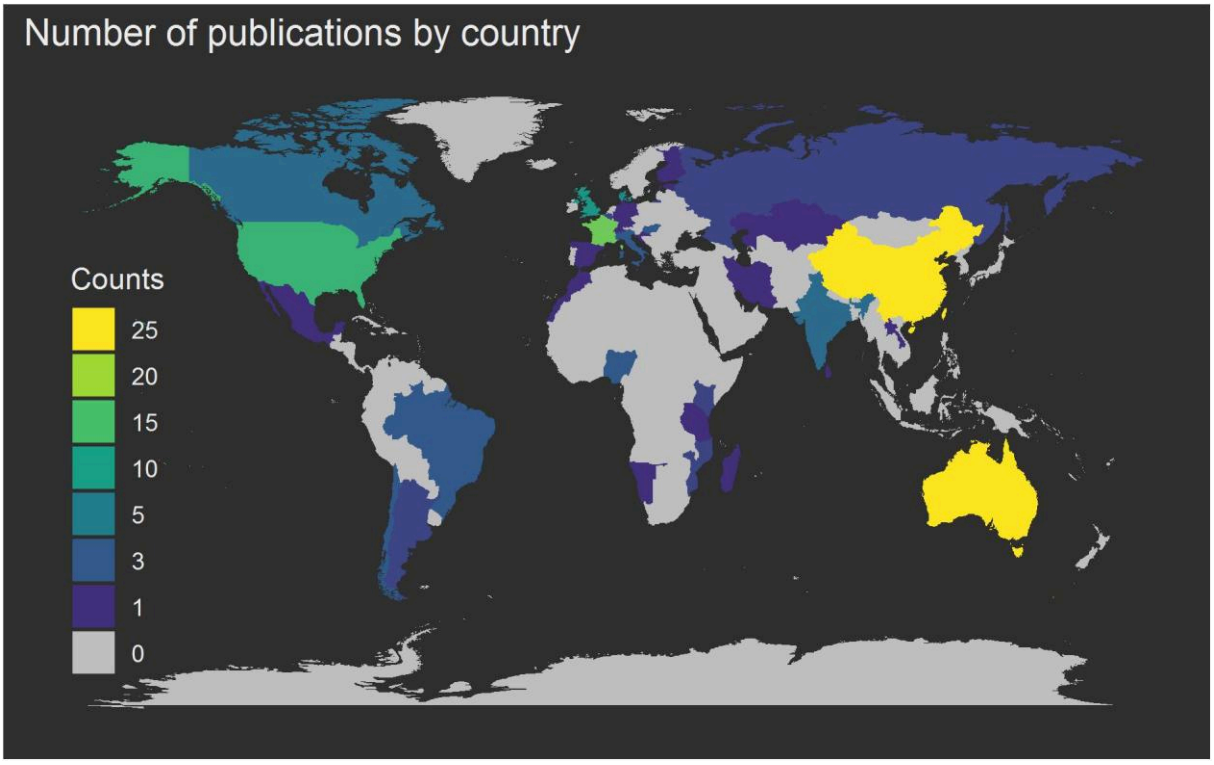


Figure 2.2 Map of publications by country

Figure 2.3 shows the frequency of journals relevant to this review. A total of 46 journals were involved, and the journals with only one count were classified as "Others". It showed that Geoderma was the most frequent journal (53) that was preferred by authors to publish their researches on broad-scale DSM. Others included 27 journals that occurred one time. Science of the Total Environment and Geoderma Regional rank the second place both having 11 publications, and they were followed by Soil Research with 8 publications. Catena, Ecological Indicators, European Journal of Soil Science and PLOS One had 6 publications on broad-scale digital soil mapping. It also showed that 28 out of 160 articles were open access, in which the majority were published after 2014 (not shown in the Figure 2.3).

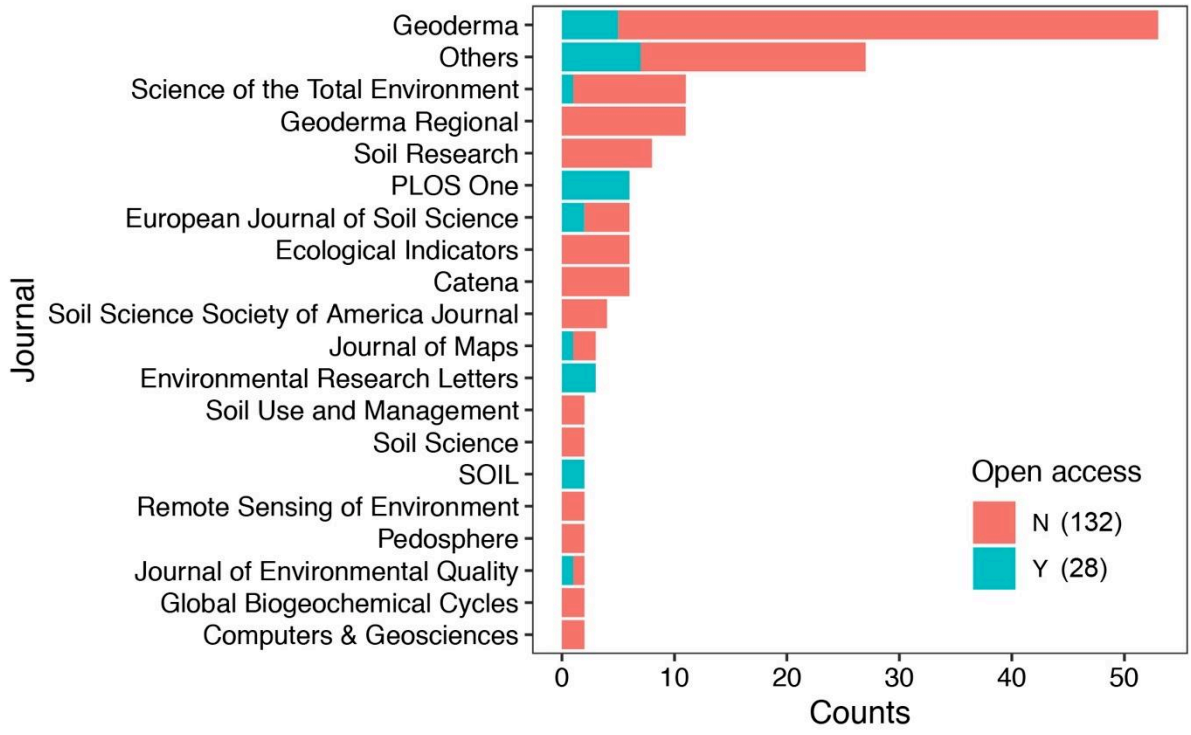


Figure 2.3 Number of publications by journal. The category of "Others" indicates the sum of the journals that only occurs one time.

The trends between spatial extent and soil sampling density is presented in Figure 2.4. The sampling density varied from 1 to 0.0001 sample per km^2. It also showed that regional or national scale studies usually had higher sampling density than continental and global scale studies.

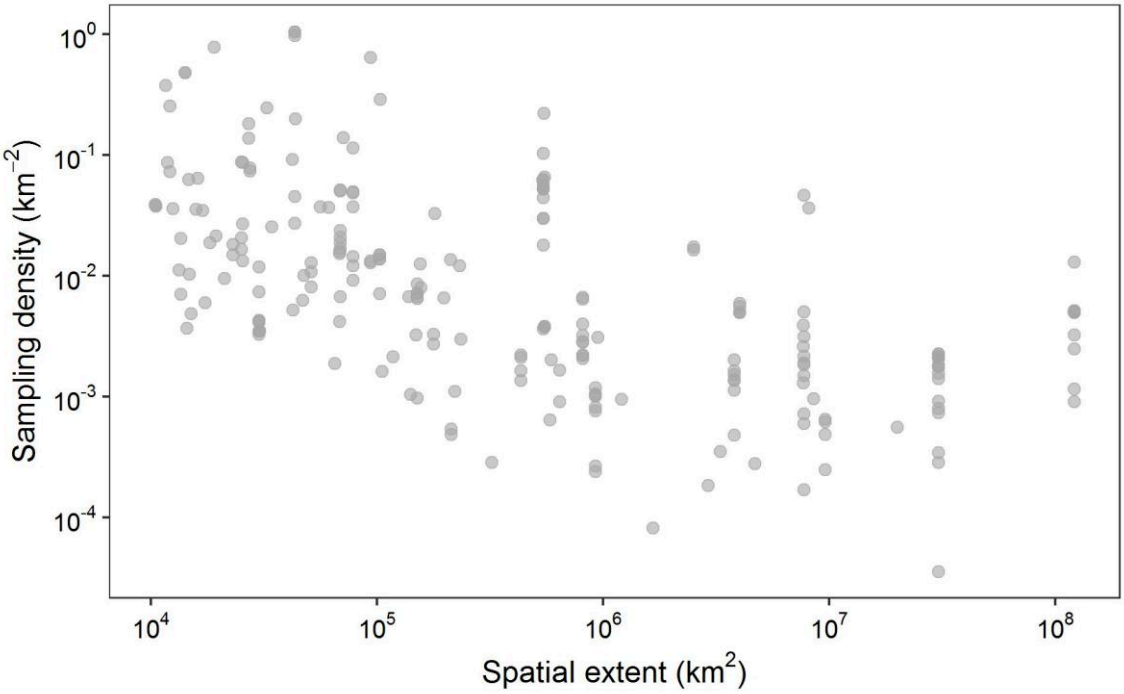


Figure 2.4 Trends between spatial extent and soil sampling density. The studies with a spatial extent larger than 10^8 km² are focused on a global scale

Figure 2.5 shows the soil sampling year reported in 160 articles. It showed a large time interval from 1920s to 2010s and 41% of these studies did not provide soil sampling year used for DSM. A total of 31% of the studies used soil data entirely collected after the year of 2000 while 28% of the studies also used the historical soil information before 2000s.

Figure 2.6 shows the soil sampling design used for soil information collection. More than half (56%) of the studies did not report how these soil data was collected, partially due to the fact that soil databases were compiled from various sources of historical soil information for different purposes. Apart from this, probability sampling was the most frequently adopted approach (29%) for soil sampling design whereas only 4% of these studies used purposive sampling design. The 'mixture' sampling (probability sampling+purposive sampling, probability sampling+no reported data, or purposive sampling+no reported data) counted for 10% of all the studies.

The frequency of validation strategy is presented in Figure 2.7. Internal validation (i.e., random holdback, data splitting) was the most frequent strategy (52%) for model evaluation in broad-scale DSM, and it was followed by cross-validation (i.e., *k*-fold cross-

validation, leave-one-out cross-validation), which accounted for 27%. We also observed that 6% of these studies used external validation and 7% provided the both results from internal validation and cross-validation. However, there was still 8% of the articles did not show any validation.

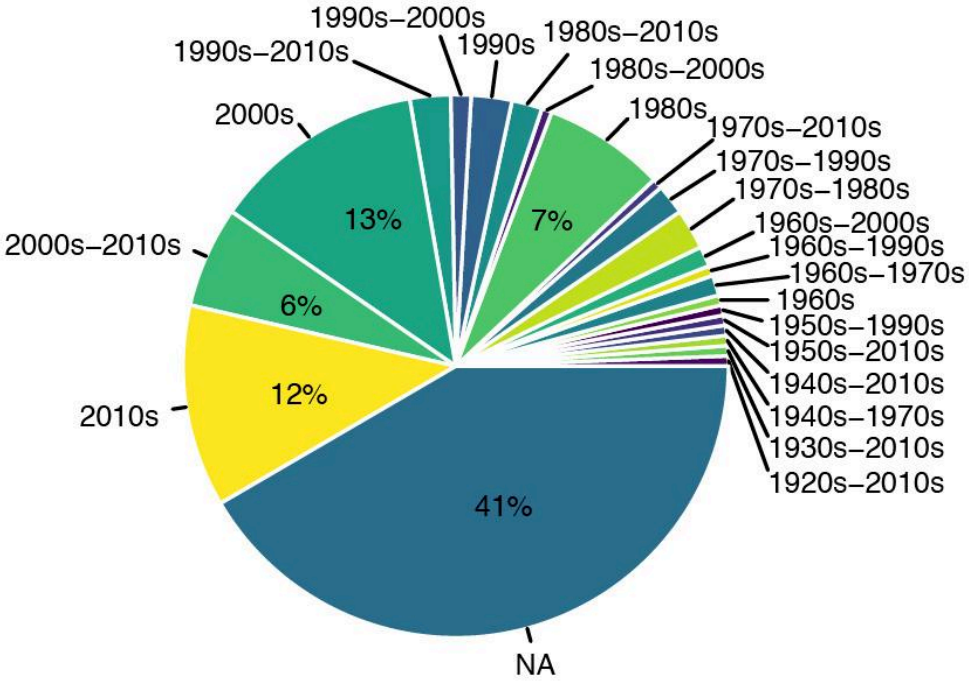


Figure 2.5 Soil sampling year

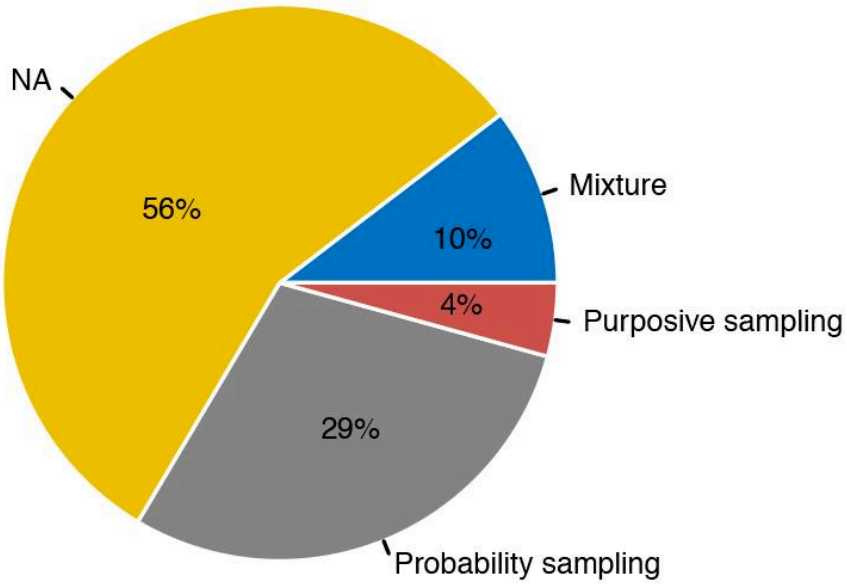


Figure 2.6 Soil sampling design

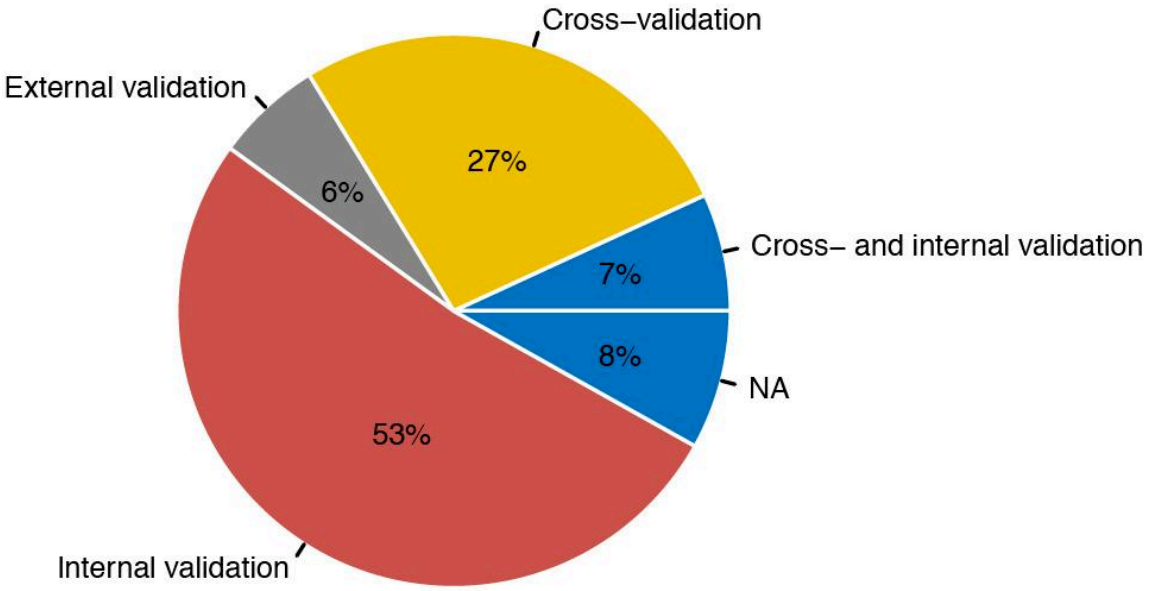


Figure 2.7 Frequency of types of validation strategy

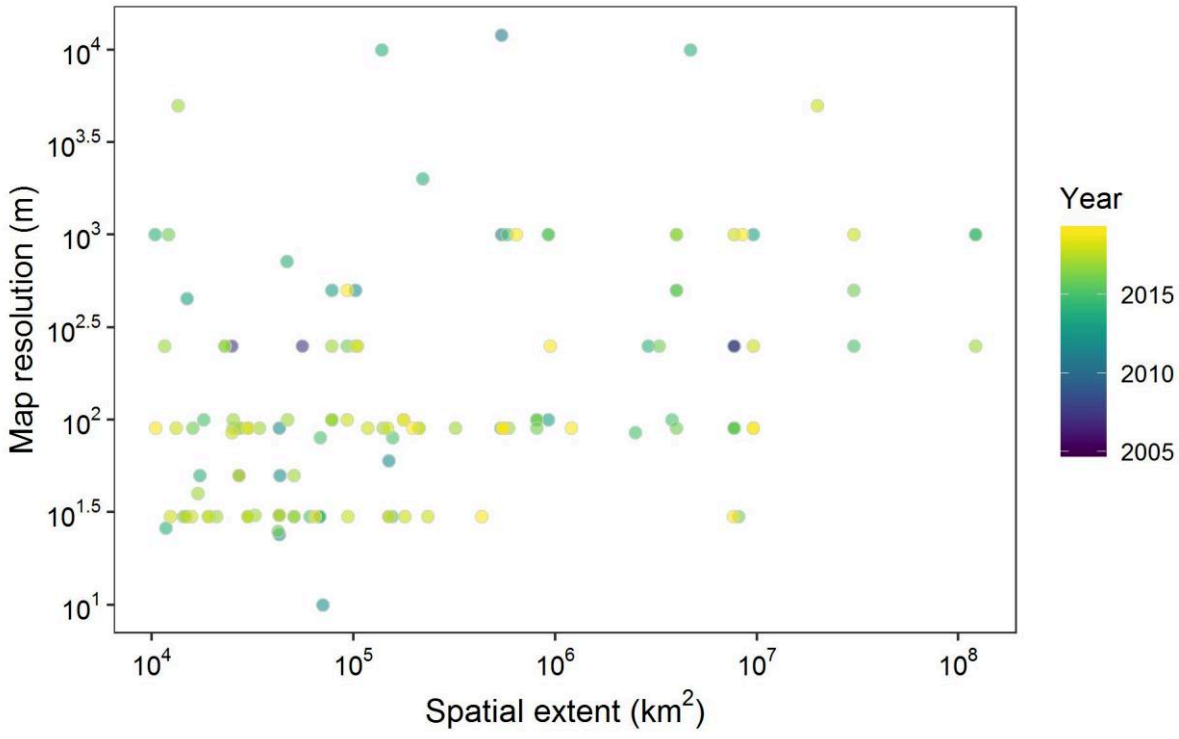


Figure 2.8 Relationship between map resolution (grid size) and spatial extent. The colour of points indicates the year of publication

Figure 2.8 presents the resolution of produced map under different spatial extents. The map resolution ranged from 10 to 10,000 m. Excluding 15 articles without the definition of map resolution, it suggested a slight increasing trend of map resolution with the increasing spatial extent. There is also a general trend that recent articles produced digital soil maps with high resolution.

The frequency of spatial predictive models is shown in Figure 2.9. The spatial predictive models were divided into six groups: (1) Geostatistical model included pure Geostatistics and spatial models such as Simple Kriging, Ordinary Kriging, Universal Kriging (Kriging with External Draft), Bayesian Kriging, Area to point Kriging, Filtered Kriging, Regression Kriging, Sequential Gaussian simulation, Geographically Weighted Regression, spatial trend and Integrated Nested Laplace Approximation with Stochastic Partial Differential Equations (INLA-SPDE); (2) Conventional model comprises conventional statistical regression and classification methods excluding machine learning algorithms, such as Multiple Linear Regression, Partial Least Square Regression, Principle Component Analysis, Nearest Neighbor, General Linear Model Fuzzy Logic model, and Structure Equation Model; (3) Machine learning included Cubist, Classification and Regression Trees, Multiple Additive and Regression Trees, Boosted Regression Trees, Bagged regression trees Random Forest, Artificial Neural Network, Multinomial Logistic Regression, Bayesian Networks, Support Vector Machine, XGBoost, Quantile Random Forest, Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (DSMART), Random Survival Forest, and Convolutional Neutral Network; (4) Hybrid model was the integration of regression model and Geostatistics in which the method first fitted a regression model (based on machine learning), then performed Geostatistics on the regression residual, and finally merged these two parts as the final predictions; (5) Others included these models that were not within previously mentioned classes, such as model averaging and taxonomic distance. As the number of articles related to broad-scale DSM was rather low before 2010, there was no clear trend among different groups. Machine learning became the dominant group since 2011, while hybrid model ranked the second between 2015 and 2017 and then decreased in the last two years. Conventional model and geostatistical model were still frequently used in the last five years.

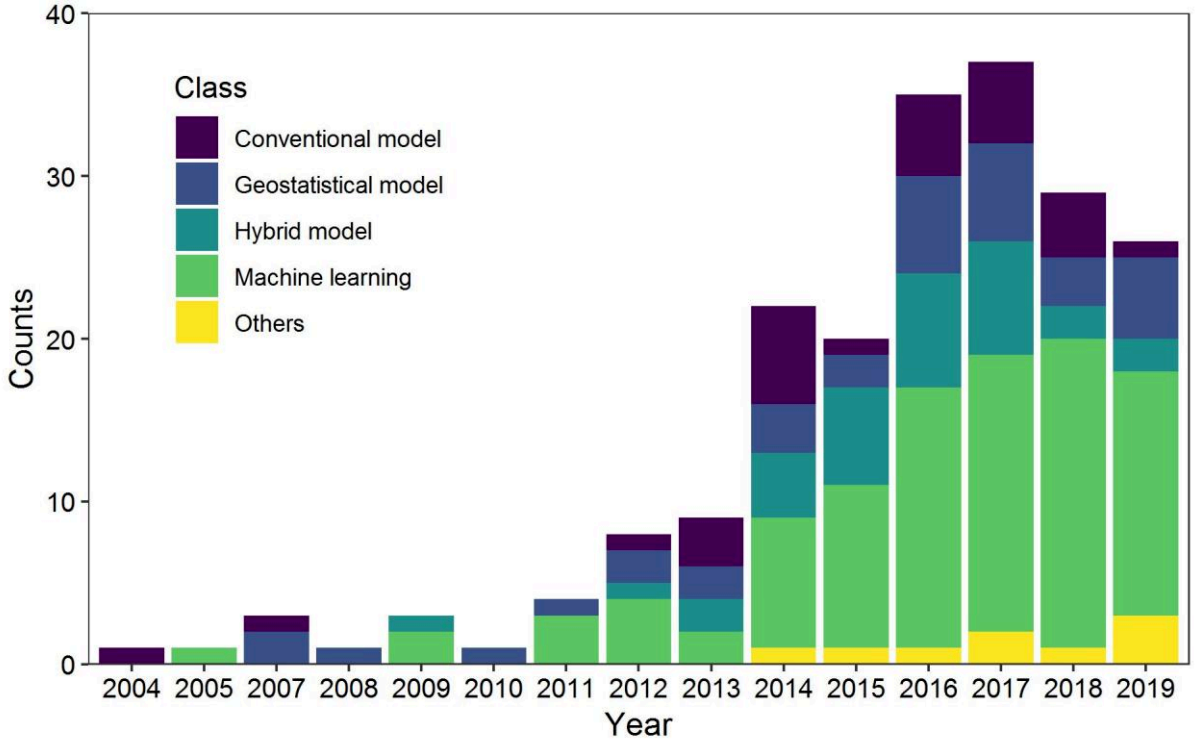


Figure 2.9 Frequency of different spatial predictive models

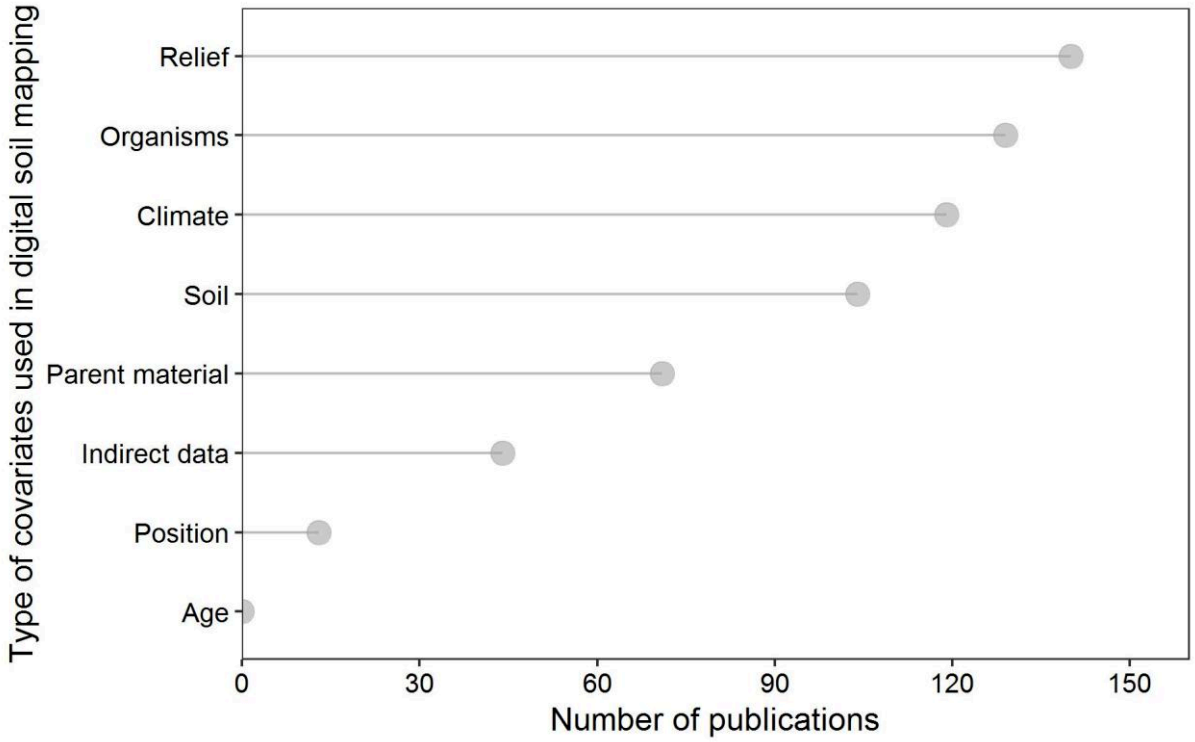


Figure 2.10 Frequency of *Scorpan* covariates used in DSM

Figure 2.10 shows the frequency of *scorpan* covariates used in DSM. The relief was used in 134 articles, making it to be the most frequent covariate. Being used in 122 and 113 articles, organisms and climate covariates ranked the second and third places. They were followed by soil and parent materials which were used in 100 and 69 articles, respectively. The indirect data here referred to the original bands information (not to index calculated from several bands) from airborne or satellite images, and it was used in 40 articles. Position was only used in 12 articles and age was not used for any studies in this review.

The target soil properties in broad-scale DSM are present in Figure 2.11, in which soil properties with a frequency less than 3 were classified as "Others". SOC content was the top one soil property of interest and it occurred in around 60 articles. Being the second place, SOC stocks and clay both had been studied in 33 articles. Soil class, sand and pH were mentioned in more than 20 articles. Other soil properties, such as silt, bulk density (BD), cation exchange capacity (CEC), soil organic matter (SOM), soil depth, coarse elements, total nitrogen (TN), and total phosphorus (TP), were relatively less predicted in broad-scale DSM.

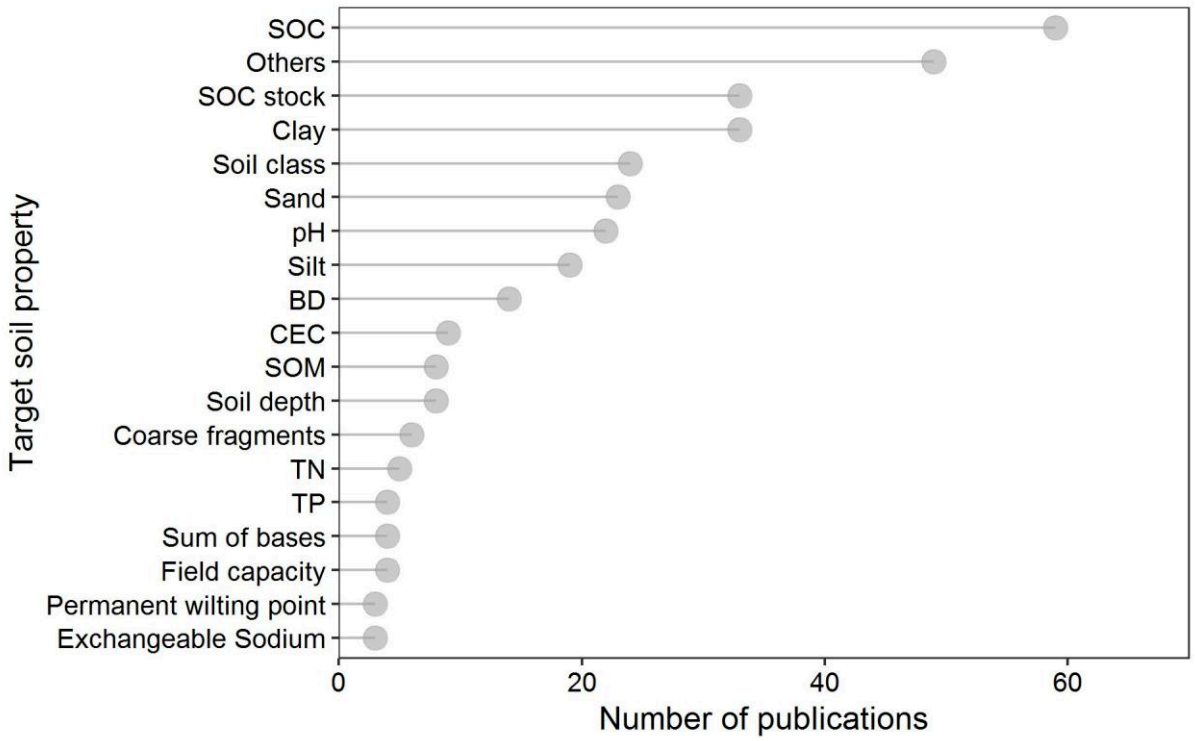


Figure 2.11 Soil properties of interest in broad-scale DSM. The category of "Others" indicates the soil properties that occur less than three times

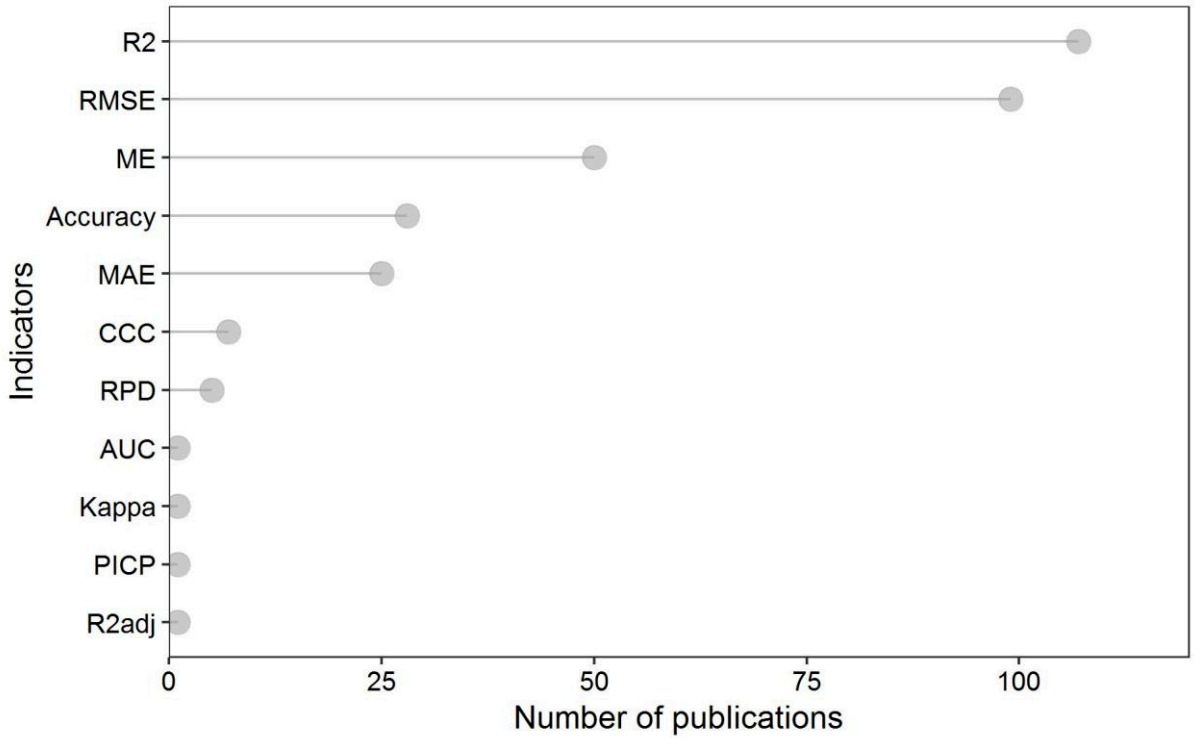


Figure 2.12 Frequency of performance indicators

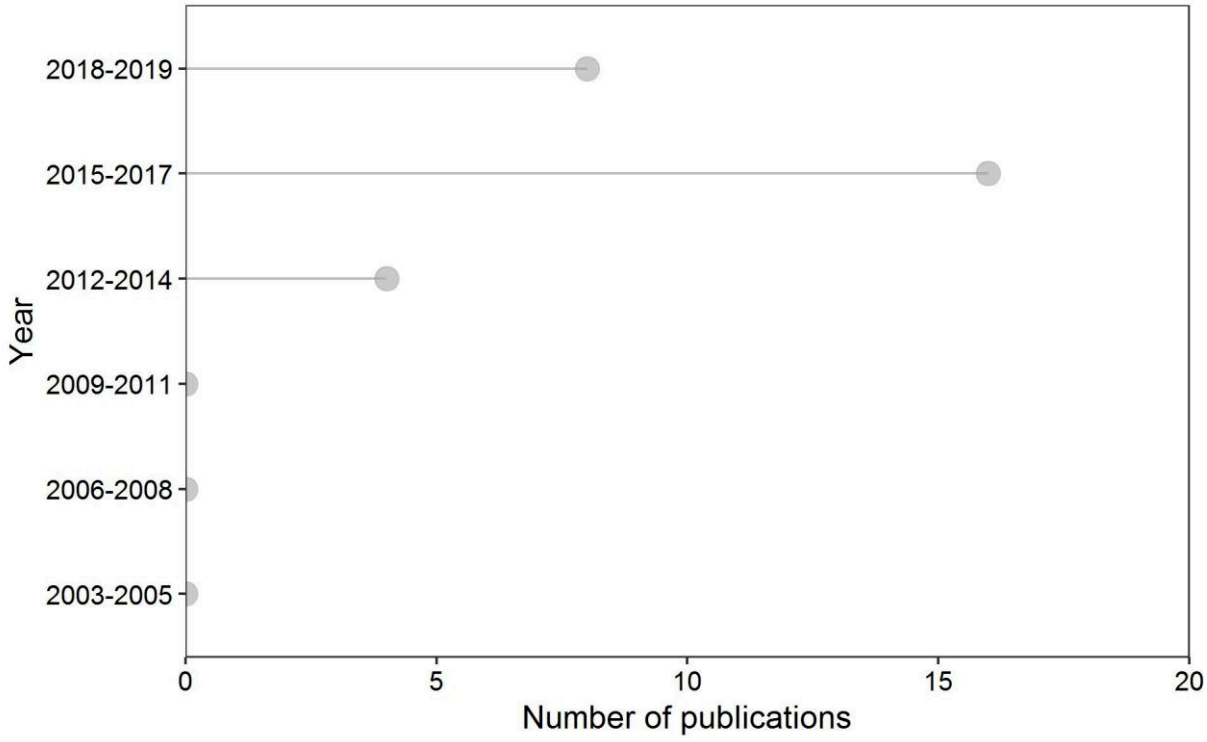


Figure 2.13 Frequency of *GlobalSoilMap* like articles

Figure 2.12 presents the common indicators used to evaluate model performance. The determination coefficient (R^2) and root mean square error (RMSE) were the top two indicators that had been used for more than 100 articles. Around 50 articles used mean error (ME) as one of the indicators, making it ranked third. Accuracy and mean absolute error (MAE) were used to for model evaluation in nearly 30 and 25 articles, respectively. Other indicators, such as Lin's concordance correlation coefficient (CCC), ratio of performance to deviation (RPD), area under the ROC curve (AUC), Kappa, prediction interval coverage probability (PICP) and adjusted R^2 (R^2_{adj}) were relatively less used.

Figure 2.13 shows the frequency of *GlobalSoilMap* like articles. Once an article met both the following two requirements, it belonged to *GlobalSoilMap* like articles: (1) it used the depth intervals defined by *GlobalSoilMap* (i.e., 0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 cm); (2) it provided the estimates of the map uncertainty. In total, 28 out of 160 articles were classified as *GlobalSoilMap* like articles, and all of them were published after 2012. In details, there were 4, 16, and 8 articles published during the time intervals of 2012-2014, 2015-2017 and 2018-2019.

2.4 Discussion

2.4.1 Data source: legacy data and sampling strategy

2.4.1.1 Time scale and sampling density of legacy data

The soil data used in broad-scale DSM can date from the 1920s and it is a common case that all the available soil data are used for modelling soil properties without taking time scale into account due to the scarcity of soil information (Figure 2.5). This practice is acceptable for some stable soil properties (i.e., soil depth, particle size fractions, soil class) in 100 years, but it can introduce a large uncertainty for other soil properties that may change in rather short time scales (from years to decades) such as SOC, pH and CEC. Generally, two solutions may help to solve this issue: (1) only use the soil data within a given period (i.e., 2000-2010) in model training; (2) incorporate the sampling year as a covariate and built a space-time model (2D+time or 3D+time). The first strategy was adopted by Stockman et al. (2015) to map SOC stocks at a global scale in 1960s, 1980s, 1990s and 2000s and to assess their spatial-temporal changes. The second strategy has not been explored in large-scale DSM so far and it needs more future studies.

It is expected that the sampling density of soil legacy data generally decreases with

the increasing spatial extent. The global studies used WoSIS soil database are exceptions due to the huge soil data from USA and the Europe. At a country level, Denmark (1.05 sample km^{-2}), Hungary (0.64 sample km^{-2}), France (0.22 sample km^{-2}), Estonia (0.20 sample km^{-2}), Australia (0.05 sample km^{-2}), and USA (0.04 sample km^{-2}) are among the countries having the highest sampling densities to produce digital soil maps at a national scale. Figure 2.2 shows that a large percentage of the countries in Africa and Asia still lack of broad-scale DSM products. However, in practice, this gap is often filled some continental (i.e., AfSIS project) and global initiatives and projects (i.e., SoilGrids).

In a review on soil legacy data rescue, Arrouays et al. (2017) mentioned that about 800,000 soil profiles were rescued in countries which responded to their survey and they were likely to be largely underestimated. Despite the great success of DSM and data rescuing, the majority of soil data is still "lost" and not digitalized in hard copy format. Therefore this effort should be pursued by using deep learning (e.g., image analysis, text recognition) to speed up the procedure in collecting legacy soil data.

2.4.1.2 Soil sampling design

Though 56% of the studies do not provide any information about soil sampling strategy in Figure 2.6, the majority of them are supposed to use purposive sampling as most of the legacy data came from historical soil surveys which were more or less designed for certain purposes and their sampling strategies were mainly based on expert knowledge. Brus (2019) noted that there is no best sampling design, and the best one depends on the techniques used for DSM. For the purpose of producing digital soil maps, systematic grid sampling or regular geographical coverage sampling is recommended when no environmental covariate is available (Walvoort et al., 2010). These two sampling designs are also in favour of the construction of soil monitoring networks which aim to monitor various soil properties simultaneously. In presence of environmental covariates, stratified random sampling offers an efficient way to cover the spatial variation of some soil properties of interest and also provides efficient statistical estimates (De Gruijter et al., 2006; Minasny et al., 2013). Proposed by Minasny and McBratney (2006), the conditioned Latin hypercube sampling (cLHS) is a modified version of stratified sampling design that enables to select sampling locations covering the distributions and combinations of the environmental covariates (feature space), and it has been applied in

many DSM studies (Mulder et al., 2013; Rad et al., 2014; Thomas et al., 2015; Omuto and Vargas, 2015). Based on the *scorpan* model, cLHS assumes that covering the feature space will allow to cover the soil properties variations, and is often used to maximize the efficiency of sampling when the number of samples is rather small. There are also other ways of stratification for maximizing sampling efficiency, such as feature space coverage sampling with *k*-means (Brus et al., 2007; Brus, 2019; Wadoux et al., 2019a). In theory, soil sampling design for DSM should both cover the feature space of the soil property of interest and the geographical space (Heuvelink et al., 2006). The feature space of the soil property of interest, however, may be unknown, and when several variables have to be mapped, their distribution will be likely different. This is why systematic sampling such as grid sampling is often used in the absence of prior knowledge. In case of geostatistical applications, it is generally recommended to add a subsets of close-pair units in geographical coverage sampling to better estimate short range variability and to fit the variogram (Marchant and Lark, 2007; Wadoux et al., 2019b). For these regions having already used legacy data for DSM, one of the outcomes could be to design more efficient supplementary sampling campaigns in the locations with greater uncertainty.

2.4.2 Prediction, modelling and mapping

2.4.2.1 Soil information of interest

Studies on mapping SOC and SOC stocks account for largest proportion of the articles in broad-scale DSM,. It results from the significant role of SOC on global C cycle and ecosystem services (e.g., food production, climate regulation, erosion control and water regulation) (Koch et al., 2013; Adhikari and Hartemink, 2016; Rumpel et al., 2018). The reliable assessment of SOC stocks is able to provide supporting information to address aforementioned issues. Due to the same reason, the Global Soil Partnership (GSP) published the first global soil map on SOC stocks (GSOCmap) in order establish a baseline with the ultimate goals to monitor the soil condition, identify degraded areas, set restoration targets, explore SOC sequestration potentials, support the greenhouse gas (GHG) emission reporting and make evidence-based decisions to adapt and mitigate to climate change.

Soil particle size fractions (i.e., clay, silt and sand) are the second frequently studied basic soil properties, which are important for soil hydrologic properties (i.e., AWC and

soil moisture), erosion, biogeochemical and crop modelling. It should be noted that a large proportion of studies predict each component of particle size fractions separately, which ignore the fact that the sum of particle size fractions is a constant of 100%. To overcome this drawback, additive log-ratio transformation (*alr*) or similar algorithms can be applied to convert three particle size fractions to two ratios between them before modelling, and this strategy has been performed in several broad-scale DSM studies (i.e., Akpa et al., 2014; Ballabio et al., 2016; Poggio and Gimona, 2017; Román Dobarco et al., 2019a). The *alr* is defined as:

$$clay_{alr} = \ln\left(\frac{clay}{sand}\right) \quad (2.2)$$

$$silt_{alr} = \ln\left(\frac{silt}{sand}\right) \quad (2.3)$$

These two *alr* transformed variables are modelled and mapped separately, and they are finally inverse transformed into three particle size fractions, which is defined as:

$$sand = \frac{1}{\exp(clay_{alr}) + \exp(silt_{alr}) + 1} \quad (2.4)$$

$$clay = \frac{clay_{alr}}{\exp(clay_{alr}) + \exp(silt_{alr}) + 1} \quad (2.5)$$

$$silt = \frac{silt_{alr}}{\exp(clay_{alr}) + \exp(silt_{alr}) + 1} \quad (2.6)$$

For difficult-to-measure soil properties, such as BD and AWC, pedotransfer functions (PTFs) are commonly used to derive data using easy-to-measure soil information (e.g., SOC, particle size fractions) before spatial modelling. However, validity domain of PTFs should be defined in order to avoid invalid spatial extrapolation and thus lead to larger uncertainty in the spatial predictive model (McBratney et al., 2002). Tranter et al. (2009) suggested the use of distance metrics (i.e., Mahalanobis distance and Standardized Euclidean distance) to determine the validity domain of PTFs and the fitted PTFs should not be applied to the samples outside of the validity domain. Several studies have recognized this issue and used the validity domain concept in predicting BD and AWC using PTFs (Chen et al., 2018; Román Dobarco et al., 2019b).

2.4.2.2 Environmental covariates

The frequency of *scorpan* factors are more or less restricted by the availability of

environmental covariates (Grunwald, 2009). Benefiting from global free-available remote sensing data, relief, organism and climate factors have been widely used in broad-scale DSM while the frequency of other factors such as soil, parent material, age and position are relatively lower (Figure 2.10).

Common relief variables derive from digital elevation model (DEM) by Shuttle Radar Topography Mission (SRTM) (Jarvis et al., 2008) and its derivatives such as aspect, slope, curvature, roughhouses, topological position index (TPI), channel network base level (CNBL), terrain wetness index (TWI), and multi-resolution valley bottom flatness (MrVBF). These DEM derivatives can be easily calculated by GIS softwares such as QGIS, SAGA GIS, GRASS GIS and ArcGIS. Recently, a new DEM product, named Multi-Error-Removed Improved-Terrain DEM (MERIT DEM, http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/) was developed Yamazaki et al. (2017) to remove multiple error components (i.e., absolute bias, stripe noise, speckle noise, and tree height bias) from the existing space-borne DEMs. The spatial resolution of this product is 3 arc-second (about 90 m at the equator).

Organisms factor is usually represented by land use/land cover (LULC), vegetation index (e.g., Normalised Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI)) and net primary productivity (NPP). The LULC data are often extracted from regional to global scales products in broad-scale DSM. At a global scale, several latest LULC products are recommended: (1) ESA CCI Land cover (http://maps.elie.ucl.ac.be/CCI/viewer/) provides 300 m annual global land cover time series from 1992 to 2015, and it describes 37 original land cover classes based on the United Nations Land Cover Classification System; (2) GLOBELAND30 (http://www.globallandcover.com) is 30m resolution global land cover data for 2010, and it describes 10 land cover classes (Chen et al., 2015).

Common climatic variables include temperature and precipitation which are either produced by spatial interpolation of the observations from meteorological stations or an integration from remote sensing data and meteorological stations. Under the soaring demand of fine-resolution DSM products, the climatic data derived from the second approach is preferred due to the requirement of high data accuracy and spatial resolution. There are several high resolution global climatic data free available for broad-scale DSM: (1) MODIS MOD11A2 (https://modis.gsfc.nasa.gov/data/dataprod/mod11.php) product

provides an average 8-day land surface temperature (LST) from 2000 to present; (2) WorldClim version 2 (https://worldclim.org/version2) provides global average monthly climate data for minimum, mean, and maximum temperature and for precipitation, solar radiation, wind speed and water vapor pressure for 1970-2000 at 1 km resolution (Fick and Hijmans, 2017). It also contains 19 bioclimatic variables calculated by the average for 1970-2000; (3) CHELSA (http://chelsa-climate.org) provides a global monthly mean temperature and precipitation for 1979-2003 at 1 km resolution as well as future climate projection (CMIP) under RCP 2.6, RCP 4.5, RCP 6.0 and RCP 8.5 (Karger et al., 2017); (4) TerraClimate (http://www.climatologylab.org/) is a global dataset of monthly climate (minimum and maximum temperature, precipitation, solar radiation, wind speed and water vapor pressure) and 7 climatic water balance for 1958–2015 at 4 km resolution (Abatzoglou et al., 2018).

In broad-scale DSM, soil factor is often characterized by soil class maps and/or soil texture maps derived from historical soil surveys (Grunwald, 2009). Soil information from proximal soil sensing can also be spatially interpolated and then served as covariates for DSM, for example, the first three principle components (PCs) of visible–near infrared (Vis-NIR) spectra have been used for producing Australian three-dimensional soil grid (Viscarra Rossel et al., 2015). Other soil information, including soil moisture and soil property maps from other sources, are also can be used as covariates in spatial modelling (Keskin et al., 2019).

Parent material is mainly derived from geological maps, and partially from airborne gamma ray in some countries/regions such as USA, Australia, and Brittany (Lacoste et al., 2011; Viscarra Rossel et al., 2015; Keskin et al., 2019). Based on the statistics in Figure 2.12, the usage of parent material is low due to the lack of data sources. With the technical advances in airborne gamma ray spectrometry and proximal gamma ray spectrometer, the data for parent material will be more available in DSM practice.

Position factor can be represented by spatial coordinates (i.e., latitude and longitude) or transformation of original spatial coordinates (e.g., the distance to coastal line). Though position factor is easy to obtain, only a few studies incorporate them as covariates in modelling, and this issue is rarely mentioned by previous reviews on DSM. As putting position factor is a simple way to ensure that spatial trends not included in the other environmental variables are not missed, McBratney et al. (2003) suggested to include this

scorpan factor in modelling. Another more classical way of taking the spatial position into account is to use hybrid models in which the regression part is fitted by machine learning approaches, and the regression residuals is fitted by geostatistical models.

Despite its significant role in pyogenesis, age is still the least used *scorpan* factor due to the difficulty of direct measurement at a broad scale (McBratney et al., 2003; Zhang et al., 2017). Considerable advance in technology (e.g., soil and material dating), geomorphology information, and expert knowledge are still needed to derive age factor, especially for broad-scale DSM.

One should keep in mind that the importance of environmental covariates may be property-specific and location-specific. As many studies concentrated their efforts on SOC, this may partly explains why relief, land use and climate are the dominant covariates that are presently used. One may think that if the soil properties of interest were, for instance, clay mineralogy, then lithology, age and climate would be more explored as controlling factors.

2.4.2.3 Spatial predictive models

Machine learning (or data mining) has become the most commonly used spatial predictive models in broad-scale DSM since 2011 (Figure 2.9), and this trend has also been confirmed by Arrouays et al. (2019, submitted) during the last Joint Workshop on Digital Soil Mapping and *GlobalSoilMaps*. It mainly results from two reasons: (1) machine learning is able to deal with complex non-linear relationship between soil property of interest and increasing number of environmental covariates, and thus it has better performance than other models; (2) the rapidly increasing computing power and techniques (e.g., parallel computing, cloud computing, and high-performance computing) makes it more efficient to produce digital soil maps on big data using DSM than ever before. However, geostatistical models are still useful as they may better capture some spatial structures than pure machine learning model, and thus the use of hybrid modelling for broad-scale DSM can be a good choice to take the merits from both machine learning and geostatistical models.

The recent advance of spatial predictive models in DSM is the introduction of deep learning (i.e., convolutional neural network), which has been explicitly described in Padarian et al. (2019), Wadoux (2019) and Wadoux et al. (2019c). Deep learning opens

new possibilities to predicting soil properties because (i) the input data for model training is a stack of spatial patterns, not spatial points; (ii) the trained model enables to provide simultaneous prediction of multiple soil properties (Padarian et al., 2019).

Despite the great advances in machine learning and deep learning, spatial predictive models seems to focus more on prediction performance and forget the importance of pedological knowledge for DSM and the use of DSM in understanding the controlling factors of soil property of interest. Therefore, future DSM should put more efforts on opening the "black boxes" of machine learning and deep learning, and integrating more pedological knowledge in both spatial predictive model and environmental covariates selection (Arrouays et al., 2019, submitted).

2.4.3 Performance validation and uncertainty estimation

2.4.3.1 Validation strategy

As mentioned in the results, the validation of digital soil mapping can be done in several ways: (1) internal validation uses random hold back or data splitting, which means a certain percentage of the data (20-40%) are randomly selected and excluded in model training (or calibration). This selected data is used to evaluate the performance of trained model; (2) cross-validation includes *k*-fold cross-validation and leave-one-out cross-validation. In *k*-fold cross-validation, the data is divided into *k* fold, the *k*-1 fold is used for model training and the 1 fold is used for model validation, and then this procedure repeated *k* times. Most of the recent broad-scale DSM use this type of validation and further use it to derivate uncertainty estimates (e.g., Mulder et al., 2016a; Kempen et al., 2019; Loiseau et al., 2019). Leave-one-out cross-validation is a special type of *k*-fold cross-validation in which only one sample is left out to validate the model trained by the *n*-1 data and it repeats *n* times, it is often used when the sampling data are rather sparse, in order to keep the maximum of data for calibration; (3) Independent validation uses additional samples different from training data to evaluate model accuracy.

Considering the random sampling, the selected validation data in internal validation may not represent the whole data and thus results in non-robust accuracy (Lagacherie et al., 2019). For overcoming this issue, repeated internal validation (i.e. 100 times) can be applied and the final validation results are calculated by the mean of all the repeats.

Brus et al. (2011) recommended the use of independent validation because internal

validation and cross-validation may not provide un-biased accuracy assessment because of the non-random sampled soil data. These additional independent data can be collected by a design-based sampling strategy involving probability sampling and design-based estimation. Due to the high cost of additional soil sampling, only a few studies used independent validation for map evaluation at a broad-scale (Thomas et al., 2015; Rial et al., 2016; Vaysse et al., 2017; Bargaoui et al., 2019).

2.4.3.2 Indicators for model evaluation

As indicated in Figure 2.12, R^2 is the most commonly used indicators for model evaluation in continuous soil properties. The use of R^2 allows to compare the accuracy for different soil properties with various units and magnitudes, and thus it is recommended to be reported in the DSM studies. It has, however, several limitations for interpretation, because it strongly depends on the number of points used to calculate it and it is very sensitive to the presence of extreme values. RPD is another indicator that eliminates the difference in units and magnitudes. However, Minasny and McBrateney (2013) suggested not quote both RPD and R^2 as they are the same measure, and ratio of performance to interquartile range (RPIQ) is a better indicator than RPD for data is not normally distributed (Bellon-Maurel et al., 2010).

RMSE is a good indicator to present prediction error, but it is not suitable in accuracy comparison for different soil properties and even for the same properties with large differences in distribution. The drawback of RMSE may be solved by the use of relative RMSE (RRMSE), which can be calculated by:

$$RRMSE = \frac{RMSE}{\bar{y}} \quad (2.7)$$

where $\bar{y}$ is the mean of validation data.

Apart from previous mentioned indicators, mean error is also suggested to be reported in DSM studies as it enables to provide the information whether the prediction is un-biased, over-estimated or under-estimated.

2.4.3.3 Estimates of map uncertainty

Compared with conventional soil mapping, one advantage of DSM is the availability of an uncertainty or measure of confidence for the predicted map. About half (81) of these studies provided the estimates of map uncertainty and large percentage of them was published after 2015 and related to *GlobalSolMap* products. The approaches used for

uncertainty quantification can be classified into following groups (Malone et al., 2017): (1) Universal Kriging prediction variance; (2) Bootstrapping; (3) Empirical uncertainty quantification through data partitioning and cross validation; (4) Monte Carlo simulation; (5) Bayesian approach. The groups 1, 4 and 5 were mainly used in geostatistical models while the groups 2 and 3 were commonly used for machine learning model. The produced map along with its associated uncertainty are useful in decision making for end users, and they also allow to quantify uncertainty propagation in some secondary soil information (i.e., SOC stocks, AWC) and digital soil assessment (Finke, 2012; Poggio and Gimona, 2014; Román Dobarco et al., 2019a, 2019b). For example, Román Dobarco et al. (2019a) found that the main sources of uncertainty for soil available water capacity map were not the pedotransfer function for predicting AWC but the input maps of coarse fragments and particle size fractions.

2.4.4 Mapping soil information changes from the past to the future

Soil monitoring network is needed for mapping soil changes over time properly (Arrouays et al., 2012; van Wesemael, et al., 2011). However, no studies related to broad-scale DSM has been reported to map soil information changes using soil monitoring scheme, which is mainly due to fact that most of the established soil monitoring networks are not old enough to have several complete sampling campaigns (a noticeable exception is the England and Wales monitoring network, Bellamy et al., 2005). Therefore, using soil data from several periods under different sampling designs is an alternative way to map soil information, though it is not the best way to eliminate the prediction error among sampling designs. Sun et al. (2012) mapped SOM in topsoil (0-20 cm) for 1980s and 2006-2007 by sequential Gaussian simulation and showed that SOM increased by 0.22% in Jiangsu, China. Schillaci et al. (2017) modelled the topsoil (0-30 cm) SOC content of the cultivated area of Sicily, Italy in 1993 and 2008, and found that SOC decreased in the areas with relatively high initial SOC, and increased in the area with high temperature and low rainfall. Song et al. (2018) produced SOC and total nitrogen (TN) stocks maps for top 100 cm soil in 1980s and 2010s using random forest, and the results showed that 7.47×10^8 t C and 1.51×10^8 t N were accumulated during the past three decades in the Songnen Plain of Northeast China. Zhou et al. (2019) mapped SOC changes in topsoil (0-20 cm) between 1980s and 2004-2005 using random forest in North and Northeast China, and showed

that SOC increased 0.094 Pg in cropland and SOC decreased 0.126 Pg in forest and grassland. All these studies used soil data collected from different soil surveys (not from a consistent soil monitoring scheme), and they were able to provide general trends of soil changes during 20 to 30 years.

Currently, most the DSM studies focus on mapping soil status at one or several particular times in the past or in present, and several studies try to predict (or project) the likely soil status change for the future, especially for SOC. Minasny et al. (2013) stated that there are two ways for this purpose, including dynamic–mechanistic simulation model and static-empirical model. In dynamic–mechanistic simulation model, the digital soil map is fed as initial soil status and then the model is simulated per each pixel under future climate, LULC and land management scenarios. In static-empirical model, the future soil changes can be predicted using fitted *scorpan* model, in which the present climate, LULC and land management are replaced by relevant future scenarios. Gray and Bishop (2016) mapped SOC changes caused by projected climate change over New South Wales, Australia until 2070, and showed a mean loss rate of 2.0 Mg ha^{-1} in 0-30 cm and a total loss of 737 Tg of CO_2 equivalent in soil down to 100 cm. Yigini and Panagos (2016) predicted present SOC and future SOC stocks using climate and land cover scenarios in Europe (EU26), and the results showed an overall increase in SOC stocks by 2050. Meersmans et al. (2016) predicted SOC changes under climate and LULC scenarios in France by 2100, and showed that climate change would have a much bigger influence on future SOC losses in mid-latitude mineral soils than land use change dynamics. Reyes Rojas et al. (2018) projected SOC distribution in central Chile using climate scenarios and found that it would experience a loss of SOC in topsoil (0-30 cm) averaging 9.7% and 12.9% for RCP4.5 and RCP8.5 scenarios by 2050. For these broad-scale DSM studies, all of them uses static-empirical model, which is mainly results from several constrains such as (1) the large disconnection between DSM and mechanistic dynamics modelling, (2) complex parameter initialization and heavy computing which is challenging for mechanistic dynamics modelling at a broad scale (Walter et al., 2006; Luo et al., 2016). These challenges require the collaboration among the scientists from multiple disciplines as well as better integration between DSM and mechanistic dynamics modelling to speed up the simulation efficiency and improve the prediction accuracy (e.g., simulate observed locations by mechanistic dynamics model and then map soil information by DSM on

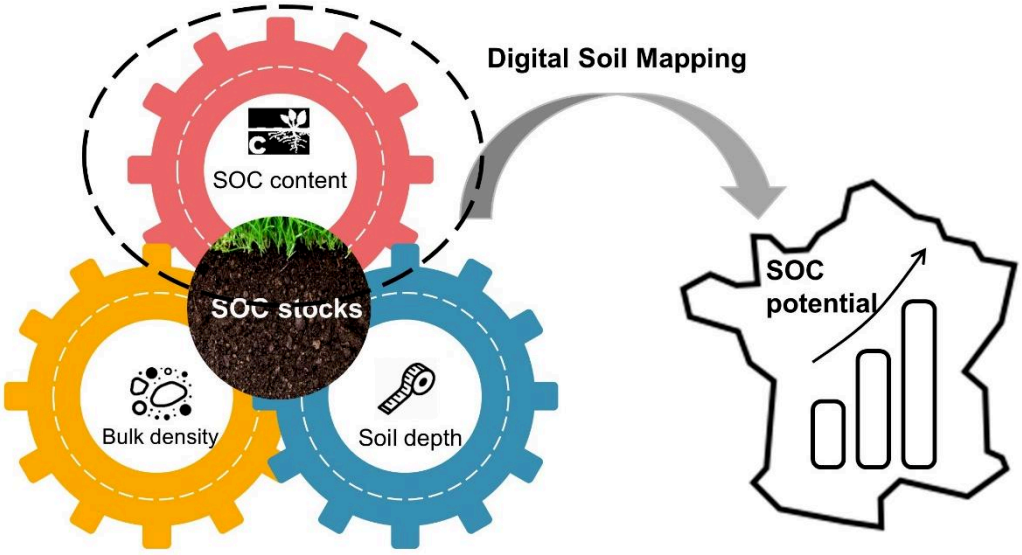
simulated data).

2.5. Conclusions

In this chapter, we reviewed 160 articles focused on broad-scale DSM from all over the world. Most of the DSM studies are clustered in several countries, such as China, Australia, USA, France and Denmark. It shows a clear trend that DSM technique has been increasing used to deliver soil information from regional to global scales. Meanwhile, DSM technique has been recognized and used by various global and governmental agencies and initiatives (i.e., GSP, FAO, 4P1000), which means DSM is shifting from pure research phase into operational use by decision and policy makers. Many legacy soil data has been rescued for the purpose of DSM, and this data rescuing should be continued together with new soil sampling campaign for a better understating of soil changes from the past to the present. Among all the soil properties of interest, SOC is the highest studied soil information due to its central role in global C cycle and ecosystem services. Soil sampling designs, environmental covariates and spatial predictive models also have been reviewed on broad-scale DSM, and we propose some related suggestions about methods and data sources for future work.

Chapter 3

Model averaging for mapping topsoil organic carbon in France



Chen, S., Mulder, V.L., Heuvelink, G.B.M., Poggio, L., Caubet, M., Román Dobarco, M., Walter, C., Arrouays, D. Model averaging for mapping topsoil organic carbon in France. Geoderma, under revision.

3.1 Introduction

Soils are crucial for maintaining ecosystem services such as food production, water regulation, erosion control, biodiversity, and climate regulation (Sanchez et al., 2009; Koch et al., 2013; Adhikari and Hartemink, 2016; Rumpel et al., 2018). To meet the increasing demand for up-to-date and fine-resolution soil information, Digital Soil Mapping (DSM, McBratney et al., 2003) has been widely adopted and is being rapidly developed across different spatial scales since the past decade (e.g., Grunwald et al., 2011; Poggio and Gimona. 2014; Viscarra Rossel et al., 2014; Hengl et al., 2015; Ballabio et al., 2016; Padarian et al., 2017; Sanderman et al., 2018; Chen et al., 2019c). At the global scale, different initiatives aim to deliver fine-resolution gridded soil information. The main examples are the recent Global Soil Parnership GSOC map (http://54.229.242.119/GSOCmap/), the *GlobalSoilMap* initiative (Sanchez et al., 2009; Arrouays et al., 2014a), and SoilGrids products (Hengl et al., 2017). SoilGrids adopts a "top-down" approach and produces soil property maps for the entire globe, which are freely distributed and available online (https://soilgrids.org/). *GlobalSoilMap* uses a "bottom-up approach" where each country produces soil property maps using its own national soil data and defined specifications (e.g., 3 arc second resolution, six standard depth intervals, quantified prediction uncertainty, Arrouays et al., 2014b). Then, these country-level soil maps are merged into a global map. There are also several initiatives producing soil property maps at the continental scale, such as LUCAS (Tóth et al., 2013) for Europe and AfSIS (Hengl et al., 2015) for Africa. As a result, there are often multiple maps available for a given soil property in a given area produced using various soil databases, environmental covariates, and DSM methods. Users may have multiple maps of the same property with different predictions and different map accuracy which may lead to confusion regarding which map should be used or whether the maps could or should be combined. It is possible to select the most suitable soil property map for a specific region, when the map accuracy can be evaluated using an independent validation dataset. When deciding to combine maps, the hypothesis is that the information provided by the maps is complementary and that a more accurate map may be obtained by merging the input maps using model averaging approaches (Caubet et al., 2019). The model averaging option needs an independent validation dataset and independent

calibration data to train the model averaging algorithm. Previous studies showed the potential of model averaging in improving the accuracy of soil property maps of pH, soil texture, and available water capacity (Malone et al., 2014; Padarian et al., 2014; Clifford and Guo, 2015; Román Dobarco et al., 2017; Caubet et al., 2019).

The choice between selecting a single map and combining multiple maps is not trivial, and many countries need to make this choice because of the increasing number of different prediction maps of the same soil property. It is particularly relevant to data-poor countries that may have very few or even no data to derive reliable country-based maps, and that could benefit from collecting a limited number of calibration samples to merge the national map with other existing products using model averaging.

The objectives of this study are to 1) evaluate the added value of applying model averaging in a data-rich country (e.g. France); 2) determine the most suitable model averaging approach for improving the topsoil (0-20 cm) SOC map of mainland France using three different SOC maps; 3) evaluate how well the model averaging approaches perform for different calibration sizes and optimize the calibration size required in model averaging; and 4) explore the potential of applying model averaging in data-poor situations.

3.2 Data

In this study, we used three SOC maps generated and harmonized from national, continental, and global DSM products and two national soil datasets in France.

3.2.1 French national soil organic carbon maps

Numerous maps have been generated for France following the *GlobalSoilMap* specifications. The most recent product (Mulder et al., 2016a) used all available point data for France, both from the French Soil Mapping and Inventory Program (Inventaire, Gestion et Conservation des Sols, IGCS) and an systematic grid aiming at monitoring French soil properties (RMQS). More details about these two datasets can be found in the study of Mulder et al. (2016a). For this study, we used the same *GlobalSoilMap* approach as Mulder et al. (2016a), but we set aside the RMQS grid to be used as an independent dataset for calibrating the model averaging algorithms and evaluating map accuracy (see Sections 3.3.3 and 3.3.4). Approximately 30,000 soil profiles from the IGCS dataset were used to generate SOC maps at the first three *GlobalSoilMap* depth intervals

(0-5, 5-15, 15-30 cm). The IGCS dataset is a compilation of soil profiles from many programs that mostly focused on agricultural soils. As a result, the soil profile density is high in some regions (Figure 3.1), whereas it is low in other regions; some land uses are over- or under-represented in the calibration dataset. SOC contents at the *GlobalSoilMap* depth intervals were obtained by applying equal area quadratic splines (Bishop et al., 1999; Malone et al., 2009) to soil profile data, as outlined in Mulder et al. (2016b). Spatially exhaustive covariates, including climate zones and meteorological data, vegetation, topography, geology, soils, and land management, were resampled to 90 m resolution. Details about these environmental covariates are given in Mulder et al. (2016a). In this study, the national SOC map (named IGCS SOC map hereafter) for the topsoil (0-20 cm) was calculated from SOC maps of 0-5, 5-15, and 15-30 cm by a weighted averaging approach, where the weights are proportional to the layer thickness (Figure 3.2a).

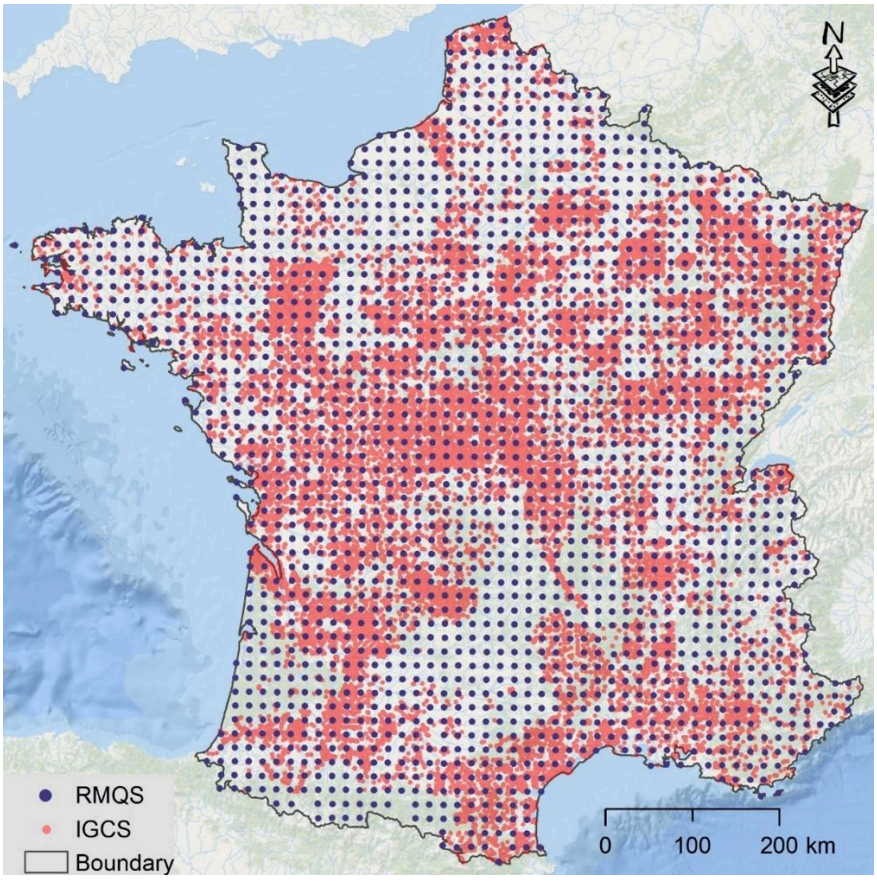


Figure 3.1 Study area (Mainland France) and soil sampling sites from IGCS and RMQS datasets

3.2.2 Continental and global scale soil organic carbon maps

In addition to the aforementioned national SOC map, we also obtained SOC maps for France from continental (LUCAS) and global (SoilGrids) soil map products.

The LUCAS SOC map (Figure 3.2b) contains SOC predictions for the topsoil (0-20 cm) at 1 km resolution for Europe (Aksoy et al., 2016). A total of 23,835 soil samples were used for model calibration. These soil samples were collected from LUCAS (19,860 samples), BioSoil (3,379 plots from forest soil), and SoilTrEC (387 samples from local soil data from six different critical zone observatories in Europe). From these datasets, about 3,500 sites were located in France. A regression kriging model was fitted to generate a SOC map using observed SOC content and 15 environmental covariates.

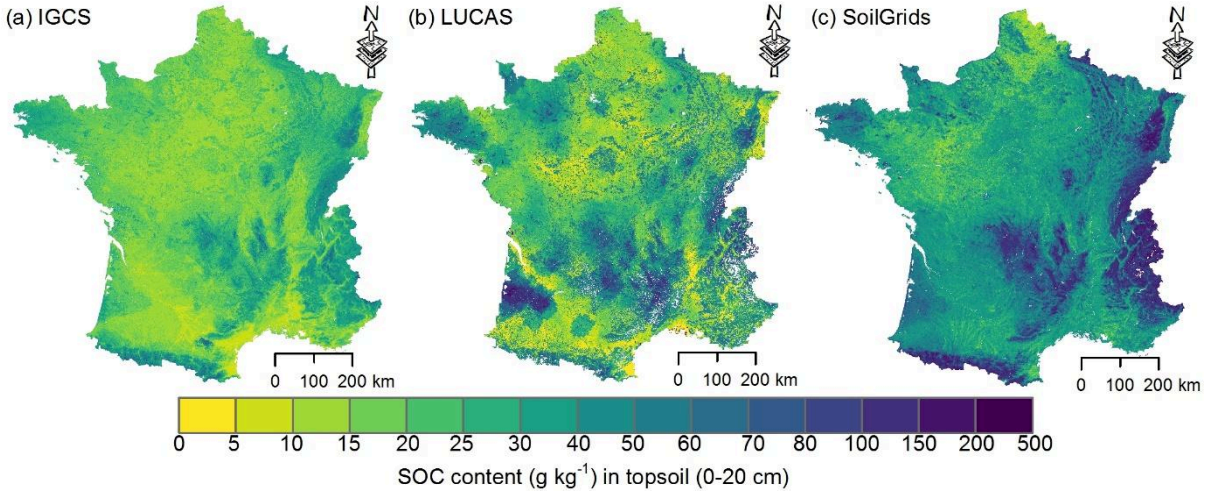


Figure 3.2 SOC maps of mainland France from IGCS (a), LUCAS (b) and SoilGrids (c)

The SoilGrids SOC map (https://soilgrids.org, v0.5.3, Figure 3.2c) was extracted from the study of Hengl et al. (2017), in which SOC was mapped at seven standard depths (0, 5, 15, 30, 60, 100, and 200 cm) at a resolution of 250 m for the globe. These SOC maps were based on about 150,000 soil profiles along with 158 remote sensing-based soil covariates. Maps were produced by fitting an ensemble prediction from random forest and gradient boosting trees. From the 150,000 soil profiles, nearly 3,000 were located in mainland France, mainly originating from the LUCAS database. For this work, the topsoil SOC map was calculated from SoilGrids SOC maps at 0, 5, 15, and 30 cm depth using trapezoidal numerical integration (Hengl et al., 2017).

The LUCAS and SoilGrids SOC maps were resampled to 90 m using bilinear

interpolation and reprojected to the Lambert 93 coordinate system to match these with the national SOC map.

3.2.3 Independent soil data for model averaging calibration and SOC map validation

To evaluate the accuracy of the input and merged maps, an independent validation dataset and an independent dataset for calibration of the model averaging algorithm were needed. These datasets were derived from the RMQS French systematic grid, which covers different soil, climate, relief, and land cover conditions (Figure 3.1). The RMQS dataset is a 16 km × 16 km square grid where sampling sites are at the centre of each grid cell, covering mainland France (Jolivet et al., 2006). For each site, 25 individual core samples were collected by a hand auger and mixed into a composite sample, both for 0–30 cm and 30–50 cm depth intervals. For more detailed information about the soil sampling design and laboratory analyses, refer to Martin et al. (2009). Because we did not have SOC measurements for a depth of 0-20 cm for the RMQS sites, we calculated these values depending on land use: 1) for most agricultural soils, SOC concentration decreases at a small rate with depth in the topsoil because of ploughing; thus, SOC content at 0-20 cm is close to that of 0-30 cm (Arrouays et al., 2001). We therefore used SOC at 0-30 cm to represent the SOC at 0-20 cm for RMQS sites under agricultural soils; 2) for natural soils (grassland and forest), SOC usually decreases with depth in the topsoil. Therefore, we first calculated SOC at 0-20 cm and at 0-30 cm by equal area quadratic splines using 5,785 grassland and forest soil profiles from the IGCS dataset. We then fitted a linear model between SOC at 0-20 cm and SOC at 0-30 cm ($SOC_{0\text{-}20\ cm} = 1.04 \times SOC_{0\text{-}30\ cm} + 0.26$, $R^2 = 0.986$). We used this model to derive SOC at 0-20 cm from SOC at 0-30 cm for all RMQS sites under natural soils.

3.3 Methods

3.3.1 Generic framework for model averaging

Figure 3.3 shows the generic framework for model averaging, which includes four steps. We first explain the procedure used for selecting the calibration and validation subsets from the RMQS dataset. To obtain spatially representative calibration and validation datasets, equal-size clustering (iterative nearest neighbour approach, Monlong, 2018) was applied to the RMQS sites (Step 1), which resulted in spatially

compact clusters. This was done for five cluster sample sizes (4, 10, 20, 50, and 100). Note that the cluster sample size is only approximately the same for all clusters because the total number of observations (i.e., 1996) is not always a multiple of the cluster sample size. Figure 3.4 shows the spatial distribution of the clusters. In Step 2, a *k*-fold cross-validation framework (*k* = 4, 10, 20, 50, 100) was used to separate a calibration set by randomly allocating one observation per cluster to each fold. Thus, the sample size of each fold was approximately 500, 200, 100, 40, and 20, for *k*=4, 10, 20, 50 and 100, respectively. In each of the *k* times, one of the folds was used to calibrate the model averaging approaches (Step 3), whereas the remaining *k-1* folds were used for model validation (Step 4, as explained in Section 3.3.2). By performing this analysis for different values of *k*, we could also evaluate the performance of the model averaging approaches for different calibration sizes (i.e. 500, 200, 100, 40, and 20).

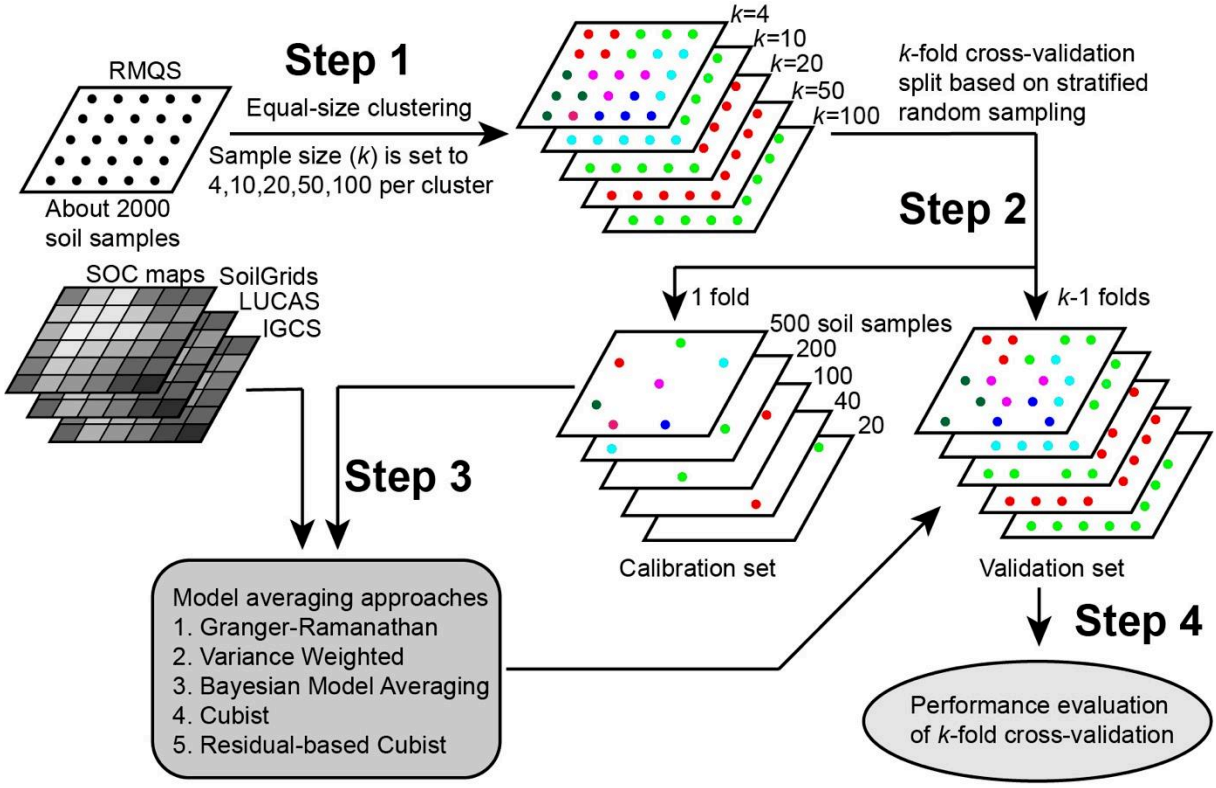


Figure 3.3 Model averaging workflow

3.3.2 Model averaging approaches

Five model averaging approaches were compared in this study. They are Granger-

Ramanathan (Granger and Ramanathan, 1984), Variance Weighted (Bates and Granger, 1969; Heuvelink and Bierkens, 1992), Bayesian model averaging (Hoeting et al., 1999), Piecewise linear decision tree (Quinlan, 1992), and Residual-based piecewise linear decision tree.

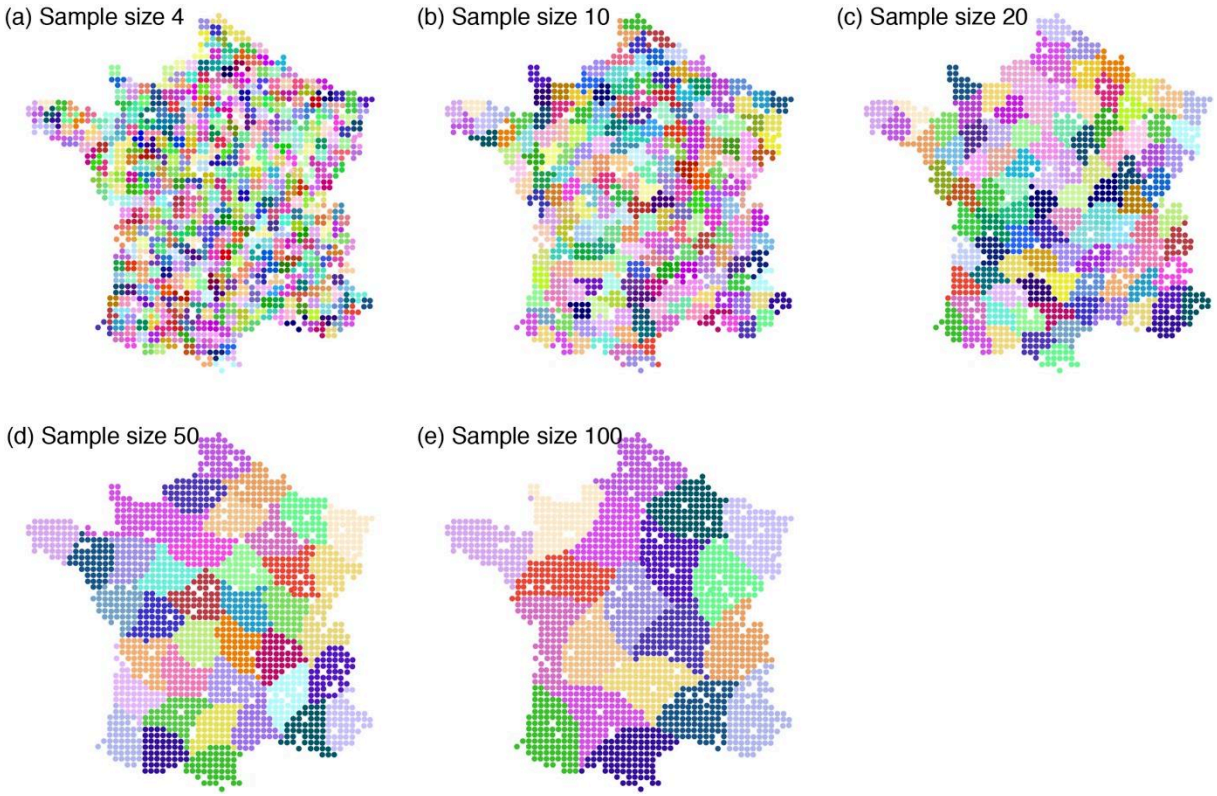


Figure 3.4 Spatial cluster distribution of RMQS sites, using equal-size clustering. The cluster sample sizes are 4 (a), 10 (b), 20 (c), 50 (d) and 100 (e)

3.3.2.1 Granger-Ramanathan

The Granger-Ramanathan (GR) approach was proposed by Granger and Ramanathan (1984). It assumes that a combination of different model predictions can be approached using a traditional Ordinary Least Square (OLS) method. In our case, a linear regression model was fitted between the measured SOC contents of the calibration set and the SOC predictions of the three SOC maps. The outcome SOC_{GR} from the GR approach can be calculated as

$$SOC_{GR}=\sum_{i=1}^{p}(\alpha_i \cdot SOC_i)+\beta \quad (3.1)$$

where α_i and SOC_i are the regression coefficient and SOC prediction of the i-th SOC map (p=3 in this study), and β is the intercept. The α and β coefficients are solved by the OLS method, and the sum of the α_i is not necessarily equal to 1.

3.3.2.2 Variance Weighted

We used the revised Variance Weighted (VW) approach from Ge et al. (2014), which is based on the error variance-covariance matrix that is estimated by comparing model predictions with observations. Thus, the outcome SOC_{VW} is calculated as

$$SOC_{VW}=\sum_{i=1}^{p}\alpha_i\cdot(SOC_i-\beta_i) \quad (3.2)$$

where α_i and SOC_i are the weight and SOC prediction of SOC map i, respectively, and β_i is the bias correction coefficient for SOC map i. The latter is calculated as

$$\beta_i=\frac{1}{m}\sum_{k=1}^{m}(SOC_{i,k}-SOC_{obs,k}) \quad (3.3)$$

where m is the number of calibration observations, and $SOC_{i,k}$ and $SOC_{obs,k}$ are the SOC prediction of SOC map i and the SOC observation at the k-th calibration site, respectively.

As described in Ge et al. (2014), the vector $\alpha=[\alpha_1\cdots\alpha_p]^T$ is calculated by minimizing the error variance of the model predictions:

$$\alpha^{\mathrm{T}}=(\mathbf{1}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{T}}\boldsymbol{V}^{-1} \quad (3.4)$$

where $\mathbf{1}$ is the p-dimensional identity matrix (recall that p=3 in this study), and $\boldsymbol{V}$ is the p-dimensional variance-covariance matrix of the prediction error. The elements of $\boldsymbol{V}$ are determined as

$$\widehat{\boldsymbol{V}}_{ij}=\frac{1}{m}\sum_{k=1}^{m}(SOC_{i,k}-SOC_{obs,k})(SOC_{j,k}-SOC_{obs,k}) \quad (3.5)$$

where i,j = 1,...,n represent SOC maps, and m is the number of calibration observations. Note that the correlations between SOC map errors are considered in the VW approach.

3.3.2.3 Bayesian Model Averaging

The Bayesian Model Averaging (BMA) approach assigns a conditional probability density function (PDF) to each model prediction (Hoeting et al., 1999). The BMA posterior distribution of the final output (SOC_{BMA}) can be expressed as (Raftery et al., 2005):

$$p(SOC_{BMA}|SOC_{obs})=\sum_{i=1}^{p} p(SOC_{BMA}|SOC_{obs},SOC_i)p(SOC_i|SOC_{obs}) \quad (3.6)$$

where SOC_{obs} are the SOC observations, p is the number of SOC maps (in this study p=3), and SOC_i denote the values of SOC extracted from the SOC map i at the locations of observations. Therefore, the BMA posterior distribution of SOC_{BMA} is a weighted average of the posterior distributions of SOC_{BMA} under each of the SOC maps, weighted by their posterior model probabilities.

The posterior model probability of SOC_i is expressed as (Raftery et al., 2005)

$$p(SOC_i|SOC_{obs})=\frac{p(SOC_{obs}|SOC_i)p(SOC_i)}{\sum_{l=1}^{p} p(SOC_{obs}|SOC_l)p(SOC_l)} \quad (3.7)$$

where $p(SOC_{obs}|SOC_i)$ is the integrated likelihood of SOC_i, and it can be calculated by BIC approximation (more details can be found in Raffery et al., 2005).

We used the R package "BMA" (Raftery et al., 2005) to apply BMA in our case study.

3.3.2.4 Piecewise linear decision tree

The Piecewise linear decision tree approach (Cubist) is based on the M5 algorithm (Quinlan, 1992). It partitions the dataset into several subsets within which inputs (independent variables) are similar. In a given subset, the standard deviation of the target values is treated as a measure of error and is used as a node splitting criterion. Every potential split is evaluated by the reduction in standard deviation. After evaluating all possible splits, Cubist chooses the one split that maximizes the reduction in error. Then, pruning and smoothing processes are performed to get the final model. More details are given in Quinlan (1992).

In the final Cubist model, partitions are defined by a list of rules, which are arranged in a hierarchy. Each rule has the following form:

if [condition] **then** [linear regression model]

else [apply next rule].

A rule indicates that whenever a case satisfies the condition of one rule, the corresponding linear regression model is used to predict the output. In this study, we used the R package "Cubist" (Kuhn et al., 2012).

3.3.2.5 Residual-based piecewise linear decision tree

The framework of Residual-based piecewise linear decision tree (Residual-based

Cubist, revised from Tao et al., 2018) is as follows: 1) calculate the arithmetic mean SOC value (SOC_{mean}) extracted from IGCS (SOC_{IGCS}), LUCAS (SOC_{LUCAS}), and SoilGrids ($SOC_{SoilGrids}$) SOC maps at locations of soil observations; 2) calculate the residuals (RES_{IGCS}, RES_{LUCAS}, and $RES_{SoilGrids}$) between SOC_{mean} and SOC_{IGCS}/SOC_{LUCAS}/$SOC_{SoilGrids}$, which are used as predictors in the Cubist model; 3) calculate the residuals (RES_{obs}) between SOC_{mean} and SOC observations (SOC_{obs}), which are used as the target variable in the Cubist model ; and 4) once the Cubist model is fitted, calculate the final SOC predictions of the Residual-based Cubist by summing up the RES_{obs} (derived from Cubist) and SOC_{mean}.

3.3.3 Evaluation of three SOC maps and five model averaging approaches using different calibration sizes

The performance of three individual soil SOC maps was assessed using all RMQS data. Based on a *k*-fold cross-validation framework explained in Section 3.3.1, we evaluated the five model averaging approaches using different calibration sizes (from 500 to 20). Three indicators, the Amount of Variance Explained (AVE), the Root Mean Square Error (RMSE), and Mean Error (ME), were used to evaluate prediction accuracy.

$$\text{AVE}=1-\frac{\sum_{i=1}^{n}(\hat{z}_i-z_i)^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2} \quad (3.8)$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i-z_i)^2} \quad (3.9)$$

$$\text{ME}=\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i-z_i) \quad (3.10)$$

where n is the size of the cross-validation dataset, z_i and $\hat{z}_i$ are measured and predicted values for the i-th observation in the cross-validation dataset, respectively, and $\bar{z}$ is the mean of the observations in the cross-validation datset.

3.3.4 The effect of national SOC maps on model averaging

The IGCS map was generated using the entire IGCS dataset (about 30,000 soil profiles), which is very large and hence is an example of a case study in a data-rich country (1 profile per 18 km²). To assess the usefulness of model averaging in data-poor situations, we applied model averaging to a case in which the national SOC map (IGCS) was generated from a much smaller number of soil profiles. To do so, we generated IGCS SOC

maps by randomly selecting 10,000, 5,000, 1,000, 800, 600, 400, and 200 soil profiles from the whole IGCS dataset. To filter out random sampling effects, we repeated this procedure 100 times for each sample size and reported the average results. These IGCS SOC maps with LUCAS and SoilGrids were finally merged only with the best model averaging approach and using the minimum necessary number of calibration sites as previously estimated. We also tested model averaging using only SoilGrids and LUCAS to test the assumption that no data were available to produce a national SOC map.

3.4 Results

3.4.1 Summary of IGCS, RMQS, and LUCAS datasets

Table 3.1 summarises SOC statistics of the IGCS, RMQS, and LUCAS (located in France) datasets. About 80% (24,596) of IGCS soil profiles were located in arable soils, and 20% (5,785) were located in forest and permanent grassland soils. In the IGCS soil database, grassland and forest soils (mean SOC of 24.88 g kg^{-1}) had higher SOC values than arable soils (mean SOC of 16.66 g kg^{-1}). Nearly half (985) of the RMQS sampling sites were located in permanent grasslands or forest soils, and the remaining half (1011) were under arable soils. In the RMQS dataset, the mean SOC was 18.19 g kg^{-1} for arable soils and 35.51 g kg^{-1} for permanent grassland and forest soils. LUCAS observations had a mean SOC of 26.20 g kg^{-1} for permanent grassland and arable soils.

Table 3.1 Summary statistics of SOC content (g kg^{-1}) in topsoil (0-20 cm) for IGCS, RMQS and LUCAS datasets.

Dataset	Land use*	N	Min.	Q1	Median	Mean	Q3	Max.	Sk.	SD
IGCS	F & G	5,785	0.39	12.75	19.86	24.88	30.83	373.00	3.42	20.97
	A	24,596	0.09	9.70	13.68	16.66	19.75	354.05	4.92	12.88
RMQS	F & G	985	3.78	18.86	28.37	35.51	44.00	266.60	2.81	26.01
	A	1,011	2.58	11.10	15.40	18.19	22.30	133.00	3.01	11.16
LUCAS	A & G	2,950	1.00	13.20	19.99	26.20	31.30	472.10	6.11	23.93

N, dataset size ; Min., minimum; Q1, first quantile; Q3, third quantile; Max., maximum; Sk., skewness; SD, standard deviation. * F, forest; G, permanent grasslands; A, arable.

3.4.2 Evaluation of SOC maps from IGCS, LUCAS, and SoilGrids datasets

The IGCS SOC map has the lowest RMSE (18.86 g kg^{-1}) and highest AVE (0.25) among

the three SOC maps (Figure 3.5). The small negative ME (-6.17 g kg^{-1}) indicates that SOC is underestimated in the IGCS SOC map. When the performance of the IGCS SOC map for arable and forest/grassland soils was separately evaluated, arable soils (AVE of 0.19 and RMSE of 10.02 g kg^{-1}) were found to have higher accuracy than forest/grassland soils (AVE of 0.09 and RMSE of 24.85 g kg^{-1}). SOC maps of LUCAS and SoilGrids have a much higher RMSE of 30.62 and 32.75 g kg^{-1}, and a negative AVE of -1.18 and -1.27, respectively. Positive ME of LUCAS (6.73 g kg^{-1}) and SoilGrids (21.81 g kg^{-1}) showed that these two maps overestimated SOC. The overestimation was larger in SoilGrids than in the LUCAS SOC map.

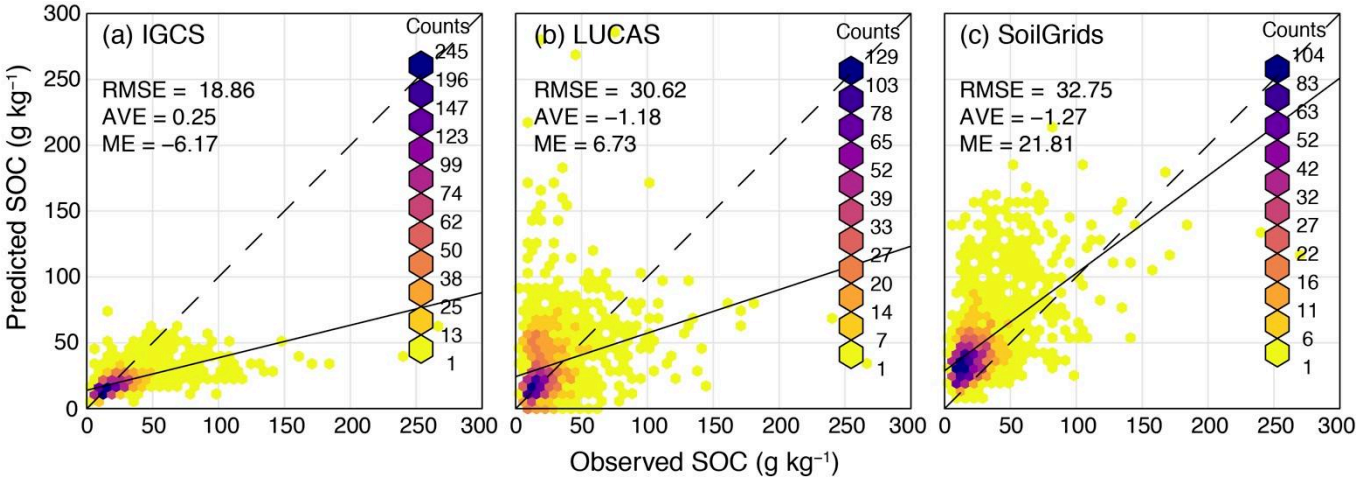


Figure 3.5 Performance of IGCS (a), LUCAS (b) and SoilGrids (c) SOC maps

3.4.3 Comparison of five model averaging approaches using different calibration sizes

The VW approach performed best among the five model averaging approaches across different calibration sizes, with the lowest RMSE (16.77-18.71 g kg^{-1}) and highest AVE (0.23-0.38) (Figure 3.6). The GR and BMA ranked second and third when the calibration size was large (100, 200 or 500), with an AVE between 0.33 and 0.38. The performance of GR substantially decreased when using a calibration sample size of 40 and 20, whereas BMA was more stable (and ranked third) when using a small calibration sample size. Cubist performed worst in the case of a large calibration sample size (100, 200, or 500) but ranked second when the calibration sample size was small (20 or 40). Residual-based Cubist did not perform well across the different calibration sample sizes. It should be

noted that VW, GR, and BMA had an ME close to 0 under different calibration sample sizes, while Cubist and Residual-based Cubist had a large negative ME.

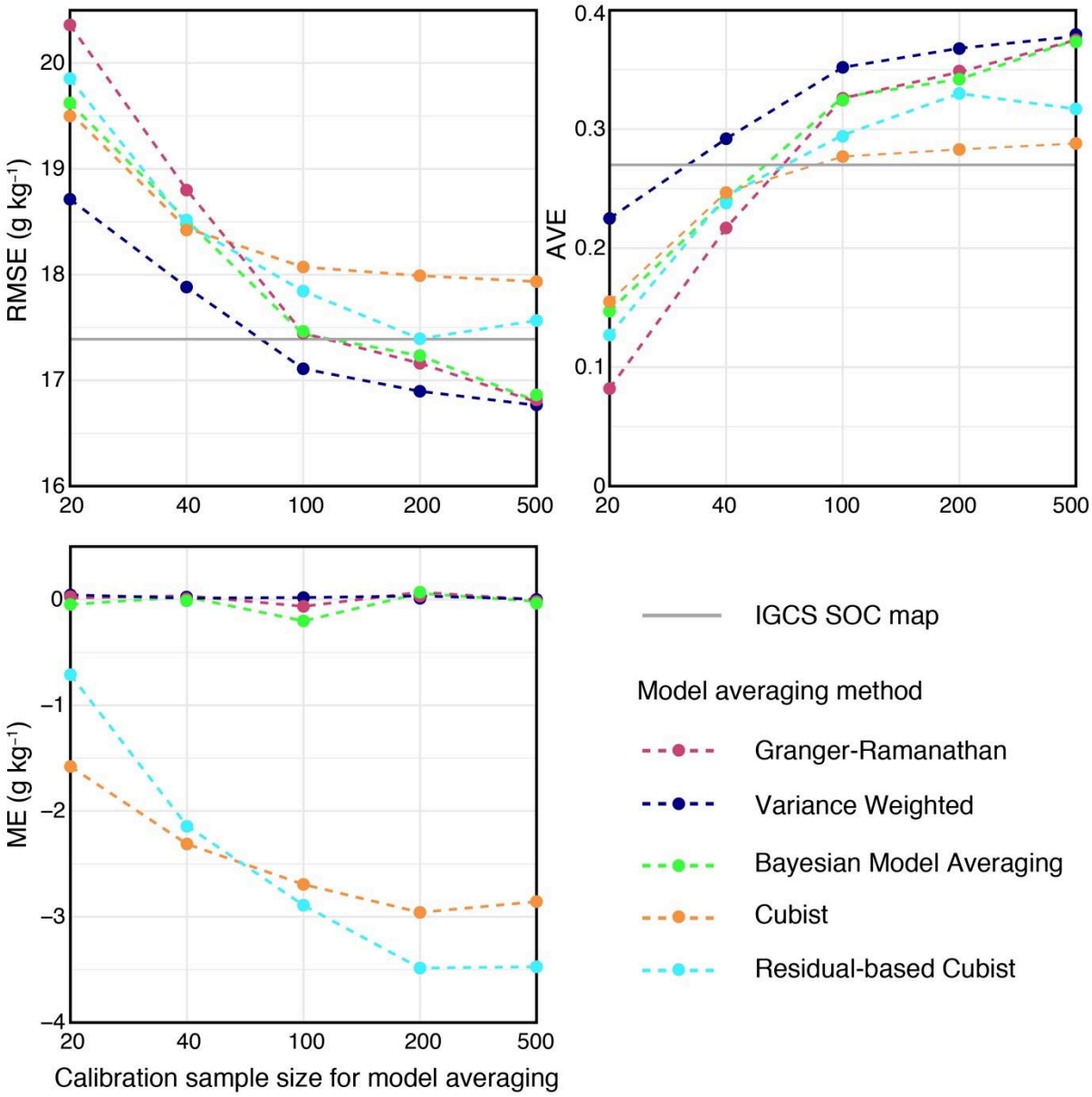


Figure 3.6 Model performance of the five model averaging approaches using different calibration sample sizes

All model averaging approaches showed better performance metrics than using the individual LUCAS and SoilGrids SOC maps for all calibration sample sizes. Improvement on the IGCS SOC map only occurred when the calibration sample size was large (100, 200, or 500), while the model averaging approaches performed worse than the IGCS SOC map when the calibration sample size was 20 or 40.

In general, the model performance of the five model averaging approaches declined

when the calibration size decreased (Figure 3.6). Being the best model averaging approach, VW had better performance than the IGCS SOC map when calibration samples were 500, 200, and 100, and it was still slightly better when only 40 calibration samples were used. However, 20 calibration samples were not sufficient to improve SOC maps using any of the five model averaging approaches. GR and BMA could improve SOC predictions when calibration sample sizes were 500, 200, and 100. However, Cubist and Residual-based Cubist only performed better than the IGCS SOC map when using a calibration sample size of 200 or more.

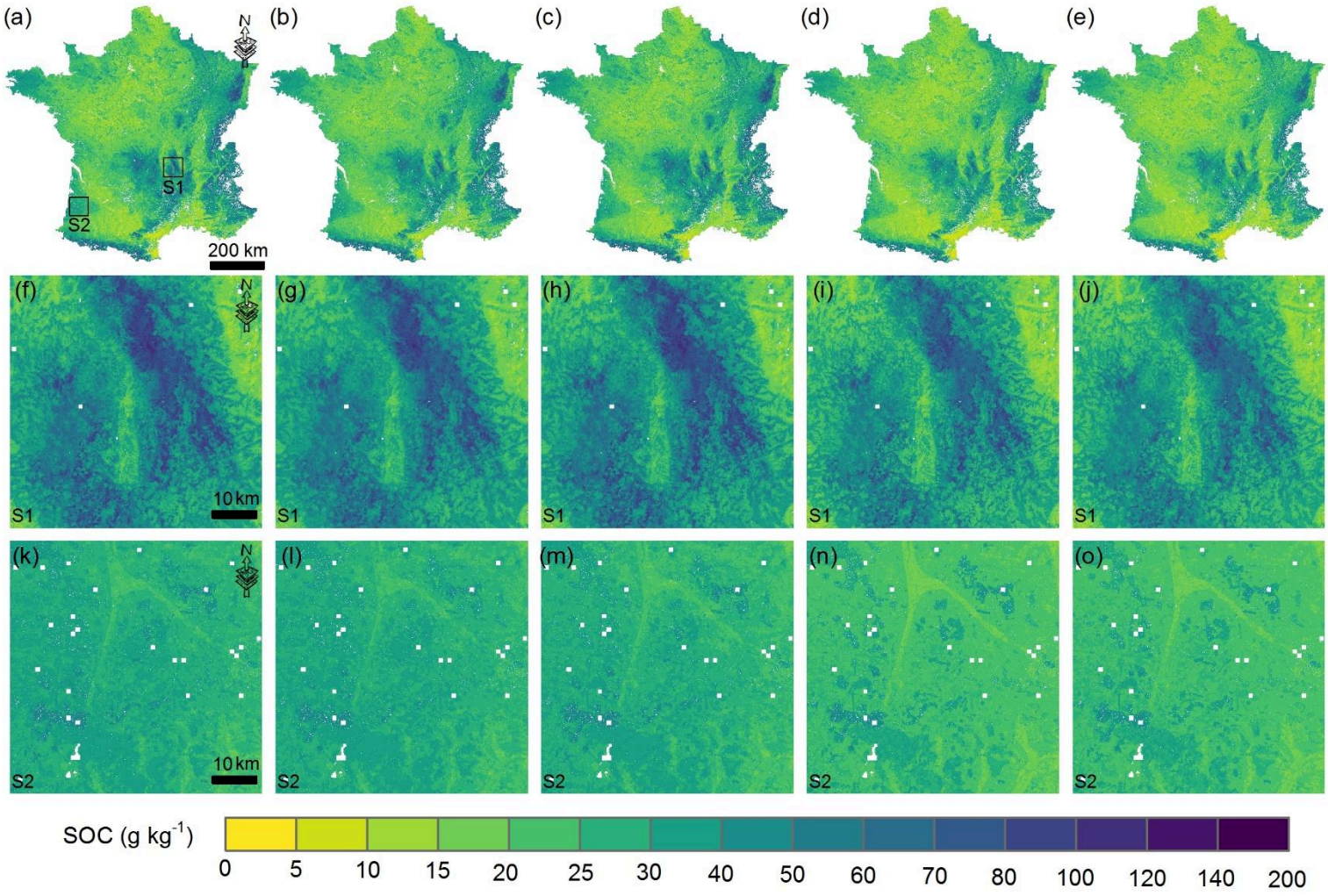


Figure 3.7 SOC maps obtained from the Granger-Ramanathan (a), Variance Weighted (b), Bayesian Model averaging (c), Cubist (d) and Residual-based Cubist (e) model averaging approaches, using all RMQS data for calibration. Local comparisons in areas S1 (f, g, h, I and j) and S2 (k, l, m, n and o) are also shown for all five model averaging approaches

As shown in Figure 3.6, only slight differences (AVE of 0.37-0.38, and RMSE of 16.77-16.90 g kg^{-1}) were observed between 500 and 200 calibration sample sizes when using VW, which was the best model averaging approach. Nevertheless, the model performance of VW showed a steady decline when the calibration sample size decreased from 200 to 20.

3.4.4 SOC maps using five model averaging approaches

Figure 3.7 shows SOC maps obtained from the five model averaging approaches using all RMQS data for calibration. The general spatial patterns of these five SOC maps were quite close, which is consistent with their similar model performance (in the case of a 500-calibration sample size) in Figure 3.6. In comparison with the IGCS SOC map (Figure 3.2a), these five SOC maps have higher SOC in mountainous regions (e.g., the Alps, the Central Massif, the Pyrenees), forests, and grasslands (e.g., the Landes of Gascony, western Brittany). As shown in Figure 3.7f to Figure 3.7o, SOC maps derived from GR, VW, and BMA had slightly higher SOC contents than Cubist and Residual-based Cubist. This is particularly visible in Figure 3.7k to Figure 3.7o, which zooms in on a square area in the Landes of Gascony forest.

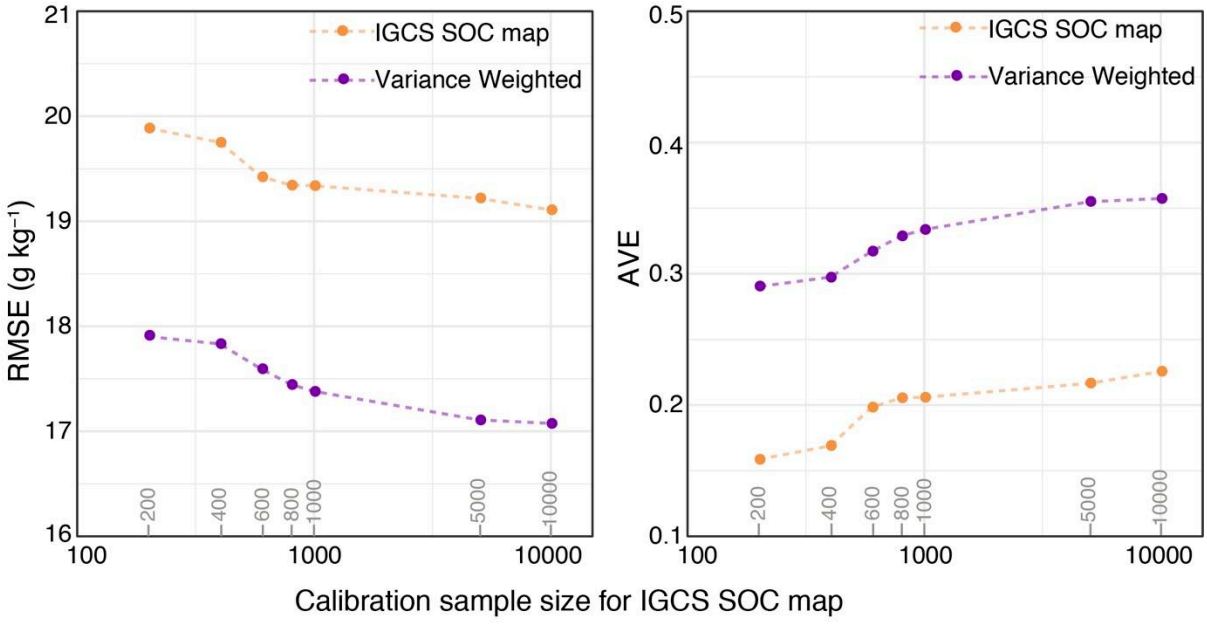


Figure 3.8 Model performance of the Variance Weighted model averaging approach when using different calibration sample sizes (200, 400, 600, 800, 1000, 5000 and 10000) for model averaging. Using only the LUCAS and SoilGrids SOC maps leads to an RMSE of 23.65 g kg^{-1} and AVE of -0.24 (points not shown). The x-axis is on log10 scale

3.4.5 Influence of national SOC maps on model averaging performance

The performance (AVE and RMSE) of the IGCS SOC maps derived from different sample sizes showed a slight decline when the number of soil profiles used decreased from 10,000 to 800 (Figure 3.8). A stronger decline in performance was observed when the number of soil profiles decreased further from 800 to 200, with AVE values dropping from 0.23 to 0.16 and RMSE increasing from 19.11g kg^{-1} to 19.89 g kg^{-1}. The performance of the VW approach showed similar declining trends as the IGCS SOC maps. However, the VW maps always performed better than the IGCS maps (ΔAVE > 0.1 and ΔRMSE < -2 g kg^{-1}). When using only LUCAS and SoilGrids for model averaging, VW performed much worse than all other SOC maps produced using IGCS, LUCAS, and SoilGrids in model averaging, with a negative AVE of -0.24 and a large RMSE of 23.65 g kg^{-1}.

3.5 Discussion

3.5.1 Performance evaluation of SOC maps from IGCS, LUCAS, and SoilGrids

The IGCS SOC map had the best performance indicators among the three source SOC products. However, it showed a slight overall underestimation and a clear tendency to underestimate large SOC values. This may be because the calibration data for generating the IGCS SOC map are dominated by cultivated soils (80% of IGCS dataset), which typically have low SOC values because of management practices (Table 3.1). As natural soils occupy 45% of the total area of mainland France (Chen et al., 2018), high SOC values are under-represented in the dataset for producing the IGCS SOC map. It consequently resulted in underestimating the effect of some controlling factors driving high SOC values (e.g., forest or grassland land uses, high elevations). Although the effects of land use and elevation are still clearly visible (Figure 3.2a), the spatial patterns of the resulting map are too smooth, as was already described by Mulder et al. (2016a; 2016b). In the French *GlobalSoilMap* product, Mulder et al. (2016a) produced national SOC maps at the first three depth intervals (0-5, 5-15, and 15-30 cm) using both IGCS and RMQS data. The AVE evaluated using 10-fold cross-validation ranged from 0.26 to 0.36 for the first three depth intervals. This shows that including RMQS data into national SOC modelling improves model performance. Nevertheless, SOC was still slightly underestimated because the IGCS dataset is almost 15 times larger than the RMQS dataset and IGCS data generally have low SOC content (Table 3.1).

The predictive performance of the LUCAS map and SoilGrids map was much worse than that of the IGCS map, as illustrated in Figure 3.2. They both have a tendency to overestimate SOC, either slightly (LUCAS) or largely (SoilGrids). The LUCAS map also exhibited more contrasted and irregular patterns than the IGCS map. Moreover, the LUCAS map showed some areas with artificially rounded boundaries (mainly in southwest France), suggesting a bias linked to the environmental covariates, predictive model, and/or interpolation method used. The SoilGrids map clearly overestimated SOC for the large majority of situations (Figure 3.5). It also clearly missed the effect of some land use types on decreasing SOC (e.g., intensively cultivated plains in northern and southwestern parts of France, vineyards in southern France). This suggests that the covariates used for global modelling could not capture these effects; e.g., land use/land cover classes used as covariates for SoilGrids were limited to cultivated land, forests, grasslands, shrublands, wetlands, tundra, artificial surfaces, and bare land cover.

Homogenising data to a common depth of 0-20 cm may have induced some additional uncertainty (Laborczi et al., 2018). We also acknowledge that resampling SoilGrids and LUCAS to 90 m resolution may have added a source of discretionality and potential uncertainty.

3.5.2 Potential and limitations of model averaging approaches

Our results demonstrate the ability of model averaging approaches to improve national SOC maps (Figure 3.5 and Figure 3.6). The improvement strongly depends on the calibration sample size used for model averaging. It is encouraging that 200 spatially stratified samples (1 sample per 2,500 km^2) were enough for producing a sufficiently accurate national SOC map (AVE of 0.37 for VW approach) when applying model averaging in France. Note also that the performance of this SOC map is comparable to that of the *GlobalSoilMap* SOC map using IGCS and RMQS datasets (Mulder et al., 2016a).

We should note that we did not map the uncertainty of SOC predictions when applying model averaging. Prediction uncertainty should be considered in future studies because it is crucial for assessing model quality and robustness and it is also a suggested product outcome, as indicated in the *GlobalSoilMap* specifications.

In addition to deriving SOC predictions using model averaging, it would be beneficial to also explicitly quantify the uncertainties associated with these

predictions. This can be done using uncertainty propagation techniques such as the Taylor series method and Monte Carlo simulation (Heuvelink, 2018; Román Dobarco et al., 2019a) provided that the uncertainties of the input maps and their correlations are quantified. This may be a useful extension of the work presented here. If it is done, it would be useful to also evaluate the validity of the uncertainty maps by computing statistics of the standardised squared prediction error (Lark, 2000) and accuracy plots (Goovaerts, 2001; Wadoux et al., 2018).

3.5.3 Comparison with previous model averaging studies

Our results suggest that map performance improves when using model averaging approaches and that the VW method is the best approach for SOC mapping in mainland France. Previous studies also showed that model averaging improves map predictions, but different approaches tend to have similar performance (e.g., Malone et al., 2014; Román Dobarco et al., 2017; Caubet et al., 2019). Caubet et al. (2019) applied two model averaging approaches (GR and VW) to improve soil texture maps (clay and sand) and showed that both model averaging approaches improved the accuracy and that GR outperformed VW. Similar results were found by Román Dobarco et al. (2017) for mapping soil texture, and Malone et al. (2014) on pH mapping. Further work could analyse the causes of these differences.

Caubet et al. (2019) also mentioned the potential use of non-linear models for improving model averaging. However, in our study, non-linear models like Cubist and Residual-based Cubist did not perform better than a linear model like GR. Perhaps this is because three SOC products are not sufficient for calibrating a regression tree or machine learning approach, and that other additional covariates (e.g., elevation, land use, and climatic variables) may be helpful to improve model performance. Especially, the example of the Landes of Gascony (see Figure 3.7k to Figure 3.7o) shows that the model does not capture the effect of forest land use well in many areas when using a rule-based model such as Cubist.

Caubet et al. (2019) found that around 200 to 300 calibration samples were sufficient for model averaging of soil texture over mainland France. This result is consistent with our finding that 200 calibration samples (1 sample per 2,500 km^2 for a total area of 550,000 km^2 and a country having a high pedodiversity (Minasny et al., 2010)) selected

from equal-size clustering are enough to improve existing SOC maps using model averaging.

3.5.4 Contribution of model averaging approaches to data-poor countries

We tested model averaging on a situation that may be considered "rich" concerning the amount of available data (Arrouays et al., 2017). In this study, we used 30,000 samples for national SOC mapping, which is 1 sample per 18 km². Although France has numerous point soil data, these data are rather clustered and irregularly cover the territory. They also over-represent some agro-pedo-climatic conditions (e.g., low elevations and intensively cultivated areas). These conditions (irregularity and non-representativeness of samples) are likely to be similar in most data-rich countries that use legacy data for DSM. Our results suggest that merging national predictions with continental and global predictions that capture some trends may help to counterbalance the effects of a national unbalanced sampling design.

The fact that the number of samples needed to calibrate the averaging model is rather low is encouraging, i.e. 200 samples for mainland France. This is cost-effective given the limited effort required to gather a fairly small number of soil samples to improve national soil maps.

The results shown in Figure 3.8 indicate that model averaging always has a substantial added value in terms of model performance compared to using the IGCS SOC map alone. Moreover, the added value of model averaging is larger than that of only increasing the number of profiles used for producing the IGCS SOC map. For example, using 200 samples for model averaging calibration results in an AVE increase of 0.12, whereas the AVE only increases by 0.07 when the number of profiles used for producing the IGCS SOC map increases from 200 to 10,000. This indicates that adding a relatively small regular grid of soil samples to merge several maps might be more efficient than expanding the database with a large number of soil samples for which the sample locations are not controlled. In many countries, soil mapping activities are frequently guided by local needs and interests. This explains why national soil datasets are often clustered and why adding more legacy data may sometimes lead to increasing sources of bias (e.g., Poggio et al., 2019). Overall, our study advocates merging predictions in both data-rich and data-poor situations and demonstrates that the added value of merging is

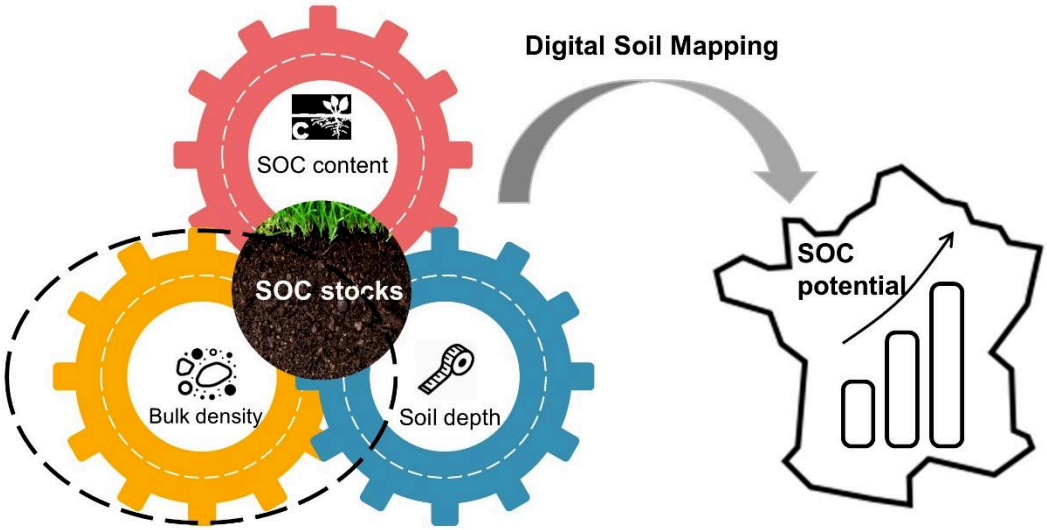
relatively higher in data-poor situations. However, notably, the performance of VW drops substantially when excluding the IGCS SOC map and when it only uses LUCAS and SoilGrids for model averaging. This indicates the importance of a national SOC map in model averaging, even if this SOC map is produced with a small dataset (i.e. 200 samples).

3.6 Conclusions

We tested the ability of five model averaging approaches for improving existing SOC maps by merging national, continental, and global SOC products. All five model averaging approaches could improve the national SOC map when more than 100 soil samples were used for calibration of the model averaging approaches. The VW approach performed better than the other four approaches. Model averaging approaches using a rather small calibration dataset (i.e. 200 observations uniformly spread over mainland France) for calibration proved to be efficient. The national SOC map was very important and drove performance when merging all SOC maps. By reducing the number of national soil samples in France for producing the national SOC map, we found that merging maps using model averaging is also applicable to data-poor situations and might thus be attractive to data-poor countries, provided sufficient soil data are available for calibration of the model averaging approach.

Chapter 4

Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over larger area



Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D., 2018. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma, 312, 52-63.

4.1 Introduction

It is well established that soil bulk density (BD) is an important property related to soil moisture availability, hydraulic conductivity, plant growth and crop yield (Dam et al., 2005). During weight-to-volume conversion for soil water, soil organic carbon (SOC), nutrients or trace elements (TE), BD measurements are often used for calculating fluxes and stocks (Poeplau et al., 2017; Martin et al., 2009; Lacarce et al., 2012). Nevertheless, for the calculation of SOC stocks alone, Poeplau et al. (2017) recently stressed that, in order not to overestimate SOC stocks, the fine soil (soil particles < 2 mm) stock of the investigated soil layer should be taken into account (see for instance Martin et al., 2011, 2014). However, bulk density might be of interest as an important soil property by itself or for calculating stocks of other elements, such as TE (Lacarce et al., 2012). Also, if bulk density and the content of coarse elements (soil particles > 2 mm) are known, the fine soil mass for a given depth can be calculated. Given its importance, BD is one of the suggested soil properties that need to be mapped at global scale according to the specifications of *GlobalSoilMap* project (Arrouays et al., 2014a).

Despite the importance of BD, it is usually lacking in soil database worldwide. This is mainly owing to the fact that determination of BD is usually time consuming and labor intensive. In addition, direct spatial modelling of BD is difficult due to a complex combination of controlling factors. In order to overcome aforementioned challenges, various pedotransfer functions (PTFs) have been applied to predict BD using easily measured and available information. Most predictive models about BD usually integrate soil chemical and physical properties including organic carbon, soil texture, sampling depth or horizon designation, and coarse element content (Jeffrey, 1970; Adams, 1973; Federer, 1983; Manrique and Jones, 1991; Tomasella and Hodnett, 1998; Bernoux et al., 1998; Kaur et al., 2002; Heuscher et al., 2005; Sequeira et al., 2014). Other less frequently used parameters include pH (Benites et al., 2007; Libohova et al., 2014; Botula et al., 2015), cation exchange capacity (Botula et al., 2015; De Souza et al., 2016), water content (Heuscher et al., 2005; Keller and Håkansson et al., 2010) and sum of exchangeable bases (Benites et al., 2007; De Souza et al., 2016). Environmental information including vegetation, topography, temperature and rainfall is sometimes added to the predictors set (Martin et al., 2009; Jalabert et al., 2010; Sequeira et al., 2014; Akpa et al., 2016; De

Souza et al., 2016). A generic issue concerning the use of PTFs is the assessment of its validity domain (Minasny et al., 1999).

In a recent study representing a variety of predictive models about BD, Nanko et al. (2014) presented the development of PTFs from physical PTFs to empirical PTFs. They compared 29 existing PTFs belonging to six groups including physical equation, radical root equation, logarithmic equation, exponential equation, decimal equation and polynomial equation, and developed revised PTFs for their database. They found that it was worthwhile to revise PTFs for the reason that the relationship between SOC and BD changed in different regions.

The final aim of PTFs is to apply these models on soil samples without BD, so in this case, the potential utility of PTFs should be taken into consideration. To avoid non-valid extrapolation, Tranter et al. (2009) explored the potential of distance metrics to identify the domain of PTFs predictions. First, they used three distance methods to determine the distance from the mean values of the calibration data set and the soil samples for which BD had to be predicted. Samples with distances exceeding a designated cutoff limit were defined distinct from the calibration data and thus were not suitable for the use of PTFs for BD predictions. Their results demonstrated that the proposed protocol was useful in excluding those samples dissimilar to the calibration data set.

The objectives of this study were four-fold: 1) Build an empirical PTFs for Region Centre of France; 2) Test PTFs' predictive ability over a much larger territory (mainland France except Region Centre); 3) Apply distance metrics on external validation data to determine the validity domain of PTFs; 4) Test distance cutoff criteria to optimize additional sampling scheme.

4.2 Materials and methods

4.2.1 Location and sampling

In this study, we used available BD data from the French Soil Inventory Program (IGCS) and French Soil Monitoring Network (RMQS) (Laroche et al., 2014; Arrouays et al., 2014a). The IGCS dataset complied data from many studies and did not arise from a single systematic sampling scheme for France. Therefore, soil data in some areas were rich while in some areas were quite sparse (Figure 4.1). Besides, the sampling depth in IGCS dataset ranged widely from 0 to 295 cm. The RMQS network is based on a 16 km ×

16 km square grid and the sites are selected at the center of each grid cell. In the case of soil being inaccessible at the center of the cell, an alternative close location with a natural soil was selected. Detailed information including land use and profile description was recorded for each site. Different from IGCS dataset, each site in RMQS dataset was sampled from topsoil (0-30 cm or less) and subsoil (30-50 cm). Finally, a total of 7090 soil samples (3,750 from IGCS and 3,340 from RMQS) were used and all these samples had BD, soil organic carbon (SOC), pH, soil clay, silt, sand and gravel content measured. Bulk density was measured using the cylinder method (AFNOR, 1992) except for stony soils where an excavation method was preferred. Coarse elements were determined by wet sieving through a 2 mm mesh. Fine earth was defined as particles with a diameter smaller than 2 mm. pH was measured in a 1:5 soil:water mixture (AFNOR, 1994) and particle-size analysis was performed with the pipette method (AFNOR, 2003). Organic carbon concentrations of the fine earth were mostly measured by the dry combustion method using an automated C:N analyzer (5,234 samples). Soils sampled before 1990 (1,856 samples) were analyzed by the wet-combustion method. No attempt was made to harmonize results from both methods, as no correction factor was available for French soils except for sandy Spodosols (Jolivet et al., 1998). We calculated the depth by determining the mid-point of upper boundary and lower boundary in each sampled layer or horizon.

The data used for establishing PTFs model was from Region Centre of France which is located in the Middle Loire basin and covers 34,151 km² (Figure 4.1). Region Centre has a continental oceanic climate with a mean annual temperature around 11.4 °C and a mean annual rainfall below 800 mm. According to World Reference Base for Soil Resources (IUSS Working Group WRB, 2006), the main soil types in Region Centre are Luvisols, Cambisols, Leptosols, Fluvisols and Podzols (Ciampalini et al., 2014). As shown in Figure 4.1, soil data in Region Centre (1,357 samples) were relatively denser than in other regions, especially for IGCS dataset (1,096 samples). As suggested by Tranter et al. (2009), robust estimates of mean and standard deviation for independent variables in the calibration data set should be reported when developing PTFs. Mean and standard deviation for soil properties in Region Centre and other regions are summarized in Table 4.1.

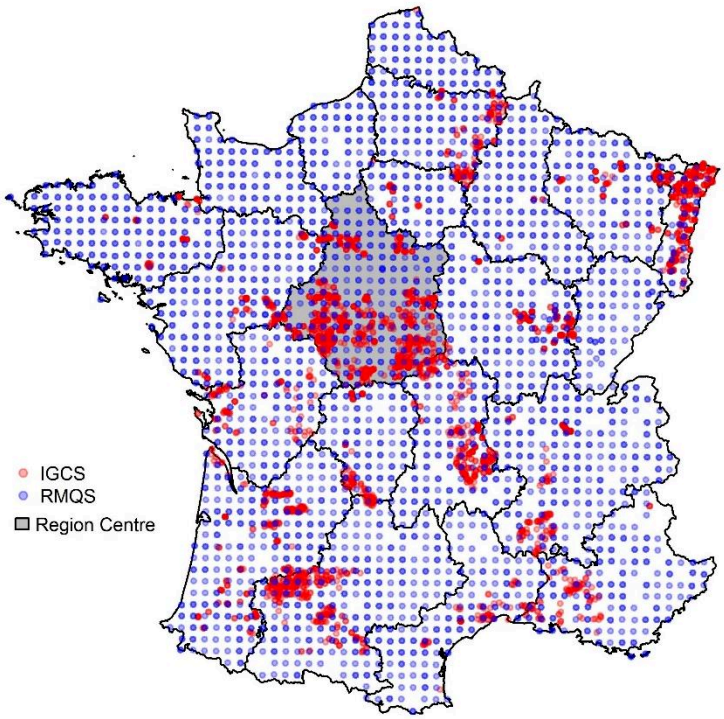


Figure 4.1 Distribution of the samples in calibration and validation data

Considering the effects from the differences (spatial distribution and soil depth) of sampling strategy between IGCS and RMQS datasets, several combinations of calibration (data from Region Centre) and validation (data from other regions) subsets were selected (Table 4.2).

4.2.2 GBM modelling

GBM algorithm (Ridgeway, 2012) was used to build PTFs for the prediction of BD. The objective of GBM is to solve predictive learning problem on estimating a function that projects a set of covariates into an output variable by minimizing a specified loss function L (Martin et al., 2009). At each iteration of GBM, the specified function L is used to fit base learners. The prediction results from each iteration are combined to get the final prediction by base learner associated weights which is called the learning rates or shrinkage parameters. GBM algorithm is a specified form of stochastic gradient boosting (Friedman, 2001) for it uses regression trees (Breiman et al., 1984) as base learners. On a

Table 4.1 Robust estimates of means and standard deviation for soil data from Region Centre and other regions

	Region Centre				Other regions			
	Whole/IGCS/RMQS				Whole/IGCS/RMQS			
	Mean	Minimum	Maximum	S.D.	Mean	Min	Max	S.D.
BD[a]	1.48/1.49/1.43	0.61/0.61/0.80	2.31/1.97/2.31	0.17/0.17/0.19	1.40/1.45/1.36	0.32/0.32/0.37	2.50/2.50/2.25	0.23/0.22/0.23
SOC[b]	0.75/0.65/1.16	0.02/0.02/0.12	8.64/8.64/6.41	0.75/0.68/0.88	1.58/1.21/1.90	0.01/0.01/0.06	26.60/26.60/26.50	1.71/1.56/1.76
pH[c]	6.78/6.84/6.55	3.40/3.40/3.70	9.06/9.06/8.64	1.04/0.96/1.31	6.77/6.94/6.63	3.70/3.70/3.70	10.00/10.00/9.20	1.24/1.19/1.27
Clay[d]	27.35/28.30/23.34	0.90/0.90/3.50	96.40/96.40/81.90	15.64/16.00/13.26	25.05/24.87/25.21	0.10/0.10/0.20	87.80/87.80/85.10	13.79/14.49/13.15
Silt[e]	39.09/38.70/40.70	0.60/0.60/4.60	85.50/85.50/79.30	19.29/18.94/20.47	41.48/41.53/41.45	0.20/0.30/0.20	87.45/87.45/81.90	17.56/17.97/17.20
Sand[f]	33.56/33.00/35.96	0.90/0.90/2.05	96.46/96.46/91.55	23.71/23.25/25.26	33.47/33.60/33.34	0.20/0.21/0.20	98.60/98.20/98.60	23.12/23.71/22.61
Depth[g]	37.92/41.14/28.49	0.00/0.00/5.00	200/200/50	26.93/28.49/12.51	29.77/36.64/31.85	0.00/1.00/0.00	280/280/60	24.31/31.85/12.31
Gravel[h]	5.52/3.96/6.17	0.00/0.00/0.00	71.10/35.00/71.10	9.05/5.90/14.60	9.81/3.98/14.84	0.00/0.00/1.40	83.10/37.00/83.10	14.38//7.19/16.90

[a] Bulk density (g cm^{-3}); [b] Soil organic carbon (%); [c] pH; [d] Clay (<2 µm, %); [e] Silt (2-50 µm, %); [f] Sand (50-2000 µm, %); [g] Depth (mid-point of the soil horizon, cm); [h] Gravel (>2000 µm, %).

Table 4.2 Combinations of calibration and validation subsets

Combination	Calibration[a]	Validation[b]	Combination	Calibration[a]	Validation[b]	Combination	Calibration[a]	Validation[b]
C1	IGCS+RMQS	IGCS+RMQS	C10	RMQS	$IGCS_{>50}$	C19	IGCS0-50	$IGCS_{0-50}$
C2	IGCS+RMQS	RMQS	C11	IGCS	IGCS+RMQS	C20	IGCS0-50	$IGCS_{>50}$
C3	IGCS+RMQS	IGCS	C12	IGCS	RMQS	C21	IGCS>50	IGCS+RMQS
C4	IGCS+RMQS	$IGCS_{0-50}$[c]	C13	IGCS	IGCS	C22	IGCS>50	RMQS
C5	IGCS+RMQS	$IGCS_{>50}$[d]	C14	IGCS	IGCS0-50	C23	IGCS>50	IGCS
C6	RMQS	IGCS+RMQS	C15	IGCS	IGCS>50	C24	IGCS>50	$IGCS_{0-50}$
C7	RMQS	RMQS	C16	IGCS0-50	IGCS+RMQS	C25	IGCS>50	$IGCS_{>50}$
C8	RMQS	IGCS	C17	IGCS0-50	RMQS			
C9	RMQS	$IGCS_{0-50}$	C18	IGCS0-50	IGCS			

[a] Calibration subset was from Region Centre; [b] Validation subset was from other regions of France; [c] Samples with a maximum lower depth between 0 and 50 cm in IGCS dataset; [d] Samples with a maximum lower depth larger than 50 cm in IGCS dataset.

given iteration or tree, only a subset of the dataset, named bag fraction is randomly selected without replacement for fitting the base learner. In addition to learning rate and number of trees, other two parameters are also important in GBM models. These two parameters mainly control the details in base learner and are: (i) tree size (or known as the maximum depth of variable interactions) and (ii) minimum number of observations in the terminal nodes of the trees. There are several options for parameter tuning in GBM models, and as the most efficient one that suggested by Ridgeway (2012), an internal 10-fold cross-validation was used in this study. These optimal parameters were used in order to avoid over-fitting problem.

Contribution of each covariate in GBM can be determined by computing a relative variable importance, which averages the relative contribution of each covariate across all the individual trees (Friedman and Meulman, 2003). For a covariate *j*, the index *I* is computed as

$$I_j^2 = \frac{1}{M}\sum_{m=1}^{M} I_j^2(T_m) \quad (4.1)$$

where *M* is number of trees in the GBM and I_j is the relative influence of the covariate *j* for the individual trees T_m.

The GBM models were fitted by Region Centre data and validated by other regions data using *gbm* function in the *caret* R package (Ridgeway, 2012; Kuhn, 2008).

4.2.3 Calibration and validation procedures

In this study, common covariates in previous PTFs, including SOC, pH, clay, silt, sand, depth and gravel, were used for the modelling.

Two tasks were accomplished in this section. Task one was model comparison between GBM and PTFs belonging to six groups described by Nanko et al. (2014). As shown in Table 4.3, six groups of PTFs were revised or refitted by SOC or soil organic matter (calculated from SOC multiplied by 1.724) from our calibration data using the Levenberg-Marquardt non-linear least-square method available in the *minpack.lm* R package (Elzhov et al., 2013). In this part, calibration data from whole Region Centre data including IGCS and RMQS while validation data was whole other regions dataset. Task two was a model comparison between combinations of calibration and validation subsets which were specified in details in section 4.2.1 (Table 4.2) and GBM algorithm was applied

Table 4.3 Summary of PTFs defined in previous research

Model*	Function	Refitted coefficients		
		a	*b*	*c*
A	$\rho=a+b\sqrt{SOC}$	1.658	-0.228	
B	$\rho=a+b\log_{10}(SOC)$	1.419	-0.197	
C	$\rho=a+be^{c(SOC)}$	0.721	0.855	-0.172
D	$\rho=\frac{1}{a+b(SOC)}$	0.635	0.059	
E	$\ln(\rho)=a+b\ln(SOC)+c\{\ln(SOC)\}^2$	0.347	-0.092	-0.021
F	$\rho=\frac{100}{\frac{SOM}{a}+\frac{100-SOM}{b}}$	-5.259	1.724	

* ρ, bulk density (g cm^{-3}); *SOM*, soil organic matter (%); *SOC*, soil organic carbon (%).

Table 4.4 Accuracy comparison among different combinations of calibration and validation subsets using GBM

	Calibration				Validation					Calibration				Validation			
	MPE	SDPE	RMSPE	R^2	MPE	SDPE	RMSPE	R^2		MPE	SDPE	RMSPE	R^2	MPE	SDPE	RMSPE	R^2
C1					0.005	0.179	0.179	0.529	C16					-0.001	0.192	0.192	0.399
C2					0.018	0.165	0.165	0.608	C17					-0.005	0.189	0.189	0.436
C3	-0.001	0.132	0.132	0.446	-0.011	0.192	0.192	0.446	C18	0.001	0.132	0.131	0.392	-0.008	0.195	0.195	0.331
C4					-0.007	0.202	0.202	0.465	C19					0.012	0.203	0.203	0.359
C5					-0.014	0.181	0.181	0.341	C20					-0.029	0.184	0.186	0.301
C6					0.012	0.182	0.182	0.431	C21					-0.064	0.201	0.211	0.294
C7					0.032	0.171	0.173	0.427	C22					-0.076	0.198	0.212	0.328
C8	-0.001	0.157	0.156	0.299	-0.013	0.193	0.193	0.352	C23	0.000	0.128	0.128	0.431	-0.048	0.202	0.208	0.356
C9					-0.013	0.201	0.202	0.322	C24					-0.071	0.215	0.227	0.281
C10					-0.012	0.181	0.182	0.263	C25					-0.024	0.184	0.186	0.307
C11					0.005	0.187	0.187	0.389									
C12					0.009	0.183	0.184	0.406									
C13	-0.001	0.128	0.128	0.409	0.001	0.192	0.192	0.338									
C14					0.014	0.201	0.202	0.341									
C15					-0.013	0.181	0.181	0.299									

in each combination here.

In the calibration procedure of GBM, Ridgeway (2012) made some recommendations for various GBM parameters; however, the optimization of these parameters depends on the dataset to a large extent. Consequently, *train* function in *care*t R package was used for the optimization procedure according to the following combinations:

1) Learning rate {0.02, 0.04, 0.06, 0.08, 0.1};
2) Number of trees {200, 400, 600, 800, 1,000};
3) Tree size {2, 4, 6, 8, 10};
4) Minimum number of observations in the terminal nodes of the trees {2, 4, 6, 8, 10}.

Optimized combination of parameters was determined by best internal 10-fold cross-validation and then was applied in the validation data set. In order to reduce modelling time, *doParallel* R package (Calaway et al., 2015) was used for parallel computing.

4.2.4 Evaluation of model performance

In both calibration and validation procedures, comparison between measured and predicted values of BD was performed using a set of indices that have been commonly suggested (Martin et al., 2009; Nanko et al., 2014): the mean prediction error (MPE), standard deviation of the prediction error (SDPE), root mean square prediction error (RMSPE) and determination coefficient (R^2). These indices are defined as follows:

$$\mathrm{MPE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{\rho}_i-\rho_i) \quad (4.2)$$

$$\mathrm{SDPE}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\{(\hat{\rho}_i-\rho_i)-\mathrm{MPE}\}^2} \quad (4.3)$$

$$\mathrm{RMSPE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\rho}_i-\rho_i)^2} \quad (4.4)$$

$$R^2=\frac{\sum_{i=1}^{n}(\hat{\rho}_i-\bar{\rho})^2}{\sum_{i=1}^{n}(\rho_i-\bar{\rho})^2} \quad (4.5)$$

where n is the number of observations, $\hat{\rho}_i$ and ρ_i are the predicted and measured BD for observation i, and $\bar{\rho}$ is the mean value of measured BD. The MPE indicates the bias of regression model, while SDPE and RMSPE evaluate the random variation of the predictions after correction for global bias (Nanko et al., 2014). The R^2 determinates the portion of variability in the predicted values explained by measured BD values. A good

model has a MPE close to 0, a smaller SDPE and RMSPE values, and also a higher R^2.

4.2.5 Validity domain of GBM model

The validity domain of GBM model was identified by distance metrics described by Tranter et al. (2009). The main idea of this method is to use distance metrics to determine the distance from the arithmetic mean of calibration data and a subject of interest. Commonly three methods including Euclidean distance, Standardized Euclidean distance and Mahalanobis distance can be used for distance calculation. Given the scale sensitivity of input variables in the Euclidean distance, Tranter et al. (2009) suggested that Standardized Euclidean distance and Mahalanobis distance would be more suitable for use in PTFs due to the large scale differences of input variables. In addition, when the covariance elements of variance-covariance matrix approach to zero, the Standardized Euclidean and Mahalanobis distances would be equal. Standardized Euclidean distance was applied in calculating distance metrics in this study.

In an n-dimensional space, Standardized Euclidean distance (d) between point x (x_1, x_2, x_3...x_n) and point y (y_1, y_2, y_3...y_n) is calculated as:

$$d = \sqrt{(x\text{-}y)^T \mathrm{A} (x\text{-}y)} \quad (4.6)$$

where T the transpose of the matrix, and A is the matrix of variance elements in order to standardize each dimension and is described by the form:

$$\mathrm{A} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} \quad (4.7)$$

The squared distances of the calibration are close to a chi-squared distribution for data with a normal distribution and 97.5% percentile of the cumulative chi-squared distribution from calibration data is a commonly used cutoff limit to identify whether a point is similar or not with major part of calibration data (Filzmoser and Hron, 2008; Rousseeuw and Zomeren, 1990). But it is rare to find a rigid normal distribution in practice, so the distance cutoff limit was tuned from 90% to 100% by an interval of 0.5% and it was optimized by a balance between the number of observations and RMSPE calculated from these validation data within distance cutoff limit.

The protocol of the aforementioned method is described in the following steps (Tranter et al., 2009):

1) Determine the means and standard deviations of the calibration data;
2) Calculate Standardized Euclidean distance (d) from the arithmetic mean of calibration data to the subject of interest;
3) Tune and optimize distance cutoff limit;
4) Exclude soil samples with distances larger than the cutoff limit from the GBM model.

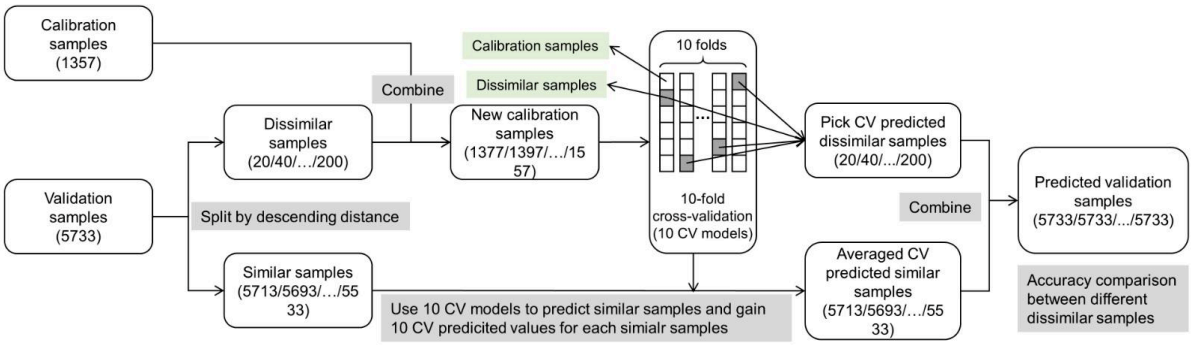


Figure 4.2 Flowchart of comparison between different numbers of additional samples

4.2.6 Additional sampling optimization

In mainland France except for Region Centre, soil samples with less similarity from calibration data were defined by different distance cutoff limits. We tested if additional sampling on these dissimilar samples made it possible to improve the predictive ability of GBM model at national scale. To take into consideration on a balance between cost of soil sampling (number of dissimilar samples) and predictive accuracy, a series of additional samples from 20 to 200 with an interval at 20 were used to assess the effect of additional sampling strategy. Figure 4.2 describes how we evaluate this strategy. Firstly, these additional samples were chosen from dissimilar samples and added into calibration data by descending distance. Then 10-fold cross-validation was performed on new calibration data. For each additional sample, a cross-validation predicted value was obtained. For each remaining similar samples in validation, 10 cross-validation predicted values were averaged into one. Finally, the performance of different additional samples was calculated on combing cross-validation predicted dissimilar samples and remaining

similar samples. There were two reasons to do so: 1) when different numbers of additional samples were excluded from validation data, the performances were not comparable; 2) the predicted BD of additional samples were from cross-validation, so it avoided over-fitting problem.

4.3 Results

4.3.1 Summary statistics about datasets

Table 4.1 lists the descriptive statistics of soil properties from two regions (Region Centre and other regions) and two datasets (IGCS and RMQS). The mean value of BD in Region Centre (1.48 g cm^{-3}) was slightly larger than that in other regions (1.40 g cm^{-3}) while standard deviation (0.17 g cm^{-3}) of BD in Region Centre was smaller. In Region Centre, mean value and standard deviation of SOC were both 0.75%, which were 1.58% and 1.71% respectively lower than in other regions of France.

There were large differences between IGCS and RMQS datasets, especially for SOC, depth and gravel. These were mainly due to different sampling schemes including sampling distribution, density and depth. A majority of data (70.06%) in IGCS dataset was sampled from arable land, which lead to the low gravel content (less than 4%) in IGCS dataset. Meanwhile, RMQS dataset covered a wide range of land use including arable land (43.14%), forest (24.86%), pasture (24.56%), natural grassland or shrubland (3.57%), and permanent crop (3.12%). Forest and pasture had relatively high mean SOC contents (2.98% and 2.12%) and accounted for nearly half of the sampling locations in RMQS dataset. As a result, mean SOC content in RMQS dataset was much higher than that in IGCS dataset.

4.3.2 Model comparison between revised PTFs and GBM model

The revised PTFs (A, B, C, D, E, and F) were developed by non-linear parameter fitting to the six groups of PTFs summarized by Nanko et al. (2014) using whole data in Region Centre. Revised coefficients for these PTFs are listed in Table 4.3. The performance of six revised PTFs on validation data and the relationship between measured and predicted BD values are shown in Figure 4.3. The R^2 values of the revised PTFs ranged from 0.122 to 0.389 while RMSPE values ranged from 0.186 to 0.278. Similar SDPE and RMSPE were found among model A, C, D and E while model C and D had a higher R^2. Consequently, Model C and D performed better than the other six revised PTFs. However, an obvious

plateau around 1.6 g cm^{-3} was found in Model C and D, which means these models always underestimated soil samples with a measured BD values larger than 1.6 g cm^{-3}.

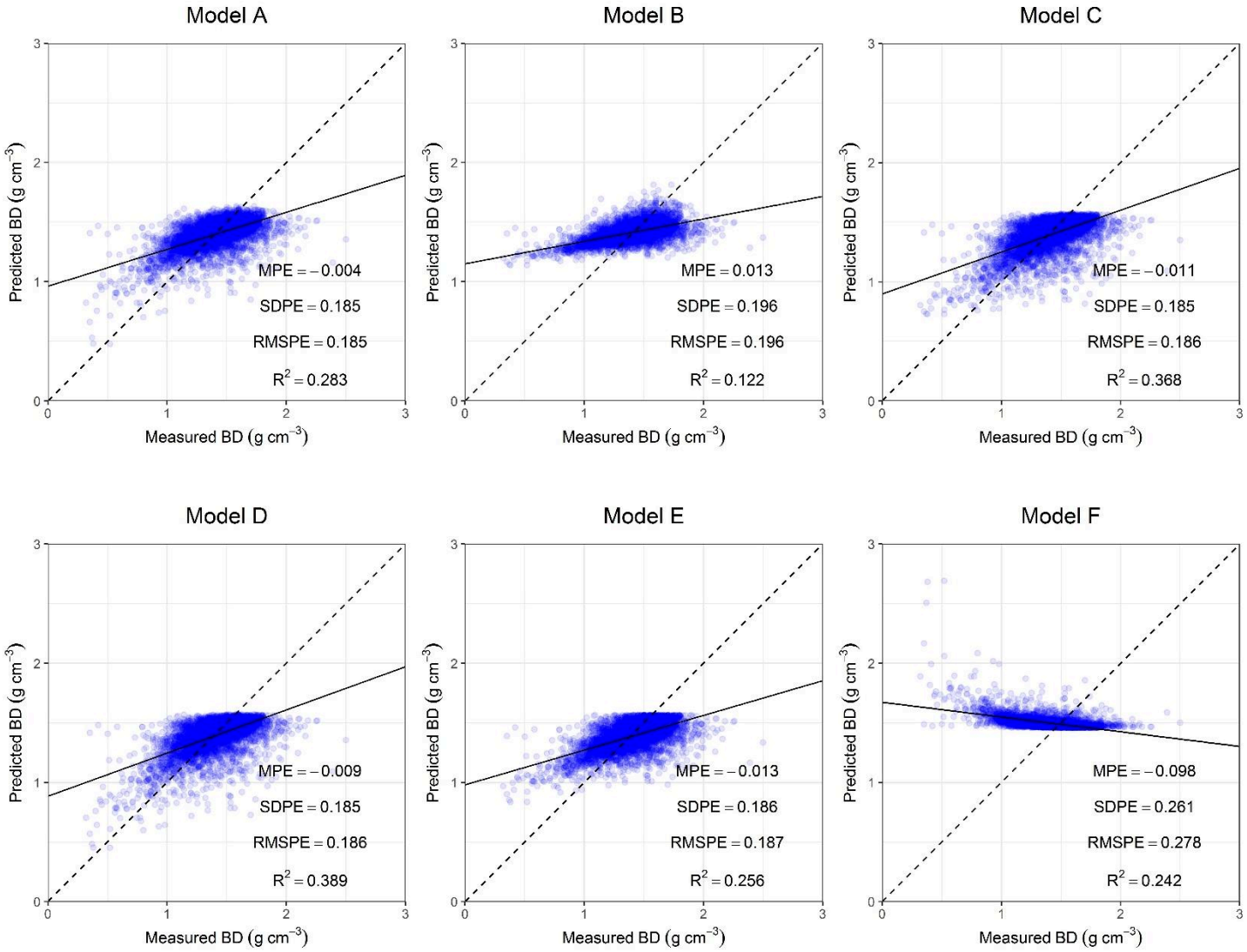


Figure 4.3 Measured vs. predicted bulk densities in six revised PTFs

Seven covariates from Region Centre data were used for GBM modelling. After 10-fold cross-validation among 625 tuned models, optimized GBM model was chosen by least RMSPE and thereafter the best GBM model was performed on other regions data for validation procedure. Figure 4.4 shows performance on calibration and validation for this best GBM model. In the calibration, GBM performed well and yielded a MPE close to 0, small SDPE and RMSPE values (both in 0.132 g cm^{-3}), and a R^2 at 0.446. Higher SDPE, RMSPE and R^2 values were gained for the validation on other regions data. SOC and clay were the top two important covariates of BD in GMB model and their importance index summed to more than 60% of contribution in modelling (Figure 4.5). Sand came the

third with 13% contribution. Gravel, pH, silt and depth did little contribution in GBM model (<10%).

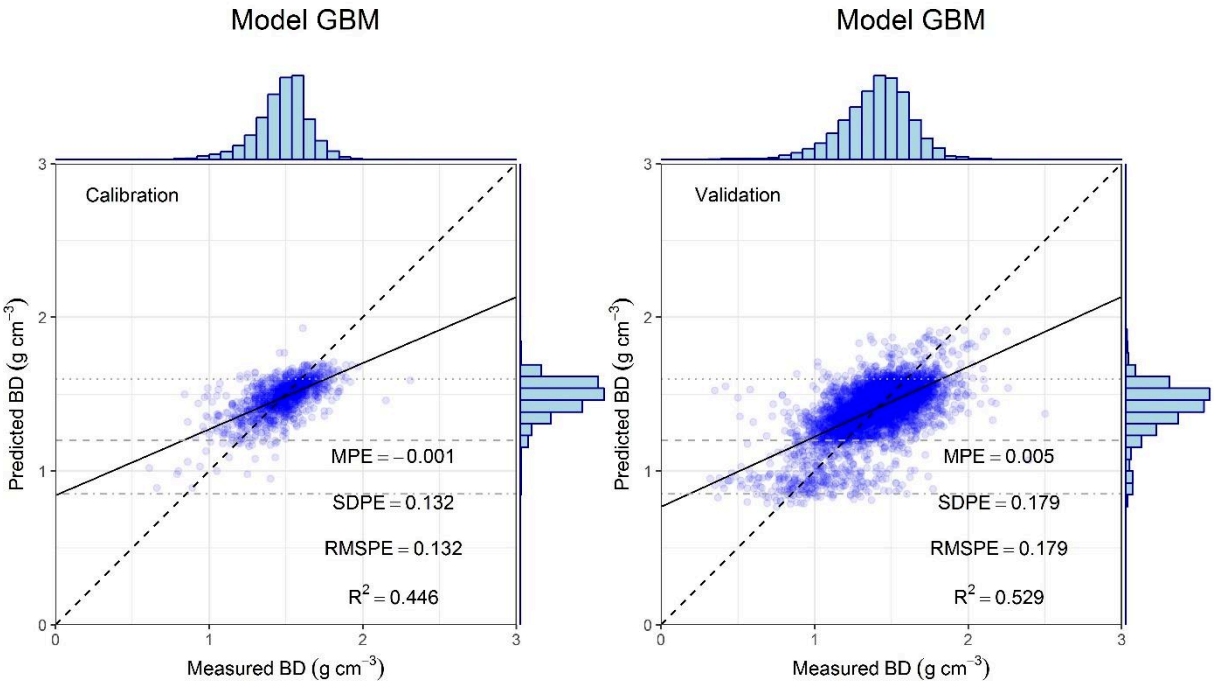


Figure 4.4 Measured vs. predicted bulk densities in calibration and validation using GBM

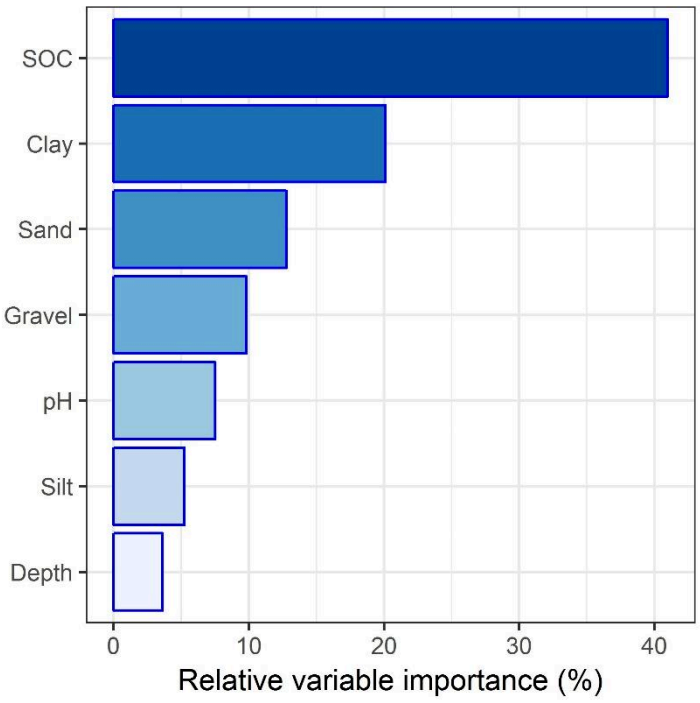


Figure 4.5 Relative variable importance for each covariate of the GBM model

There was no doubt that GBM model performed better with a smaller SDPE, RMSPE and higher R^2 than six revised PTFs. Besides, GBM model did not have the problem of a predictive threshold so it had a wider predictive range than the six revised PTFs.

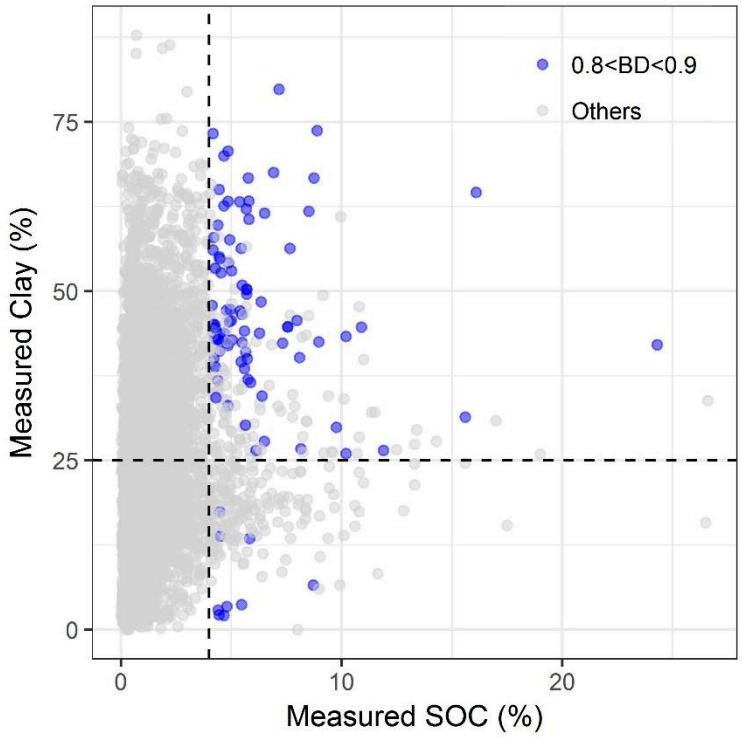


Figure 4.6 Distribution of soil samples with predicted BD between 0.8 and 0.9 cm-3 in validation data

There were two interesting results about the distribution of GBM validation results. High density of predicted BD values assembled between 1.20 and 1.60 g cm^{-3} (grey dash line and dot line respectively in Figure 4.4). It was mainly caused by the similar distribution of predicted BD in calibration; in other words, when a majority of predicted BD is located within a specified range of values in calibration data set, most of the predicted BD in validation data set would be most likely in this range too. It is a characteristic of tree based models that they make prediction based on averaging within terminal nodes groups of individuals, thus they can't make predictions outside the learning datasets ranges, unlike linear model. There was a weak clustering of predicted BD around 0.85 g cm^{-3}. In order to determine the likely cause, we identified these soil

samples with predicted BD values between 0.8 and 0.9 g cm^{-3} on the two-dimensional space based on the two most contributing covariates, namely SOC and clay (Figure 4.6). It shows that a majority of these soil samples had a SOC content larger than 4% (black dash vertical line) and a clay content larger than 25% (black dash horizontal line). In the calibration data set, there were only six similar soil samples and their mean BD value was 0.88 g cm^{-3} (grey dash dot line in Figure 4.4). Therefore, calibration data's distribution in the feature space of covariates had a great effect on GBM modelling.

4.3.3 Performance on combinations of GBM models

According to the definition in Table 4.2, GBM models with 25 combinations of calibration and validation subsets were evaluated by four indices (Table 4.4). Four calibration subsets including RMQS+IGCS, IGCS, $IGCS_{0-50}$ and $IGCS_{>50}$ gained similar SDPE (0.128-0.132 g cm^{-3}), RMSPE (0.128-0.132 g cm^{-3}) and R^2 (0.392-0.446), and the performance of RMQS calibration subset was lower with SDPE, RMSPE and R^2 in 0.157 g cm^{-3},0.156 g cm^{-3} and 0.299 respectively.

Performance of GMB models on five validation subsets differed. When comparing the accuracy on RMQS+IGCS validation subset, C1 was found to have the lowest SDPE and RMSPE values, and highest R^2; C11, C16 and C21 performed worse, with a higher SDPE (>1.85 g cm^{-3}), a higher RMSPE (>1.85 g cm^{-3}), and a lower R^2 (<0.40); though C6 had comparable SDPE and RMSPE values with C1, but its R^2 was 0.43 and much than that of C1 (0.53). RMQS validation subset had a similar trend with RMQS+IGCS. For IGCS, $IGCS_{0-50}$ and $IGCS_{>50}$ validation subsets, there was no significant difference of SDPE and RMSPE between five calibration subsets; however, C3, C4 and C5 gained the better R^2 than others. Indeed, GBM model calibrated by whole Region Centre data (RMQS+IGCS) yielded better indicators of performance in all validation subsets than other models (Table 4.4).

4.3.4 Assessment of the validity domain of the GBM model

4.3.4.1 Validity domain and goodness of prediction

Distance cutoff limits between 90% and 100% with an interval of 0.5% were tested in order to find a balance between the validity domain of GBM model and predictive accuracy. When a specified distance cutoff limit was defined, if these validation samples had a larger distance than distance cutoff limit, then the accuracy was evaluated without

these dissimilar samples. Figure 4.7 lists the changes of RMSPE, of the number and of average distance of remaining samples within cutoff limit when distance cutoff limit decreases. When distance cutoff limit decreased from 100% to 97%, the RMSPE dropped from 0.177 to 0.163 g cm^{-3} and the number of samples within cutoff limit diminished from 5669 to 5002. Afterwards, there was a moderate decrease on RMSEP from 97% to 90% of the distance cutoff limit, while the number of samples within cutoff limit also showed a moderate rate of decrease from 5,002 to 4,230. It's understandable why the trend of average distance of remaining samples was close to that of accuracy, because the accuracy was highly affected by these dissimilar samples.

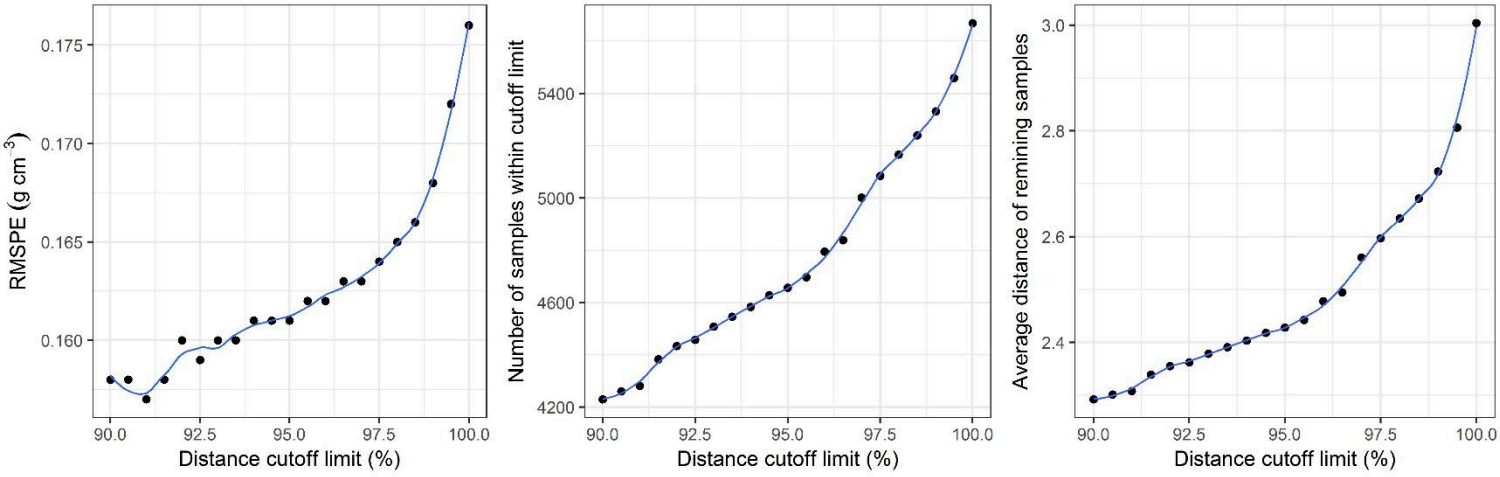


Figure 4.7 Changes of RMSPE, of the number and of average distance of remaining samples within cut-off limit when distance cutoff limit decreases

Figure 4.8 describes how the spatial distribution of dissimilar samples changes when distance cutoff limit decreases from 100% to 90%. When cutoff limit was 100%, a few samples that were most different from Region Centre data were excluded and did not show any clear spatial trend. There was an obvious enlargement of dissimilar samples when cutoff limit was set at 98%, and meanwhile a large part of these dissimilar samples located in mountainous region. With the decrease of cutoff limit, high density of dissimilar samples started to appear in other regions. When cutoff limit was 90%, dissimilar samples almost located everywhere, with a clear southeast-northwest gradient.

4.3.4.2 Additional sampling with validity domain extension

To take economic cost into account, additional samples from dissimilar samples were added to Region Centre data by an interval at 20 stepwisely and 10 additional GBM models were fitted accordingly. Figure 4.9 shows the relationship between predictive accuracy and the number of additional sampling. First, RMSPE dropped from 0.179 to 0.171 g cm^{-3} when additional samples increased to 80. When additional samples increased to 100, RMSPE remained the same at 0.171 g cm^{-3}. Then RMSPE decreased slowly from 0.170 to 0.169 when additional samples increased from 120 to 200.

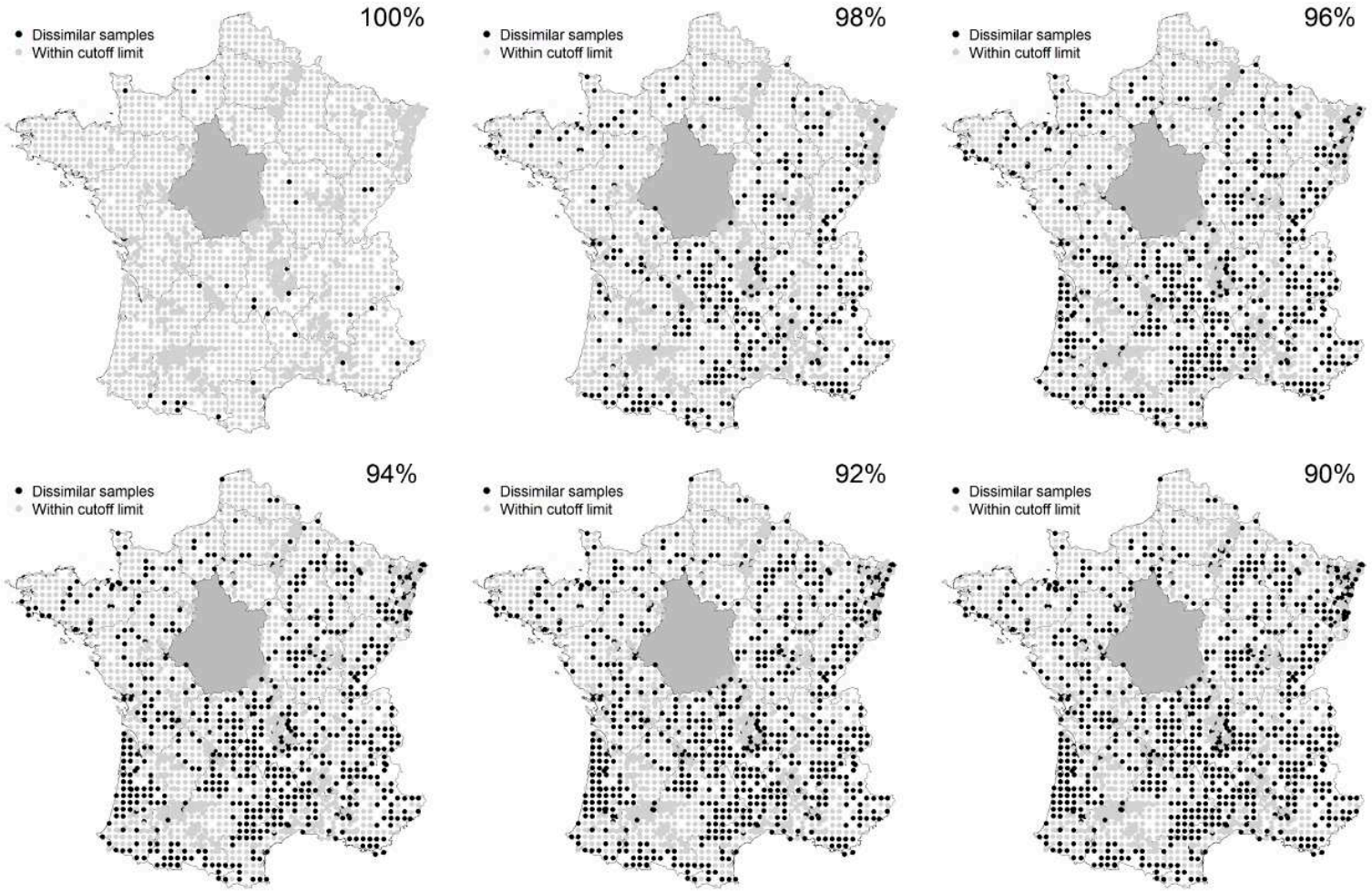


Figure 4.8 Spatial distributions of dissimilar samples when distance cutoff limits are set at 100%, 98%, 96%, 94%, 92% and 90%

4.4 Discussion

4.4.1 Performance of GBM model

GBM model was calibrated with whole data in Region Centre, the predictive accuracy on whole data from other regions was acceptable (RMSPE=0.179 g cm^{-3}). The predictive accuracy of GBM is similar with these of Nemes et al. (2010) (0.17 g cm^{-3}) and Benites et al. (2007) (0.19 g cm^{-3}), but are less than these of Tranter et al. (2007) (0.153 g cm^{-3}), Martin et al. (2009) (0.123 g cm^{-3}), Nanko et al. (2014) (0.137 g cm^{-3}), Akpa et al. (2016) (0.107 g cm^{-3}) and De Souza et al. (2016) (0.15 g cm^{-3}). These differences could be attributed to three reasons. The first one results from the differences between our dataset and others. For instance, the mean value of BD in Nanko et al. (2014) and De Souza et al (2016) were 0.60 and 1.28 g cm^{-3}, which were much smaller than our data (Table 4.1). The second important reason is the fact that all these results except those from Benites et al. (2007) and Tranter et al. (2007) were obtained from cross-validation, and no independent validation dataset was used for accuracy assessment. In Benites et al. (2007) and Tranter et al. (2007), evaluation dataset was randomly split from whole dataset but this strategy was still quite different from evaluation procedure used in this study where all validation data was spatially separated from calibration data. And lastly, it may be more challenging to apply a regional model to a large area than to spit a dataset from the same area.

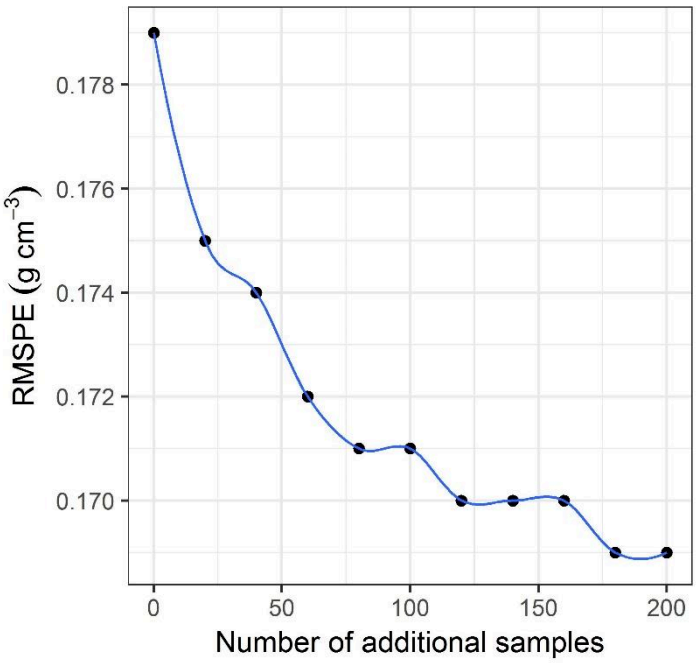


Figure 4.9 Validation accuracy with different additional sampling strategies

There was no doubt that SOC ranked first among all covariates in GBM model. In many studies, SOC was the only or most efficient covariate in modelling BD (Jeffrey, 1970; Adams, 1973; Alexander, 1980; Federer, 1983; Manrique and Jones, 1991; Nanko et al., 2014). As shown in Figure 4.3, only SOC-based recalibration models using simple PTFs could yield a good predictive accuracy. Therefore, these only SOC based PTFs could be an auxiliary method to be used in incomplete soil databases. The effect of clay was consistent with some previous studies (Tomasella and Hodnett, 1998; Kaur et al., 2002). Gravel was not as important as mentioned on RMQS data by Martin et al. (2009), and this might be explained by the fact that the majority of data in Region Centre was from IGCS dataset and lower relationship was found between BD and gravel in this dataset. The low score of the depth covariate confirmed the results of Martin et al. (2009). The weak decreasing BD trends with depth might be linked to the majority percentage of soil samples in arable among the calibration data set and the management practices tend to erase the differences between the surface and deep soil layers.

4.4.2 Limitations linked to data

As previously mentioned, calibration data comprised of IGCS and RMQS datasets. There were two obvious differences between these two databases: sampling depth and spatial sampling strategy. Consequently an important issue that should be investigated is whether sampling depth or/and spatial sampling strategy largely influenced the model predictive ability.

All data in RMQS had a maximum sampling depth of 50 cm, so it could be classified as topsoil here. For RMQS topsoil, models C7 and C17 based on grouping calibration data by soil depth showed poorer results than ungrouped calibration model C2 (Table 4.2 and Table 4.4). No significant difference were found for IGCS topsoil (sampling depth between 0 and 50 cm) between models C4, C9 and C19. Models C5, C15 and C25 also showed close predictive accuracy in IGCS subsoil (sampling depth > 50 cm).Thus, these results showed that grouping calibration data by soil depth did not improve predictive accuracy of BD (De Vos et al., 2005; Botula et al., 2015). That is to say, sampling depth did not influence the model predictive ability in this study.

Spatial sampling strategy often relates to the coverage of spatial variability and thus controls feature space of soil properties. We evaluated the effect of spatial sampling

strategy by comparing the models where IGCS and RMQS datasets were used in calibration and validation procedures separately (Table 4.4). C8 was calibrated with RMQS dataset and validated by IGCS dataset, and it had a similar RMSEP in validation with that of C3. Conversely, C12 was calibrated with IGCS dataset and validated by RMQS dataset while its RMSPE in validation was much higher than that of C2. As shown in Table 4.1, RMQS dataset had a larger feature space than IGCS dataset especially for SOC, so previous results proved that if a larger coverage of sampling distribution is linked with a larger feature space, spatial sampling strategy could improve model predictive ability.

Overall, using legacy data such as IGCS lead to improved prediction, suggesting that it is worth rescuing such legacy data besides different vintages (Arrouays et al., 2017). Using however, IGCS data for validation could lead to biased estimates of performance as it was built upon purposive sampling strategy. However, evaluation conducted using IGCS remained close to these conducted with RMQS, showing that bias, although present, did not lead to an overestimation of the performance.

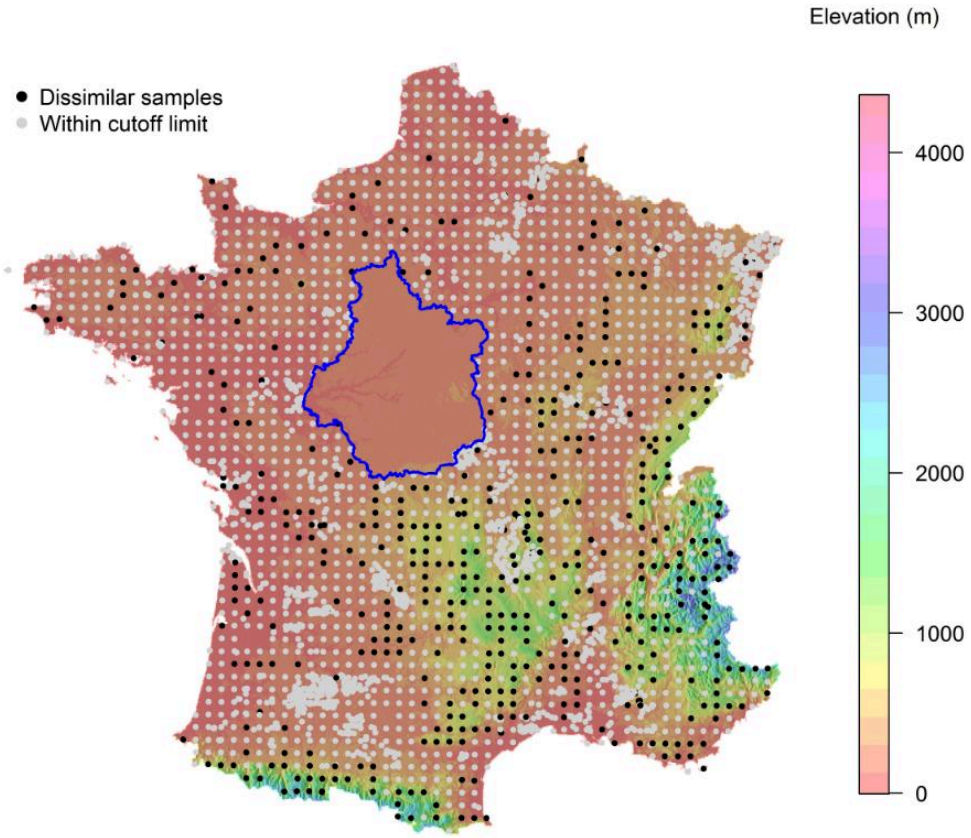


Figure 4.10 Spatial distribution of dissimilar samples when distance cutoff limit is set as 97%

4.4.3 Relationship between dissimilar samples elevation and land use

The distance cutoff was optimized to 97% after a balance between predictive accuracy and between the validity domain of GBM model, which was close to the cutoff value of 97.5% suggested by Rousseeuw and Zomeren (1990). It is interesting that most of dissimilar samples originated from mountainous regions including the Pyrenees, the Massif Central, the Alps, the Burgundy, the Jura and the Vosges (Figure 4.10). We calculated the dissimilar samples' percentage by the number of dissimilar samples dividing the number of validation data at given elevation ranges (interval at 200 m) (Figure 4.11). The results showed that, with the increase of the elevation, high proportion of dissimilar samples were found. When the elevation was higher than 1,000 m, more than 50% of samples in validation data were dissimilar samples.

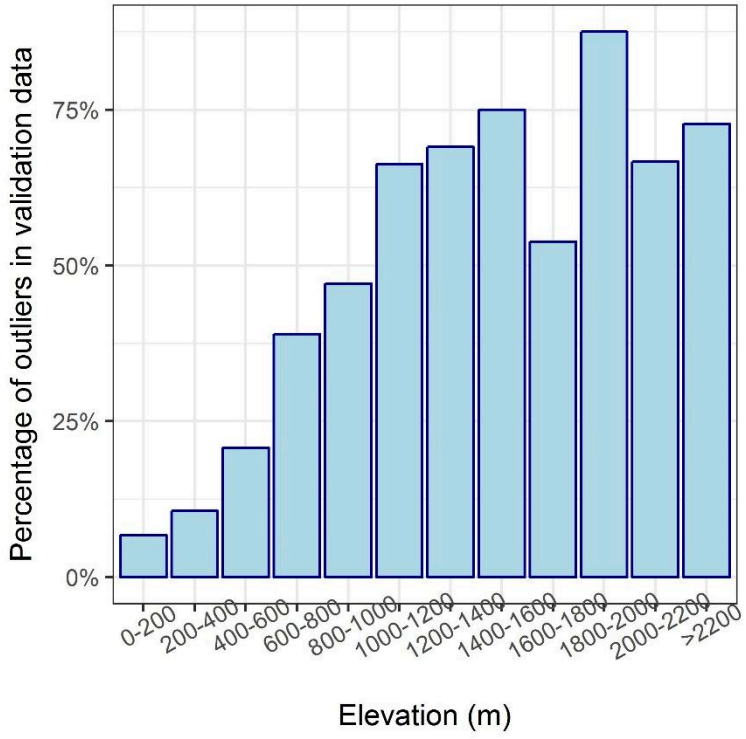


Figure 4.11 Percentage of dissimilar samples in validation data at given elevation intervals

We further analyzed the percentage of land use for these dissimilar samples (17 samples without land use records were excluded), forest accounted for the largest proportion of the dissimilar samples, with nearly 37%; pastures came the second with

27%; around 22% and 10% of dissimilar samples were in arable land and grassland respectively while about 5% of them were under other land use such as shrubland, permanent crops and wetland (Figure 4.12). These dissimilar samples had high SOC content (mean SOC > 4 g kg^{-1}), especially for these samples in forest, grassland and pastures (Figure 4.12), and they were extremely different from data in Region Centre (mean SOC of 0.75 g kg^{-1}). Meanwhile, forest, grassland and pastures accounted for nearly 75% of dissimilar samples and most of them located in mountainous regions with very shallow soil lower boundary but high SOC content (Figure 4.12). Considering the fact that calibration data was located in the Middle Loire basin and 70% of them were from arable land, a low mean SOC as well as a small proportion of samples with high SOC were found in calibration data. Therefore, the large difference of SOC associated with land use and elevation in other region lead to the interesting spatial distribution of dissimilar samples in Figure 4.10. Though Martin et al. (2009) showed that land use was a poor predictor of bulk density compared to SOC, information such as land use and elevation can improve sampling design in order to cover a wide range of SOC content.

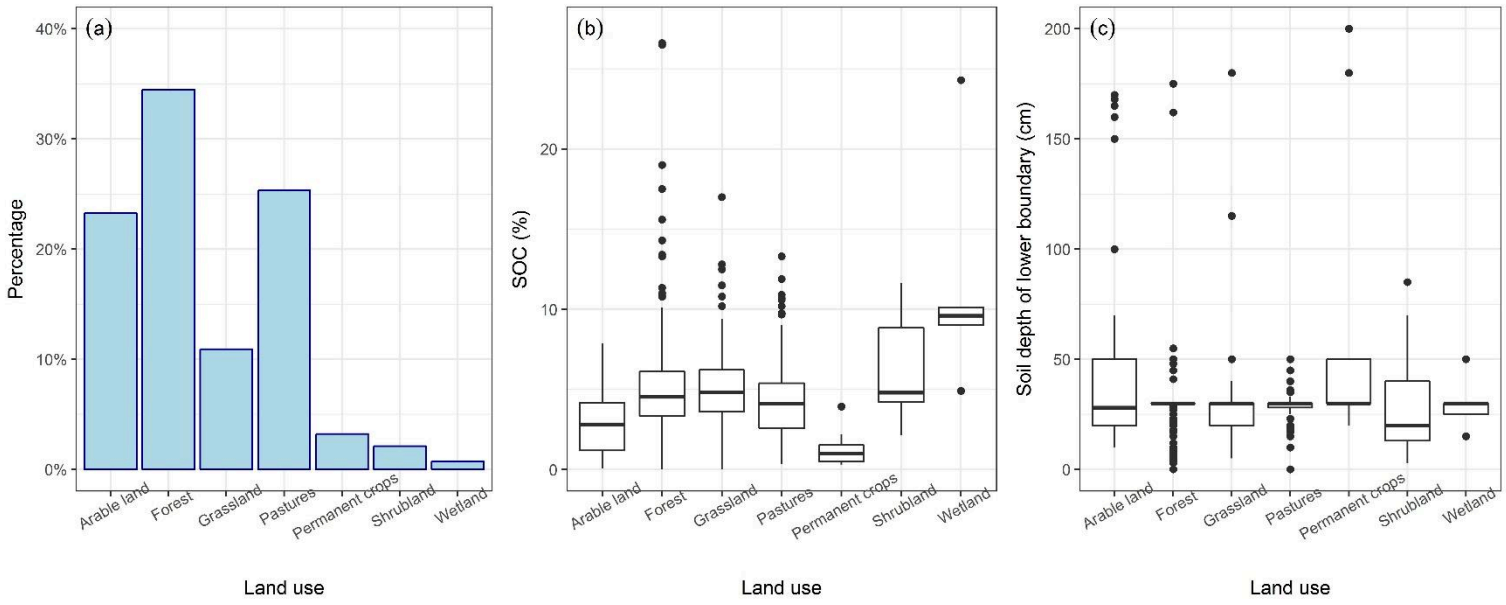


Figure 4.12 Percentage of dissimilar samples on different land use at 97% distance cutoff limit (a) and statistics of SOC (b) and soil depth of lower boundary (c) on these land use

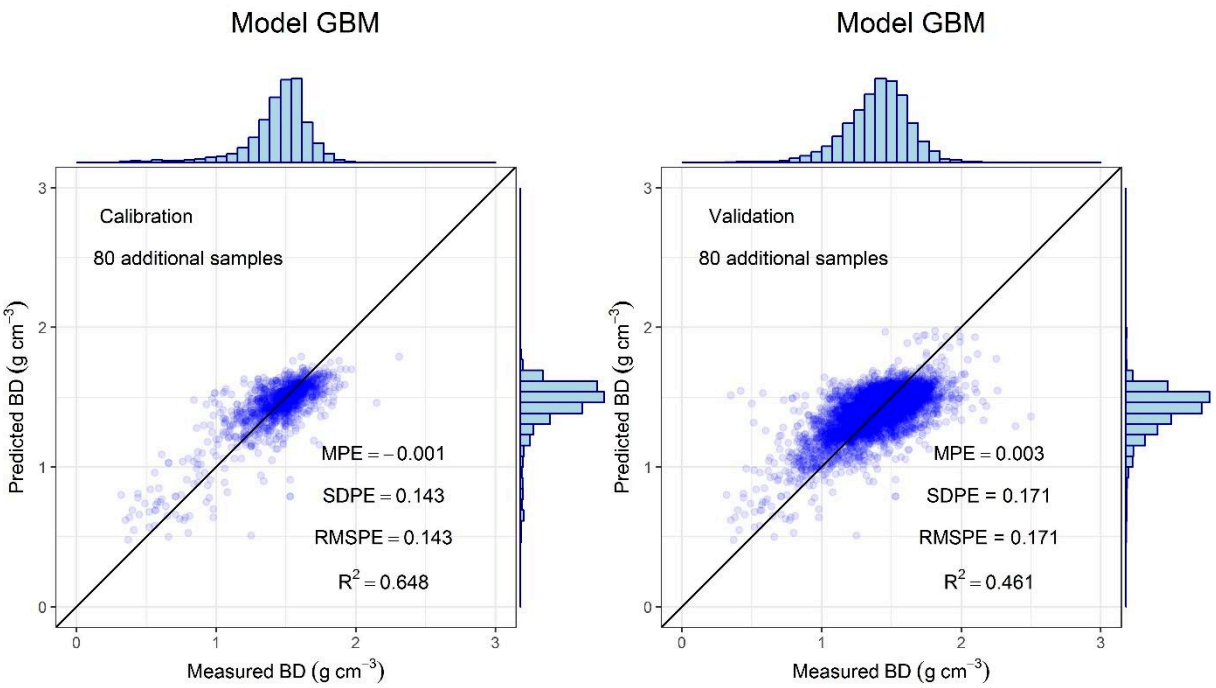


Figure 4.13 Measured vs. predicted bulk densities in calibration and validation using GBM after adding 80 additional samples

4.4.4 Usefulness of additional sampling strategy

In our study, 80 additional samples seemed an optimized option to satisfy the balance between economic costs and predictive accuracy (Figure 4.9). After adding 80 additional samples, new calibration model had a comparable accuracy with original model, but new calibration model had a much wider range of measured BD than before, especially for low BD contents (Figure 4.4, Figure 4.13). In validation, RMSEP decreased to 0.171 g cm^{-3} after adding 80 additional samples. The weak clustering problem in original validation (grey dash dot line in Figure 4.4) was solved after adding additional samples into calibration procedure, and these low BD samples well distributed along 1:1 line. These soil samples with measured BD near 0.5 g cm^{-3} were still over estimated to around 1.5 g cm^{-3} after adding 80 additional samples. It showed that BD of some soil samples differs from these of other soil samples though they have similar covariates range. As a result, we should consider that the use of distance metrics makes sense only if all the main controlling factors on BD are similar between calibration data and soil samples without BD. If some regions have different controlling factors, these soil samples without

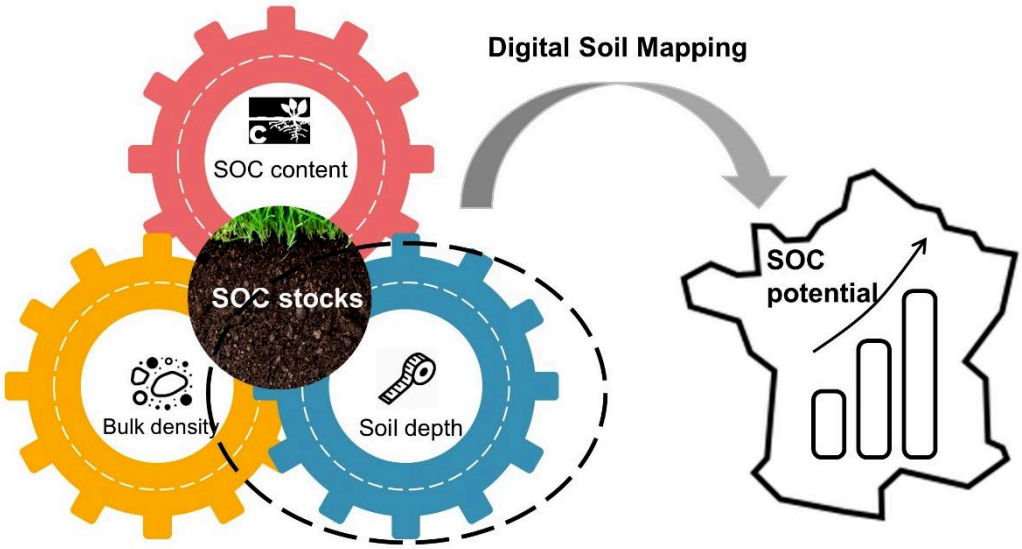
BD will be poorly predicted. Therefore, in order to mitigate this problem when using distance metrics to determine validity domain of PTFs, two solutions are recommended. One is to design a purposive sampling strategy based on the controlling factors of variance and conditional Latin Hypercube Sampling algorithm (Minasny and McBratney, 2006) are highly suggested. The other strategy is based on systematic and dense sampling of the spatial space. This is the case for some national/pan European existing database such as soil monitoring networks (SMNs). Though sampling density of SMNs differs between countries, the SMNs provide wide geographical and bibliographical coverage of soil information (Arrouays et al., 2012), and therefore, the SMNs should improve the application of distance metrics.

4.5 Conclusions

We established a new GBM model for predicting soil bulk density in Region Centre of France and tested its validity domain on mainland France using distance metrics. Compared to six groups of revised PTFs, GBM model performed better with a reasonable predictive accuracy. In GBM model, geographical coverage of soil samples had a large effect on goodness of prediction. Distance metrics successfully excluded those samples dissimilar to the calibration data and the optimization of distance cutoff limit is suggested. An additional sampling strategy based on these samples exceeding validity domain improved predictive ability of GBM model. Our study also suggests that, when determining validity domain of PTFs for BD or other soil properties, a purposive sampling strategy or a systematic and dense sampling strategy will improve the robustness of modelling.

Chapter 5

Probability mapping of soil thickness by random survival forest at a national scale



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

5.1 Introduction

Soils are of great importance in supporting, provisioning, and regulating ecosystem services, such as food production and climate change mitigation (Clothier et al., 2011; Keesstra et al., 2016; Millennium Ecosystems Assessment, 2005). As stated by recent studies (e.g., Bouma, 2018; Groshans et al., 2018; Marx et al., 2019), there is a rising demand for up-to-date and ecosystem service relevant soil information. Therefore, substantial effort is needed to communicate soil information among diverse audiences and produce fine resolution soil maps to support practical land management. In this study, we use the *GlobalSoilMap* project specifications. These specifications focus on delivering consistently produced high-resolution soil property information throughout the world by predicting mean values and their prediction intervals (PIs) (Arrouays et al., 2014a, 2014b; Sanchez et al., 2009). Among the twelve soil properties to be predicted following the recommendations of *GlobalSoilMap*, soil thickness (ST) is a key property. In this study, in line with *GlobalSoilMap*, ST is defined as 'the depth (cm) from the soil surface to the lithic or a paralithic contact' (Soil Survey Division Staff, 1993). The ST is highly relevant for soil hydro-mechanical modelling (Tesfa et al., 2009; Wang et al., 2006), soil erosion impact, landscape evolution, vegetation growth (Heimsath et al., 2001; Meyer et al., 2007), and for calculating soil functions (e.g., available water capacity (Leenaars et al., 2018; Román Dobarco et al., 2019a), soil structure (Rabot et al., 2018; Vogel et al., 2018), and soil organic carbon stocks (Batjes, 1996; Chen et al., 2019a). Despite the great importance of accurate ST information, the large spatial variability and high cost of ST measurements make ST determination difficult (Lacoste et al., 2016). Discordance in the definition of ST also hampers ST modelling, especially when data are collected from various projects (Lacoste et al., 2016). The observed ST recorded in soil information systems for some profiles are often less than the actual ST (i.e., right censored data).

ST results from the mass balance between soil formation from the bedrock and soil transport by erosion and sedimentation (Heimsath et al., 1997; Heimsath et al.,1999); thus, it varies as a function of physical, chemical, and biological processes (Román et al., 2018). ST can be related to these processes by modeling the relationship between the main soil-forming factors, i.e., Jenny's Soil-Landscape paradigm (Jenny, 1941): parent material, climatic conditions, organisms, terrain relief, and time (Dietrich et al., 1995;

Minasny and McBratney, 1999). More recently, McBratney et al. (2003) formulated the concept of the *scorpan* model, which also includes also soil information and spatial location.

Various approaches for ST modelling and mapping have relied on modelling the relationship between the main soil forming factors. The majority of these approaches can be broadly classified into two groups: 1) physically based and mechanistic models, which predict ST using soil process models based on the rates of weathering, denudation, and accumulation (Bonfatti et al., 2018; Dietrich et al., 1995; Minasny and McBratney, 1999; Pelletier and Rasmussen, 2009); and 2) empirical models, including statistical and geostatistical methods (Kuriakose et al., 2009). These models rely on the empirical relationships between ST and explanatory covariates of inferential attributes (e.g., plant species, precipitation, and parent material).

For the latter, a wide range of statistical methods have been previously applied in ST modelling, including canonical correspondence analysis and principal component analysis (Odeh et al., 1991), multiple linear regression (Moore et al., 1993), expert knowledge and fuzzy logic (Zhu et al., 2001), Generalized Additive Models and Random Forest (Tesfa et al., 2009), and Cubist and Gradient Boosting Modelling (Lacoste et al., 2016; Mulder et al., 2016a).

Within the field of geostatistics, various kriging techniques have often been used to predict and spatially interpolate ST from point samples. Ordinary Kriging was most commonly used among these kriging techniques (Penížek and Borůvka, 2006; Vanwalleghem et al., 2010). The prediction variance was typically reduced when including additional prediction variables using regression kriging (Kuriakose et al., 2009; Odeh et al., 1995) or Kriging with External Drift (Bourennane et al., 1996; Kempen et al., 2015).

None of the studies referred to above addressed the issue of having right censored data entries in their soil databases. However, it is often the case that the actual ST is thicker than the observed ST, which can mainly be attributed to practical constraints, such as the standard auger length (120 cm), and time constraints. In fact, in soil sciences very few studies consider the effect of right censored data; the issue is often ignored or processed by adding a fixed value (e.g., 30 cm) in ST modelling (Knotters et al., 1995; Vaysse and Lagacherie, 2015; Lacoste et al., 2016; Shangguan et al., 2017). Some previous works dealt with left censored data, especially regarding data below detection limits (e.g.,

de Oliveira, 2005; Fridley and Dixon, 2007; Orton et al., 2009; Orton et al., 2012; Villaneau et al., 2011). Ignoring the presence of right censored data entries within a database and relying on the observed ST for those entries will result in an underestimation of modelled ST (Vaysse and Lagacherie, 2015; Shangguan et al., 2017).

However, right censored data are commonly used in statistics and medical research, especially in survival analysis. Several models have been used to deal with right censored data in survival analysis, including the Kaplan Meier method (Kaplan and Meier, 1958), Cox regression (Andersen and Gill, 1982), and Random Survival Forest (RSF, Ishwaran et al., 2008). The Kaplan Meier method and Cox regression mainly deal with linear effects, but RSF is capable of handling complex non-linear effects that may exist between predictor variables (Mogensen et al., 2012). Therefore, as previously suggested by Styc and Lagacherie (2016), RSF may have the best potential for identifying and correcting right censored data used for Digital Soil Mapping (DSM).

In this study, the potential of RSF was evaluated for ST mapping in mainland France. The main objectives of this study are noted below:

1) Apply RSF for mapping the probability of exceeding a certain ST using both actual and right censored ST data from the French Soil Monitoring Network (RMQS) and
2) Derive the 90% confidence intervals of the specific ST using bootstrapping.

5.2 Material and methods

5.2.1 Soil dataset

We used ST data from the RMQS soil database that were gathered between 2001 and 2009 (Jolivet et al., 2006), covering different soil, climate, relief, and land cover conditions (Figure 5.1). The RMQS dataset is based on a 16 km × 16 km square grid where all sites are selected at the centre of each grid cell. When sampling the exact location was not possible, a site was selected as close as possible to the grid centre. A soil pit was dug, and the surrounding information (land use and geomorphology) and a detailed description of the soil profile were recorded for each site, including soil horizon depth and ST. Auger boring was recommended (but not mandatory) to complete the soil profile when the soil pit was not thick enough to determine the ST. For more detailed information about the soil sampling design and laboratory analysis, see Chen et al. (2018).

Among 2,109 RMQS sites, ST was explicitly recorded for 1,020 sites (down to a lithic or paralithic contact), while the remaining 1,089 sites were right censored data. The ST for nine RMQS sites was set to 0, as these sites were identified as mountainous sites with bare rock.

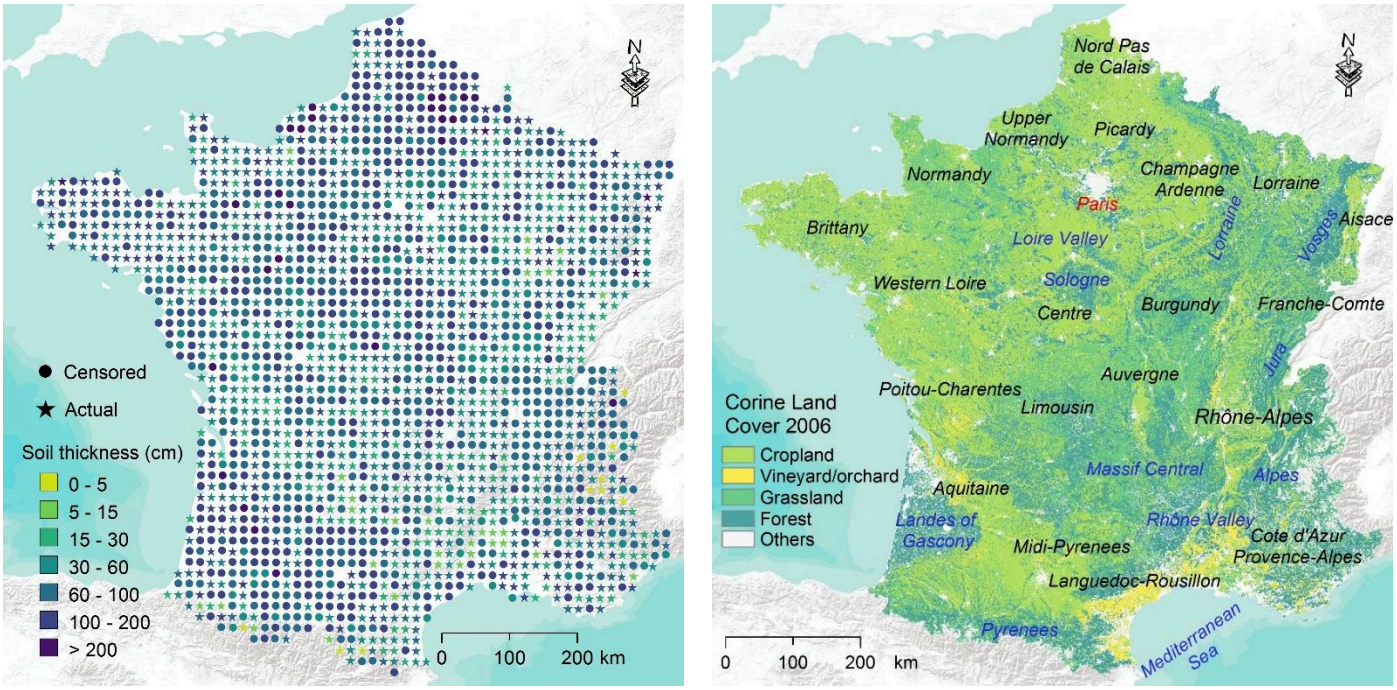


Figure 5.1 Locations of RMQS sites with actual (dotted) and censored (star) ST values. For each site, ST is classified based on the *GlobalSoilMap* standard depths. Corine Land Cover map of 2006 of mainland France (right) with the administrative regions (black italics) and natural geographic regions (blue italics)

5.2.2 Exhaustive covariates

We used a DSM framework (McBratney et al., 2003) to model the relationships between ST and ancillary covariates (Table 5.1). These covariates cover a series of soil formation related environmental factors, including soil, climate, organisms, relief and parent material. Before modelling, these covariates were re-projected to Lambert 93 (official projection for mainland France) and resampled to a 90 m resolution (in raster format) using a bilinear interpolation (numeric covariates) or nearest neighbour (categorical covariates).

Table 5.1 Exhaustive covariates used for ST modelling (after Mulder et al., 2016b)

Variable	Abbreviation	Scale/resolution	Soil forming factor	Reference
Elevation	ELEVATION	90 m	Relief	Jarvis et al. (2008)
Compound topographic index	CTI	90 m	Relief	Jarvis et al. (2008)
Curvature	CURVATURE	90 m	Relief	Jarvis et al. (2008)
Exposition	EXPOSITION	90 m	Relief	Jarvis et al. (2008)
Roughness	ROUGHNESS	90 m	Relief	Jarvis et al. (2008)
Slope	SLOPE	90 m	Relief	Jarvis et al. (2008)
Slope cosines	SLOPECOS	90 m	Relief	Jarvis et al. (2008)
Slope position	SLOPEPOS	90 m	Relief	Jarvis et al. (2008)
Topographic wetness index	TWI	90 m	Relief	Jarvis et al. (2008)
Gravimetric data (Bouguer anomaly)	GREVIMETRY	4 km	Relief	Achache et al. (1997)
Soil type[a]	SOIL	1:1000000	Soil	IUSS Working Group WRB (2006)
Erosion rates	EROS	1:1000000	Soil	Cerdan et al. (2010)
Rate of river network development and persistence	IDPR	1:50000	Soil and parent material	Info Terre – Site cartographique de référence sur les géosciences (2014)
Parent material	PM	1:1000000	Parent material	King et al. (1995)
Mean annual net primary production	NPPMEAN	1 km	Organisms	NASA LD (2001)
Forest type	BDFOREST	Min area 2.25 ha	Organisms	Inventaire Forestier National (2006)
Land cover from Sentinel-2	LCS	10 m	Organisms	Inglada et al. (2017)
Corine land cover 2006	CLC06	250 m	Organisms	Feranec et al. (2010)
ECOCLIMAP land use	ECOCLIM	1 km	Organisms	Faroux et al. (2013)
Climatic zones	TYPO	1 km	Climate	Joly et al. (2010)
Mean annual precipitation	RAINFALL	1 km	Climate	Hijmans et al. (2005)
Mean annual temperature	TEMPMEAN	1 km	Climate	Hijmans et al. (2005)

[a] Soil type defined by World Reference Base (WRB)

5.2.3 Random survival forest for probability modelling of soil thickness

5.2.3.1 General introduction

RSF is an ensemble tree method for modelling right censored survival data (Ishwaran et al., 2008). RSF is an extension of Breiman's (2001) random forest (RF), known as an

ensemble learning approach that is improved by injecting randomization into the base-learning process. For example, using a human analogy, a specific status introduced in RSF aims to distinguish whether it is a death case (status = 1) or a survival case (status = 0) at a given observed time. The RSF does this by estimating the presence-absence probability. Applying this censoring concept to ST, at a given site there are two possibilities for a thickness to be recorded: i) the actual ST has been observed down to lithic or paralithic contact as defined before (status = 1) or ii) the lithic or paralithic contact has not been reached and the actual ST remains unknown (status = 0). For the latter, it is only known that the actual ST is greater than the observed ST. RSF uses a new type of predicted outcome that contains a cumulative hazard function (CHF, see Section 5.2.3.2).

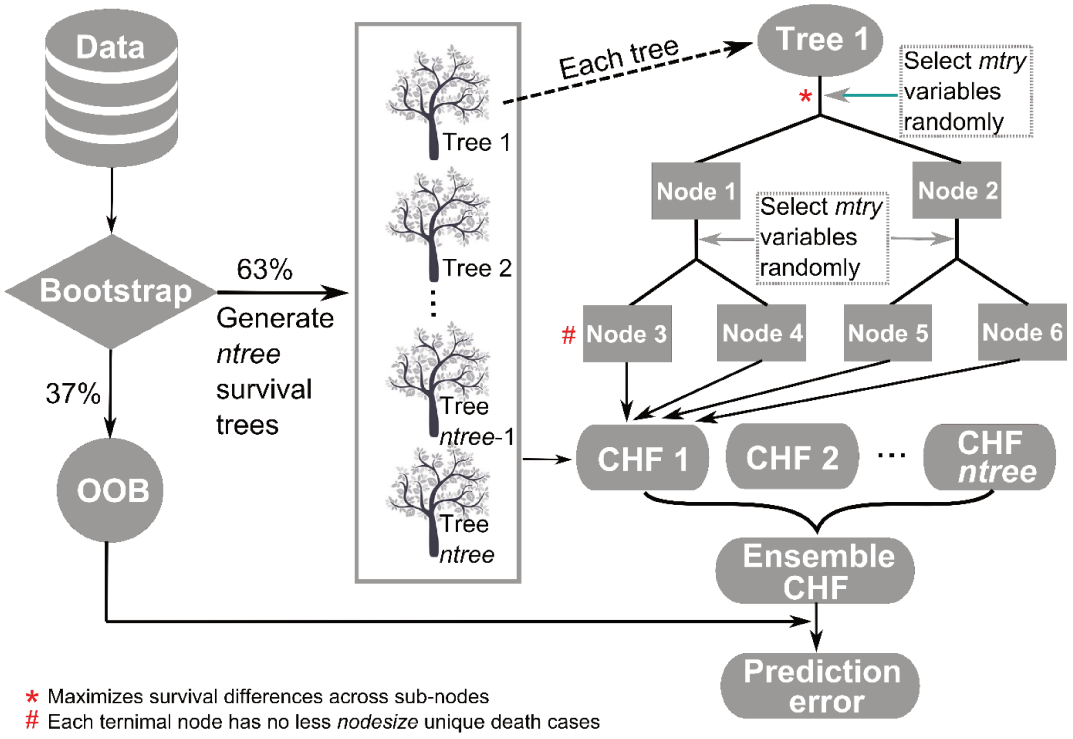


Figure 5.2 Random survival forest workflow

5.2.3.2 Random survival forest model fitting

RSF model fitting involves the following steps (Ishwaran et al., 2008), as shown in Figure 5.2.

1) Select *ntree* bootstrapped samples from the calibration data. Approximately 37% (e^{-1}) of the calibration data are excluded in each bootstrapped sample, which are so-called out-of-bag (OOB) data.
2) Grow a survival tree for each bootstrapped sample. At each node of the survival tree, randomly select *mtry* covariates for splitting the data. Survival splitting criteria are then used, and each node is split on that covariate, which maximizes survival differences across sub-nodes.
3) Grow the survival tree to full size under the constraint that a terminal node should have no less than *nodesize* unique actual ST samples.
4) Calculate a CHF for each survival tree and obtain the ensemble CHF by averaging all the survival trees for each sample.
5) Calculate the prediction error of the ensemble CHF based on OOB data.

Ensemble cumulative hazard function and ensemble survival function

Constructing the ensemble CHF is crucial for RSF. Hereafter, we provide details about the procedure for a better understanding.

For a survival tree, let $(ST_{1,h}, \delta_{1,h}), \ldots, (ST_{n(h),h}, \delta_{n(h),h})$ be the observed ST and the 0–1 censoring status (δ) for n samples in a terminal node h. Here, let $ST_{1,h} < ST_{2,h} < \cdots < ST_{n(h),h}$ be the different observed ST in the terminal node. The CHF estimate for h is then defined by the Nelson–Aalen estimator $\widehat{H}_h$:

$$\widehat{H}_h(st) = \sum_{st_{l,h} \le st} \frac{a_{l,h}}{Y_{l,h}} \quad (5.1)$$

where $a_{l,h}$ and $Y_{l,h}$ are the number of actual ST samples and all samples at observed ST $st_{l,h}$, respectively. All the samples within the terminal node h have the same CHF.

Each sample i has a *mtry*-dimensional covariate $\mathbf{x}_i$ that will belong to a unique terminal node h. Therefore, the CHF for i is the Nelson–Aalen estimator for $\mathbf{x}_i$' terminal node:

$$H(st|\mathbf{x}_i) = \widehat{H}_h(st) \quad (5.2)$$

Equation 5.2 describes the CHF from an individual tree. The ensemble CHF is computed by averaging over *ntree* trees. The bootstrap ensemble CHF for sample i is defined below (the definition of OOB ensemble CHF please refer to Ishwaran et al., 2008):

$$H_e(st|\mathbf{x}_i)=\frac{1}{\text{ntree}}\sum_{n=1}^{\text{ntree}} H_n(st|\mathbf{x}_i) \quad (5.3)$$

where $H_n(st|\boldsymbol{x})$ is the CHF for a tree grown from the n^{th} bootstrap sample.

The survival function is a probability density function that describes the survival probability at a given ST. In RSF, the ensemble survival function (S_e) could be derived from ensemble CHF (Mogensen et al., 2012):

$$S_e(st|\mathbf{x}_i)=\exp\left\{-\frac{1}{\text{ntree}}\sum_{n=1}^{\text{ntree}} H_n(st|\mathbf{x}_i)\right\} \quad (5.4)$$

Here, the survival probability at a given ST is equal to the probability of exceeding a given ST or censored probability at a given ST. The probability of exceeding a given ST ranges from 0 to 1, and when it is close to 1, the location has a high probability of being censored. Therefore, in this latter case, the actual ST has a high probability of being thicker than the censored ST.

Node splitting rule

The node splitting rule is another important parameter in RSF. There are several choices for splitting rules, including the log-rank splitting rule, conservation splitting rule, log-rank score rule, and fast approximation to the log-rank splitting. Here, the log-rank splitting rule is used as the default splitting rule, as suggested by Ishwaran et al. (2008). We define $st_1 < st_2 < \cdots < st_i$ as the ST intervals and $x_{i,j}$ and $a_{i,j}$ as the number of samples and number of actual ST samples at ST st_i in the sub-nodes j (1 or 2), respectively. Here, $x_i = x_{i,1} + x_{i,2}$ and $a_i = a_{i,1} + a_{i,2}$. The log-rank test for a split at the value n of the covariate c is defined as

$$L(c, n)=\frac{\sum_{i=1}^{I}(a_{i,1}-x_{i,1}\frac{a_i}{x_i})}{\sqrt{\sum_{i=1}^{I}\frac{x_{i,1}}{x_i}(1-\frac{x_{i,1}}{x_i})(\frac{x_i-a_i}{x_i-1})a_i}} \quad (5.5)$$

where the value $|L(c, n)|$ is the measure of node split, and $x_{i,1}$ and $a_{i,1}$ are the number of samples and number of actual ST samples, respectively, at ST st_i when c is less than n. The larger the $|L(c, n)|$ value, the larger the difference between two sub-nodes and a better split. The best split at each note is determined by searching the optimized covariate c^* and split value n^* to maximize the $|L(c, n)|$ value.

Prediction error

In survival analysis, Harrell's concordance index (Harrell Jr et al., 1982) is commonly used for estimating prediction error as it does not depend on choosing a fixed time for model evaluation and specifically accounts for censoring (May et al., 2004). The concordance index (*C* index) is calculated by the following steps in ST modelling.

1) Generate all possible pairs of samples over the data.
2) Remove pairs whose lower ST is censored. Remove pairs *i* and *j* if $st_i = st_j$ unless at least one is an actual ST sample. The total number of permissible pairs is recorded as *Per*.
3) For each permissible pair where $st_i \neq st_j$: if the thinner ST has worse predicted outcome (higher cumulative hazard value), count 1; ii) otherwise, count 0.5. For each permissible pair where $st_i = st_j$ and both are actual ST samples: i) if predicted outcomes are equal, count 1; ii) otherwise, count 0.5. For each permissible pair where $st_i = st_j$ and not both, are actual ST samples: i) if the actual ST sample has a worse predicted outcome, count 1; ii) otherwise, count 0.5. The sum of all permissible pairs is recorded as *Con*.
4) The *C* index is defined by the ratio of *Con* to *Per*.

In RSF, the *C* index is computed via OOB data using the steps mentioned above, and it ranges between 0 and 1. The prediction error is calculated by the1-*C* index, so it is also between 0 and 1. A lower prediction error represents better model performance for the calibration model.

5.2.3.3 Assessing the main controlling factors for ST modelling

To assess the main controlling factors for ST in France, the variable importance of the ST predictors (i.e., covariates used) in the RSF model were evaluated. In RSF, the variable importance of a covariate *c* is calculated by dropping OOB samples down their in-bag survival tree. A sub-node is randomly assigned when encountering a split for *c*, and then an average of the CHF obtained from these trees is calculated. The variable importance for *c* is calculated as the difference of prediction error between the new ensemble obtained using randomized *c* assignments and the original ensemble. A larger variable importance value indicates a higher contribution to the model for a covariate.

5.2.4 Soil thickness probability mapping and bootstrapping for determining prediction uncertainty

As introduced in Section 5.2.3.2, the RSF model outcome entails a function between the survival (censored) probability and ST for each prediction. In other words, the censored probability can be calculated over the full soil profile (0 to the maximum depth of actual ST samples) for any position in mainland France from RSF. As an example, the censored probabilities for the six *GlobalSoilMap* standard depths were extracted from the survival probability function (Figure 5.3); those depths are 5, 15, 30, 60, 100, and 200 cm, which we refer to hereafter as ST5, ST15, ST30, ST60, ST100 and ST200, respectively. From this, we derived a probability map for each *GlobalSoilMap* standard depth in mainland France.

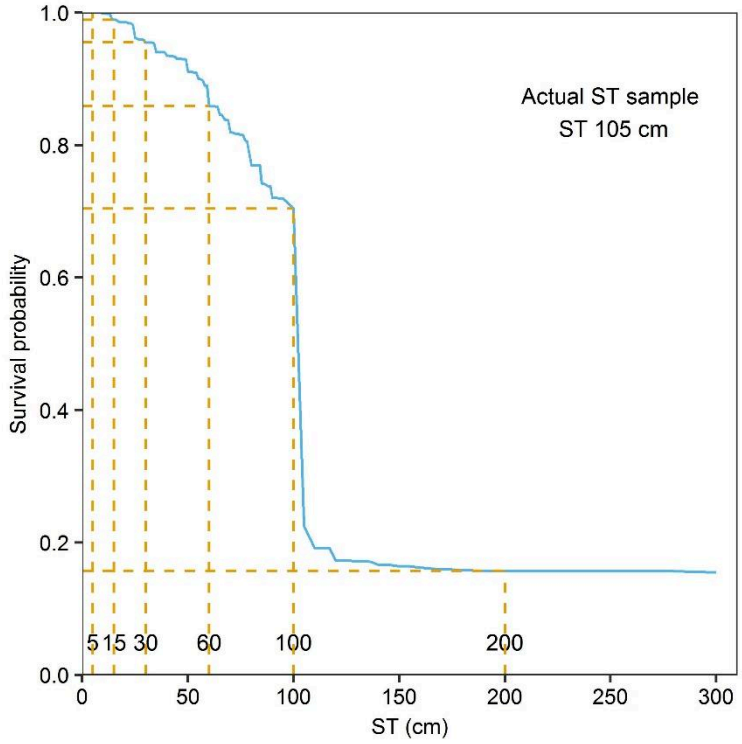


Figure 5.3 Survival probability curve (blue solid line) for one location predicted by RSF. The orange dashed vertical lines indicate the six *GlobalSoilMap* standard depths, and the orange dashed horizontal lines indicate their corresponding censored probabilities that are derived from the survival probability curve

Bootstrapping was applied to determine the average and 90% Confidence Intervals

(CIs) of the RSF model. Hence, we did not determine the 90% PIs as is recommended by the *GlobalSoilMap* specifications; instead we estimated the 90% CIs. This was deemed suitable, as we were not able to identify the random error in the RSF model. Consequently, the estimated 90% CIs would be narrower than 90% PIs. The bootstrap samples were drawn 50 times by repeated random sampling with replacement of the RMQS sites; the RMQS sites not used in each bootstrap sample were used to evaluate the model performance of each bootstrap RSF model (details in Section 5.2.5). Note that the bootstrap sample used here corresponds to the initial data used in the RSF framework (Figure 5.2), not the bootstrap sample used to generate trees. Finally, using these bootstrap samples, 50 bootstrap RSF models were generated, from which 50 probability functions between the censored probability and ST could be exhaustively predicted for mainland France. After several iterative model calibrations leading to the final prediction model, we choose 50 bootstrap models because it is time-consuming to make predictions at a 90 m resolution for mainland France (RSF produces a probability function rather than a value for each pixel, so it takes 2 weeks for 50 bootstrap RSF models under parallel computing that make full use of a computer with 8 cores and 32 GB of RAM). A robust estimate of the probability of exceeding each standard *GlobalSoilMap* soil depth was determined by averaging the bootstrap predictions. Their lower and upper 90% CIs were calculated by the averaged bootstrap predictions minus and plus 1.645 times (Z score for 90% CIs) the standard deviation of bootstrap predictions, respectively. Surface area percentages of five probability intervals (0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1) were calculated from the averaged bootstrap predictions of probability maps at six *GlobalSoilMap* standard depths. The mean probability was computed by averaging all pixels of the probability map for each *GlobalSoilMap* standard depth.

5.2.5 Model performance

In addition to the CIs, the model performance of each *GlobalSoilMap* standard depth was evaluated using the RMQS sites that were not used in the bootstrap samples, which referred to an evaluation dataset from each bootstrap RSF model. For a given *GlobalSoilMap* standard depth (st_s), the prediction performance was evaluated based on the confusion matrix in which the misclassification rate was calculated based on whether the data was censored or not. Hence, given a sample with observed ST (st_o): 1) when $st_s \leq$

st_o, if the probability exceeds 0.5, the sample is correctly predicted, otherwise, it is incorrectly predicted; and 2) when $st_s > st_o$, if the probability is less than 0.5, the sample is correctly predicted, otherwise, it is incorrectly predicted.

Subsequently, the confusion matrix was calculated as the mean counts of OOB samples with actual ST and censored ST separately from 50 bootstrap predictions.

All of the statistics and modelling were performed in R (R Core Team, 2016). R package *randomForestSRC* was used for RSF modelling (Ishwaran and Kogalur, 2017).

5.3 Results

5.3.1 Summary statistics of the ST dataset

Among 2108 RMQS sites, more than half were right censored for ST (Figure 5.4). The actual ST ranged from 0 to 300 cm, with a mean value of 64 cm. The first quantile, median and third quantile were 39, 59, and 80 cm, respectively, indicating a large percentage of soils thinner than 60 cm. The censored ST ranged from 50 to 270 cm, with the mean ST (104 cm) being higher than the actual observed RMQS sites. For the censored RMQS sites, the first quantile, median, and third quantile were 75, 95, and 120 cm, respectively.

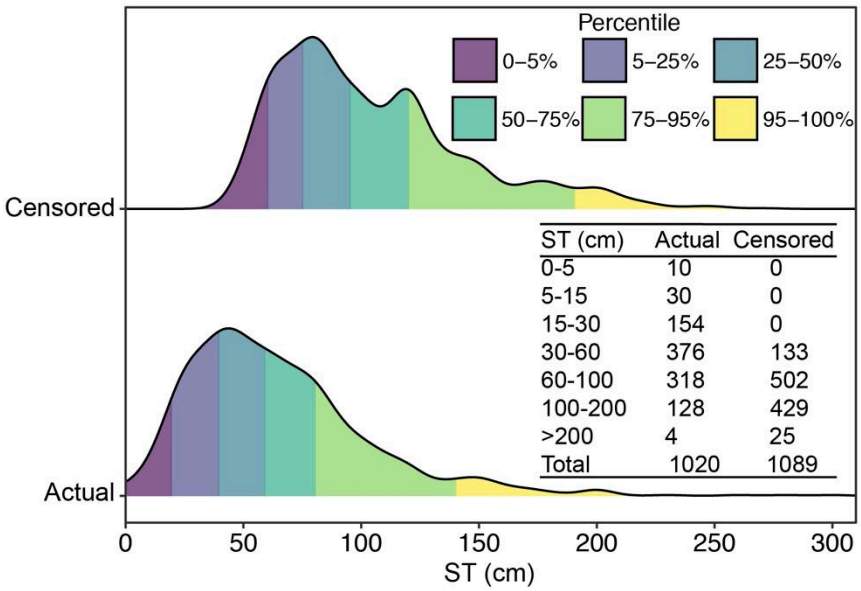


Figure 5.4 Density distribution of STs for actual and censored RMQS sites. Counts of actual and censored samples within *GlobalSoilMap* depth intervals are provided

5.3.2 Model performance

The prediction error of the calibrated RSF models decreased from 0.27 to 0.15 as the number of trees increased up to 50 (Figure 5.5). After 50 trees, the prediction error decreased slightly and became more stable as the number of trees increased (max. 300 trees). This indicated that 50 trees were sufficient for this study to produce a stable model while accelerating the prediction efficiency for big data.

The prediction performance differed when evaluated at the six *GlobalSoilMap* standard depths (Table 5.2). For the actual RMQS sites, the overall accuracy decreased from 0.989 to 0.546, when depth increased from ST5 to ST60. The overall accuracy then gradually increased up to 0.793 for ST200. The overall accuracy for censored RMQS sites were 1, 0.998, and 0.995, respectively, for SD5, SD15, and SD30, the accuracy then decreased to 0.825 for SD60 and subsequently dropped to 0.534 for ST100 and 0.563 for ST200.

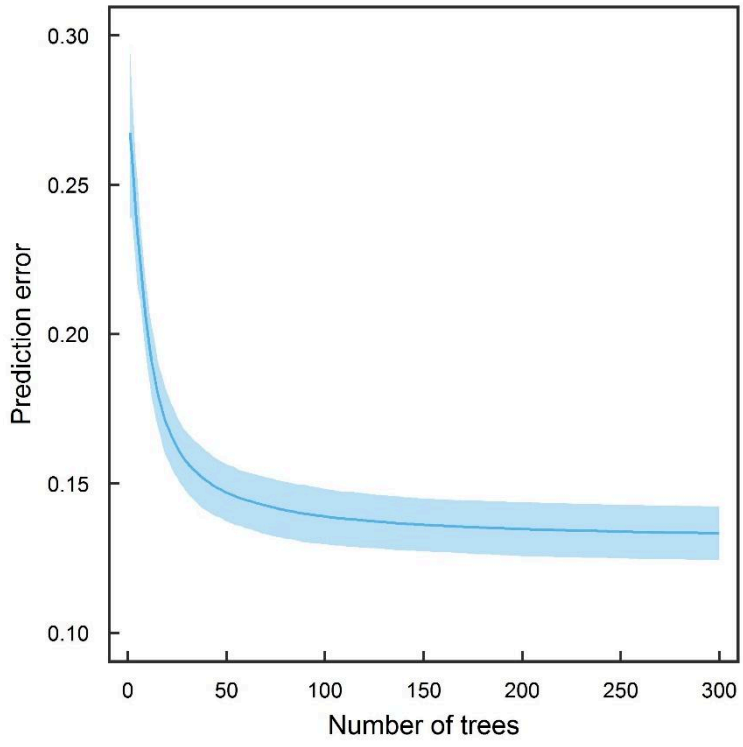


Figure 5.5 Mean and 90% confidence intervals of the prediction error, given different numbers of trees from 50 bootstrapping random survival forests

5.3.3 Controlling factors of ST modelling

Parent material (PM) and climatic zones (TYPO) were the two most important variables affecting the ST probabilities in RSF models, based on the average bootstrap RSF (Figure 5.6). The difference in prediction error between the new and the original ensembles was most affected by these two variables, despite the large 90% CIs. Roughness, precipitation, elevation, slope, gravimetry and Net Primary Production (NPP) also had large contributions in ST modelling. The remaining covariates contributed less to the RSF model and had smaller CIs.

Table 5.2 Model performance of actual and censored RMQS sites per each *GlobalSoilMap* standard depth, based on out of bag samples. The count of correctly classified sites is marked **bold**, and the overall accuracy is marked *italic underlined*

ST (cm)		Actual RMQS sites			Censored RMQS sites		
	Predicted / Observed	Thin	Thick	Accuracy	Thin	Thick	Accuracy
5	Thin	**2**	2	0.500	**0**	0	1
	Thick	0	**367**	1	0	**400**	1
	Reliability	1	0.995	*0.989*	*n.a.*	1	*1*
15	Thin	**2**	12	0.143	**0**	0	*n.a.*
	Thick	0	**356**	1	1	**399**	1
	Reliability	1	0.967	*0.962*	0	1	*0.998*
30	Thin	**5**	65	0.063	**0**	0	*n.a.*
	Thick	1	**300**	0.997	2	**398**	1
	Reliability	0.833	0.843	*0.822*	0	1	*0.995*
60	Thin	**58**	150	0.279	**7**	43	0.140
	Thick	18	**144**	0.889	27	**323**	0.923
	Reliability	0.763	0.490	*0.546*	0.205	0.883	*0.825*
100	Thin	**203**	120	0.628	**97**	138	0.413
	Thick	19	**28**	0.596	49	**117**	0.705
	Reliability	0.914	0.189	*0.624*	0.664	0.459	*0.534*
200	Thin	**294**	76	0.795	**219**	172	0.560
	Thick	1	**1**	0.500	3	**6**	0.667
	Reliability	0.997	0.013	*0.793*	0.986	0.034	*0.563*

n.a. Not available.

5.3.4 ST probability maps and associated confidence intervals

Figure 5.7 presents the ST probability maps of exceeding the six *GlobalSoilMap* standard depths and their 90% CIs for mainland France. Overall, the average probability

of exceeding the *GlobalSoilMap* standard depths of 5, 15, 30, 60,100, and 200 cm were 0.99,

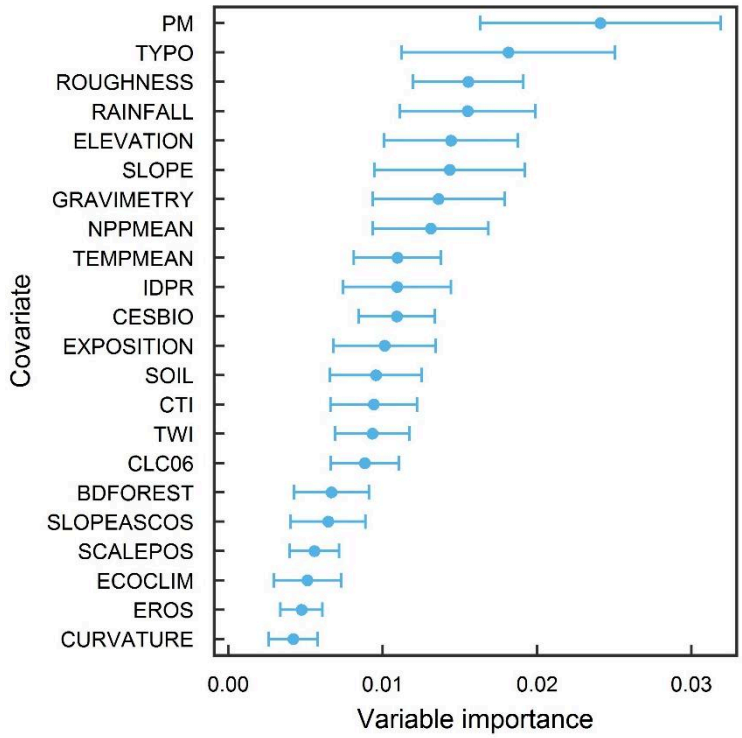


Figure 5.6 Mean and 90% confidence intervals of variable importance from 50 bootstrapping random survival forests

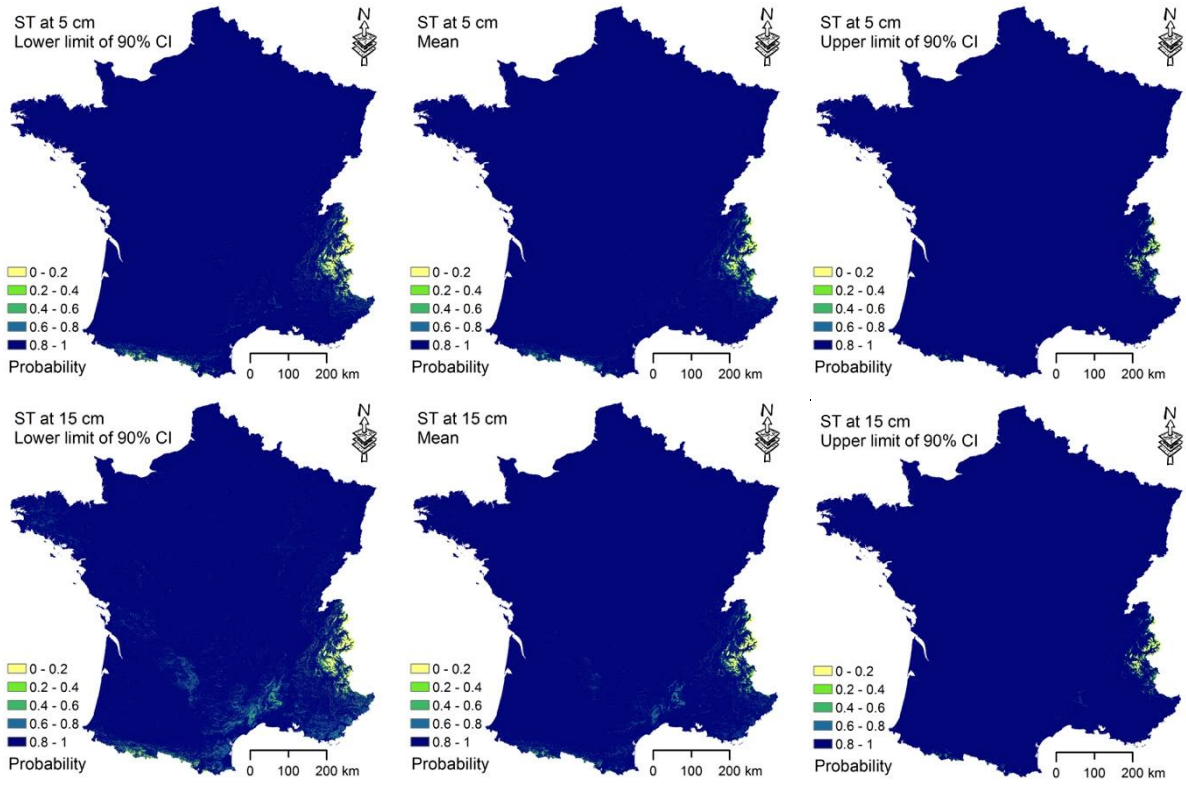


Figure 5.7 Probability maps of exceeding the six *GlobalSoilMap* standard depths (middle) and their associated 90% confidence intervals (left and right)

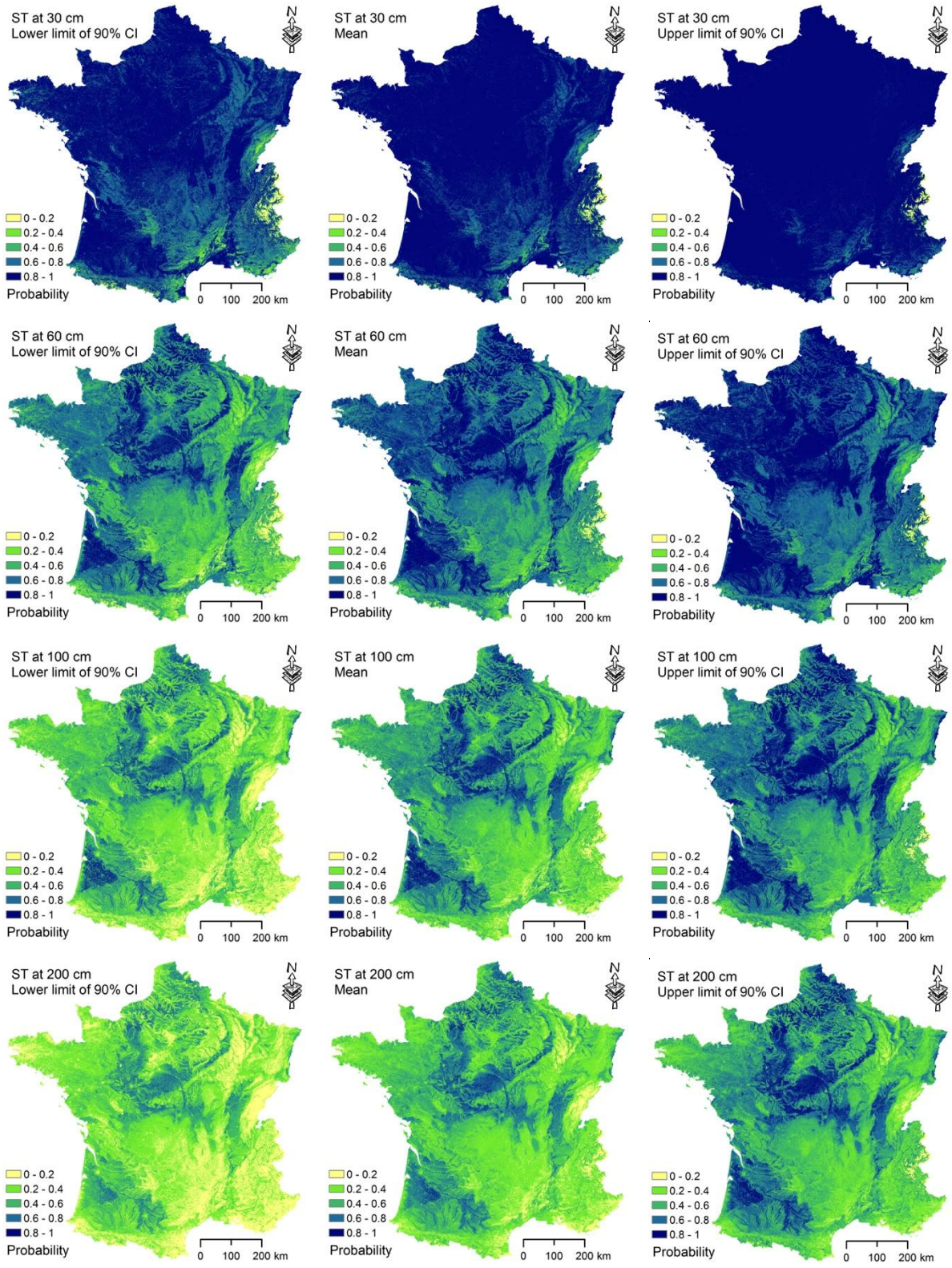


Figure 5.7 (continued) Probability maps of exceeding the six *GlobalSoilMap* standard depths (middle) and their associated 90% confidence intervals (left and right)

0.97, 0.88, 0.68, 0.51, and 0.42, respectively.

The probability of exceeding ST5 was close to 1 across the whole country, except for eastern (the Alps) and southwestern France (the Pyrenees). The 90% CI was very narrow (0.02 ± 0.06), indicating low model uncertainty and thus robust estimates for the ST5 map.

A similar spatial distribution was observed when ST increased to ST15. The low probability in southern France (the Massif Central) showed that this region had a high probability of having STs less than 15 cm. The difference between the lower and upper limits of the 90% CI was still low (0.05 ± 0.08), indicating a robust estimate. Moreover, the surface area percentages for the five probability intervals were also quite close to those of ST5 (Figure 5.8).

When the ST depth criteria was further increased to ST30, in addition to previously mentioned locations, low probabilities were found in eastern France (the Jura Mountains, Figure 5.7). Moreover, the CIs substantially increased (give numbers) compared to ST5 and ST15. This indicates a larger prediction uncertainty and a lower model robustness. In comparison with ST15, a slight increase (2%) was observed for the surface area with probabilities between 0.4 and 0.6. The surface area having a probability between 0.6 and 0.8 increased from 1 to 14%, while the area with a probability between 0.8 and 1 decreased from 98 to 83% (Figure 5.8).

Moving from the ST30 up to the ST200 thickness criteria, substantial changes in spatial patterns and the probability of surpassing the ST criteria became apparent. Most notable is how the surface area with probabilities between 0.8 and 1.0 continuously decreased, from 83% (ST30) to 2% (ST200). ST60 corresponded with a probability of 27% and ST100 with a probability of 7% (Figure 5.8).

For ST200, more than 50% of the territory of mainland France had a low probability (<0.4) of exceeding ST by 2 m, while less than 17% of the areas had a high probability (>0.6) of exceeding the ST by 2 m (Figure 5.7). The areas with a high probability were mainly located in southwestern France (the Landes of Gascony), central France (Sologne), and northern France (thick loess deposits).

5.4 Discussion

Several ways to perform probability mapping have been proposed in the literature

since the 1990s. For instance, Bell et al. (1994) applied discriminant analysis with a maximum-likelihood classification function to map the soil drainage probability in south-central Pennsylvania, USA. von Steiger et al. (1996) mapped the probability of exceeding the maximum tolerable heavy metal concentrations by Disjunctive Kriging in northeast Switzerland. Richer-de-Forges et al. (2017) used Logistic Regression Kriging in probability mapping of iron pan presence in sandy podzols in southwest France. The largest differences between the methods used in previous studies and RSF can be summarized in two aspects: 1) RSF is able to deal with right censored data while others are not, and 2) RSF can potentially produce probability estimates of any ST value, whereas other methods deal with presence/absence at a given threshold for the soil attributes of interest. Moreover, others used multiple sequential indicator simulations to model this type of distribution (e.g., Cattle et al., 2002). The survival analysis we used is a similar approach, except that it models the survival function rather than the empirical distribution function.

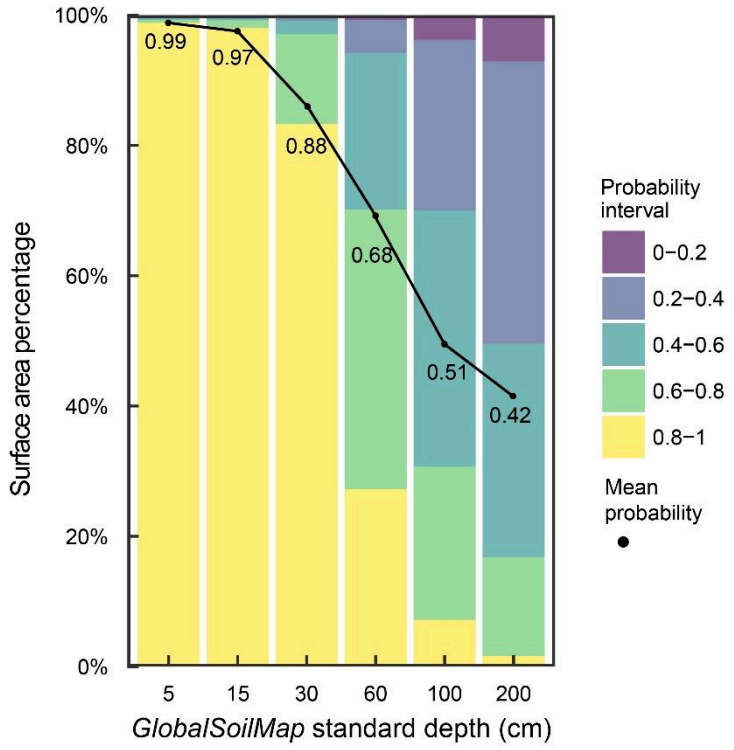


Figure 5.8 Surface area percentage of the probability of exceeding the ST of each *GlobalSoilMap* standard depths. The mean probability is calculated by averaging all the pixels in the probability map for each *GlobalSoilMap* standard depth

In the RMQS dataset, right censored ST observations entail more than half of the observations. Using them for ST modelling with traditional DSM approaches would result in highly underestimated ST estimates. Lacoste et al. (2016) proposed adding a fixed value of 30 cm to censored samples before modelling, which may help lower the underestimation but does not really solve the problem. Moreover, as actual ST values of these censored sites remain unknown, adding a fixed value may even add more noise to the data, and thus enlarging the prediction uncertainty. As shown in Figure 5.9, the probability of exceeding the observed ST for each censored RMQS site was mainly between 0.5 and 1, with a median value of 0.78. Thereafter, as outlined in the methodology Section 2.3.2, RSF makes use of the probability function derived from right censored information, thereby avoiding underestimating ST at these censored positions.

The mean probability of exceeding an ST of 100 cm across mainland France was 0.51 (Figure 5.8), which means that all locations have a 50% possibility of being observed with an ST thicker than 100 cm. This result implies that the median ST in mainland France is approximately 100 cm, which is in line with previous work by Lacoste et al. (2016), showing that 48 and 54% of surface areas were below 100 cm when using Gradient Boosting Modelling and Cubist models, respectively.

The results showed that the prediction performance decreased from 0 to 60 cm and subsequently increased up to 200 cm for censored RMQS sites (Table 5.2), implying that the predicted probability of exceeding a given ST from the RSF model is more reliable for extreme values (i.e., a thin ST or thick ST). Indeed, due to the soil forming conditions in mainland France, except for steep slopes in mountainous areas, very thin soils are quite rare, and thus the probability of exceeding a very thin ST is high. Conversely, very thick soils are concentrated in (former) depositional areas (valleys, aeolian sand, or loess deposits) that can be easily mapped using some of the covariates (e.g., parent material and terrain parameters). For the censored RMQS sites, the overall accuracies for ST100 and ST200 were approximately 0.5, in which a large percentage of thin ST samples were misclassified as being thick. This can be explained by the fact that we used observed ST of censored RMQS sites in calculating the confusion matrix. Consequently, we may overestimate the percentage of misclassification mentioned before and thus underestimate the overall accuracies of ST100 and ST200.

Parent material and climatic zones were the most important variables for predicting

ST in France using the bootstrapped RSF, but roughness, precipitation, elevation, slope, gravimetry, and NPP also substantially contributed to the ST model. These results are in line with previous findings reported by Lacoste et al. (2016). Lacoste et al. (2016) stated that the most important covariates of ST modelling in mainland France were soil properties, climate covariates and land use. Considering the variable importance and the variables acting as controlling factors for ST, parent material, climatic zones, precipitation, and gravimetry are direct drivers for the weathering process. Roughness, elevation, slope, and NPP are more related to sediment transport dynamics.

Future research should aim to derive an ST map using RSF, instead of the currently presented ST probability map of exceeding a given depth. There are three ways to determine the actual soil ST from the unique probability function produced by RSF for each location of interest: 1) use the ST extracted from the median probability in the predicted function; 2) use the ST extracted from a fixed probability, allowing the classification of censored and actual ST at high accuracy among RSF calibration datasets; 3) perform a derivative analysis on the probability curve. Moreover, it will be interesting to combine RSF with geostatistical methods. For example, kriging of residuals (Hengl et al., 2004) that are not captured by RSF and/or sampling optimizing for future campaigns to reduce the prediction variance at locations where it is highest. Alternatively, the presented probability maps can be used directly for additional ST sampling campaigns, aimed at ST modelling in mainland France. For example, the regions with a high probability (>0.8) of ST200 have a large chance of being censored. Integrating those high probability regions with parent material and climatic zones would yield an efficient and effective sampling design using conditioned Latin hypercube sampling (cLHS, Minasny and McBratney, 2006) to obtain more representative samples of all physiographic contexts. RSF is able to provide a probability at any depth and thus will be helpful for decision making in geotechnical engineering regarding, for example, laying out drains, pipes, and tubes (Zhang et al., 2005).

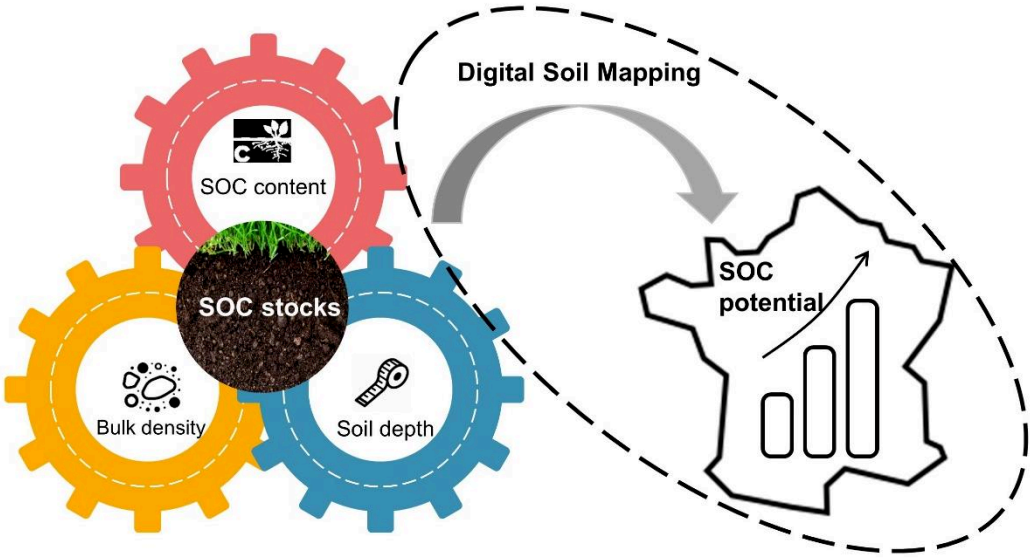
5.5 Conclusions

This study introduced the use of RSF in ST probability modelling to deal with right censored data for Digital Soil Mapping. RSF produced a probability function of ST for each soil sample included in the database. This function allowed the estimation of a

probability of exceeding a given ST, indicating each soil location was right censored or not. Robust estimates were made by bootstrapping the RSF model to quantify an averaged bootstrap prediction and 90% CI for each *GlobalSoilMap* standard depth (5, 15, 30, 60, 100 and 200 cm) using the RSF survival probability functions. The model evaluation indicated an overall good performance (overall accuracy from 0.546 to 0.989) of RSF to predict the probability exceeding the six *GlobalSoilMap* standard depths. The RSF proved suitable for using right censored soil data for digital soil mapping, and thereby this work introduced a new approach capable of using both right censored and actual data for modelling ST accordingly.

Chapter 6

Fine resolution map of top- and subsoil carbon sequestration potential in France



Chen, S., Martin, M.P., Saby, N.P.A., 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.

6.1 Introduction

The Paris Climate Agreement reached at the COP21 aims at limiting global warming to 2°C above pre-industrial levels before the end of the century. To achieve this goal, global annual emissions need to be limited at 9.8 Gt C at 64% probability (Meinshausen et al., 2009). Soil organic carbon (SOC) sequestration can make a significant contribution to offset CO_2 increase in the atmosphere by transferring it into long-lived soil C pools (Lal, 2004; Paustian et al., 2016). Consequently, at the COP21, the initiative "4 per 1000 carbon sequestration in soils for food security and the climate" (4 per 1000, https://www.4p1000.org/understand) was launched with an expectation to increase global SOC stocks by 0.4% y^{-1} as a compensation for global GHG emissions (Minasny et al., 2017; Soussana et al, 2015). The 4 per 1000 initiative also states that increasing SOC contributes to combat soil degradation, increases food security and enhances agriculture adaptation to climate change (Soussana et al., 2015). In order to achieve the 4 per 1000 target, the annual soil sequestration rate should be 0.6 t C ha^{-1} y^{-1} globally, and could be achieved largely by restoring and improving degraded agricultural lands and changes in crop rotations and agricultural practices (Batjes and Sombroek, 1997; Dignac et al., 2017). This soil sequestration rate cannot be reached everywhere due to the high spatial heterogeneity of SOC stocks and sequestration potential but global studies suggested that 0.2 to 0.5 t C ha^{-1} y^{-1} is feasible at many locations in the world (Paustian et al., 2016; Minasny et al., 2017). Being constrained by agronomic, economic and social challenges (e.g., need for dramatic changes in crop management, tradeoffs with agricultural production), the feasibility of achieving the 4 per 1000 target may be questionable (Paustian et al., 2016; Zomer et al., 2017).

It is generally accepted that there is an upper limit of soil stable C storage, which is referred to as SOC saturation (Hassink, 1997; Six et al., 2002; Stewart et al., 2007). Organic C saturation mainly depends on the intrinsic soil potential to stabilize soil organic matter (SOM) against microbial mineralization, though non-microbial degradation also matters during tillage (Baldock and Skjemstad, 2000; Balesdent et al., 2000). Mechanisms responsible for C stabilization in soils are diverse, variable and still not fully understood. However, the fine mineral fraction is considered to play a major role (Baldock and Skjemstad, 2000; Arrouays et al., 2006) and used as a proxy for soil C

stabilization potential (Hassink, 1997). Hassink (1997) proposed an equation to describe the relationship between stable C saturation and the soil fine fraction (<20 μm, clay and fine silt) using a statistical approach based on a wide range of topsoils from temperate and tropical regions. The C saturation deficit or sequestration potential can be calculated as the difference between the theoretical C saturation and the actual SOC stored in the fine fraction. This equation has been used in several studies to calculate sequestration potential at regional or national scales (Angers et al., 2011; Wiesmeier et al., 2014b). Angers et al. (2011) estimated the sequestration potential of agricultural topsoils in France based on 1.5 million legacy soil data from soil tests requested by farmers and then mapped them at the administrative unit level. Wiesmeier et al. (2014b) estimated the sequestration potential of topsoil in southeast Germany and quantified the total sequestration potential stocks based on the bulk density and land area under different land covers. These previous studies used a relatively coarse resolution and did not consider the subsoil. Because of their generally lower SOC content, subsoil horizons are generally believed to offer a large potential for C sequestration (Lorenz and Lal, 2005). However, the C saturation deficit of subsoil horizons has seldom been estimated (Castellano et al., 2017; Reis et al., 2014), and to our knowledge, never mapped. In order to improve land management and identify the locations with high potential to sequester C, it is necessary to develop a better understanding and detailed spatial distribution of SOC sequestration potential at national scale, including the subsoil horizons.

The objectives of this study were three folds:

(1) Determine the SOC sequestration potential for topsoil and subsoil in France;

(2) Build prediction models of SOC sequestration potential for topsoil and subsoil based on relationships with soil-forming environmental covariates;

(3) Produce high resolution maps of SOC sequestration potential for topsoil and subsoil.

6.2 Materials and methods

6.2.1 Site specific soil data

The soil data used in this study were obtained from 2,092 sites from the first

campaign of the French soil monitoring network (RMQS) between 2001 and 2009 (Jolivet et al., 2006), which covers entire metropolitan France (around 550,000 km²) including different soil, climate, relief and land cover conditions (Figure 6.1). The RMQS is based on a 16 km × 16 km square grid and all sites are selected at the center of each grid cell. When sampling the exact location was not possible, a site was selected as close as possible to the grid center (Martin et al., 2011). On basis of a unaligned sampling design with a 20 m × 20 m square, 25 individual core samples were collected from topsoil (0-30 cm) and subsoil (30-50 cm) by a hand auger. These individual core samples were mixed into a composite sample for each soil layers. Then composite samples were air-dried (controlled at a temperature of 30 ºC and an air-moisture of 30%) and sieved to 2 mm before laboratory analysis at Soil Analysis Laboratory of INRA in Arras, France. Apart from these composite samples, a soil pit was dug at 5 m from the south border of the 20 m × 20 m square, from which the main soil characteristics were recorded and six bulk density measurements were collected, three within the topsoil layer and three within the subsoil layer (Martin et al., 2009). The topsoil thickness was taken as 30 cm for forest and pasture soils, and deepest tillage depth for arable soils. For some sites, soils were so thin that subsoil did not exist. SOC was determined by dry combustion using a CHN elemental analyzer (Thermofisher NA2000). Particle-size distribution was determined for clay (0-2 μm), fine silt (2-20 μm), coarse silt (20-50 μm), fine sand (50-200 μm) and coarse sand (200-2000 μm) by the pipette method (NCRS, 2004).

6.2.2 Calculation of C saturation and sequestration potential

The C saturation of particle-size < 20 μm was calculated according to the equation proposed by Hassink (1997):

$$C_{sat}=4.09+0.37\times FineFraction \quad (6.1)$$

where C_{sat} is the C saturation (g kg^{-1}) and *FineFraction* is the content of particle-size < 20 μm (%).

As the C saturation deficit is calculated by the difference between C saturation and the measured C of fine fraction (C_{fine}), an approach for estimating the C_{fine} from the total SOC in our database had to be developed. Based on previously published data (Angers et al., 2011; Balesdent 1996; Jolivet et al., 2003), the C_{fine} content was assumed to comprise 85% of the total SOC in cultivated topsoil (cropland and

vineyard/orchard). To derive more reliable C_{fine} proportions in total SOC for forest, grassland in both topsoil and subsoil, we gathered a few existing data from France, summarized related studies from countries with similar climate to France, and assigned weighted average values for topsoil and subsoil under different land uses (Table 6.1). Limited by available data sources, the definition of fine fraction varied from 0-20 μm to 0-63 μm, but there were no significant differences between them (Balesdent et al., 1998; McNally et al., 2017). In the end, C_{fine} of forest and grassland topsoil was assumed to comprise 66% and 69% of the total SOC while values were 75%, 86% and 93% for forest, grassland and cultivated subsoil respectively. Averaged values from aforementioned land uses were used for other land uses in topsoil (73%) and subsoil (85%).

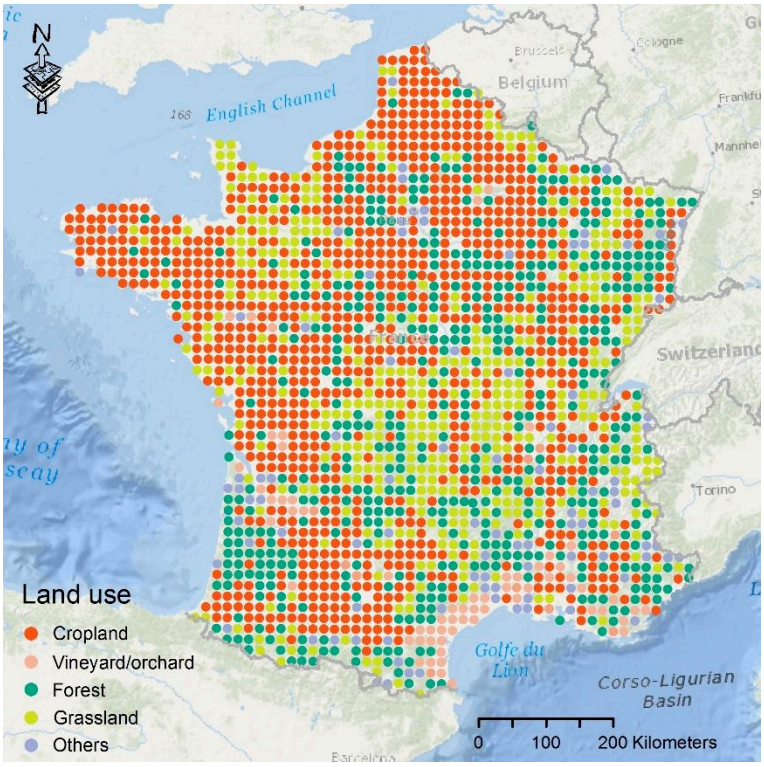


Figure 6.1 Distribution of RMQS sites in mainland France

The C saturation deficit was calculated as follows:

$$C_{sd}=C_{sat}-C_{fine} \quad (6.2)$$

where C_{sd} is the C saturation deficit (g kg^{-1}) and C_{fine} is C of fine fraction (g kg^{-1}). Our

Table 6.1 Fine fraction SOC proportion in total SOC content for topsoil and subsoil under different land use

Land use	Country	Sampling depth (cm)	Number of sampling sites	0-20 μm[a]	0-50 μm[b]	0-53 μm[c]	0-63 μm[d]	Reference
Topsoil								
Forest	Canada	0-15	1			69%		Carter et al. (1998)
	France	0-30	1	66%	67%			Balesdent et al. (1998)
	France	0-24	5		68%			Jolivet et al. (2003)
	Germany	0-24(30)	2	65%				Rumpel et al. (2004)
	Germany	0-25	1			56%		John et al. (2005)
	Germany	0-20	14				66%	Wiesmeier et al. (2014c)
Grassland	USA	0-20	2			68%		Cambardella and Elliott (1992)
	Canada	0-15	2			65%		Carter et al. (1998)
	USA	0-20	4			80%		Conant et al. (2003)
	Belgium	0-20	6		58%			Accoe et al. (2004)
	France	0-30	2		85%			Chenu et al. (2004)
	Germany	0-30	1			88%		John et al. (2005)
	Switzerland	0-20	2				60%	Leifeld and Fuhrer (2009)
	UK	0-18	1		69%			Coppin et al. (2009)
	Germany	0-20	11				70%	Wiesmeier et al. (2014c)
Subsoil								
Cropland	Spain	30-40	3			86%		Álvaro-Fuentes et al. (2008)
	France	28-44	1		95%			Moni et al. (2010)
	France	22-61	1		92%			Moni et al. (2010)
	Canada	30-70	1		91%			Poirier et al. (2014)
	USA	40-60	5			97%		Beniston et al. (2014)
Forest	Germany	24-50	1	70%				Rumpel et al. (2004)
	Germany	30-55	1	78%				Rumpel et al. (2004)
	Germany	25-40	1			77%		John et al. (2005)
Grassland	USA	20-40	4			89%		Conant et al. (2003)
	China	30-60	4	82%				Steffens et al. (2011)

[a] proportion of SOC in 0-20 μm among total SOC content; [b] proportion of SOC in 0-50 μm among total SOC content; [c] proportion of SOC in 0-53 μm among total SOC content; [d] proportion of SOC in 0-63 μm among total SOC content.

assumption about the proportion of fine fraction SOC might result in negative C saturation deficit, especially for locations with high C content.

The C sequestration potential density (or saturation deficit density) was calculated using the following equation:

$$C_{spd}=p\times C_{sd}\times BD\times(100-ce)\times10^{-2} \quad (6.3)$$

where C_{spd} is the C sequestration potential density (kg m^{-2}) in topsoil layer (0-30cm) or subsoil layer (30-50 cm), BD, C_{sd} and ce are the bulk density (kg m^{-3}), C saturation deficit (g kg^{-1} or ‰) and percentage of coarse elements (%) in these horizons, and p is the thickness of these horizons (m) within topsoil or subsoil. Besides, C density of fine

fraction ($C_{fine\text{-}den}$) and C saturation density ($C_{sat\text{-}den}$) were calculated by similar equations.

The Degree of C sequestration potential or C saturation deficit (%) was defined using the following equation:

$$Degree C_{spd} = \frac{C_{spd}}{C_{sat\text{-}den}} * 100 \quad (6.4)$$

6.2.3 Digital soil mapping approach

In 1940s, Jenny (1941) proposed the well-known soil forming equation which was later extended by McBratney et al. (2003) named *scorpan*-SSPFe (soil spatial prediction function with spatially autocorrelated errors). This method fits quantitative relationships between soil properties or classes and seven *scorpan* factors, which can be written as:

$$S = f(s,c,o,r,p,a,n) + e \quad (6.5)$$

where S is soil classes or soil properties. The s refers to soil information either from prior maps, or from remote or proximal sensing data. The c is climatic properties of the environment at a point. The o is organisms including vegetation or fauna or human activity. The r refers to relief. The p is parent material or lithology. The a is age, which is regarded as time factor. The n refers to space or spatial position. The e is spatially correlated residual.

The *scorpan*-SSPFe method has been widely used in mapping various soil properties (e.g., SOC, pH, soil texture) from local scale to global scale (e.g., Hengl et al., 2017; Malone et al., 2011; Minasny et al., 2006; Mulder et al., 2015; Viscarra Rossel et al., 2014).

6.2.4 Scorpan covariates

The covariates that are at the same time responding to *scorpan* model at fine level and available at good resolution for the study area are listed in Table 6.2. The covariates provided information related to five *scorpan* factors including climate, organisms, soil, parent material and topography. The spatial position *n* and spatially correlated residual *e* were taken into account in Kriging phase (details in section 6.2.5). Due to different resolution/scale on original covariates, data pre-processing was performed in two steps: (1) reprojection of the coordinate system to Lambert 93 (official projection for mainland France); (2) resampling covariates into 90 m resolution raster images using a nearest

neighbour interpolation. We selected 90 m resolution because it is the target resolution suggested by the *GlobalSoilMap* consortium for mapping selected soil attributes (Sanchez et al., 2009; Arrouays et al., 2014a).

Table 6.2 Covariates used for modelling C sequestration potential density

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

*Soil type is defined by World Reference Base (WRB)

6.2.5 Spatial predictive modelling

To construct the spatial predictive model between the C sequestration potential density and *scorpan* covariates, we used an ensemble learning method Random Forests (RF, Breiman, 2001). As described by Breiman (2001), RF is applicable to regression and classification and it consists of multiple trees generated by a combination of bagging and random selection of features applied at each split of the trees. The final prediction result is the mean of the outputs of all trees when RF is applied in the regression modelling. RF is rather robust to noise and irrelevant features, which makes it a favorable choice for soil property modelling (Viscarra Rossel and Behrens, 2010).

RF is able to provide model variable importance, which means it may rank controlling factors of the variate of interest (C sequestration potential density in our case). Firstly, for each tree, the mean square error (MSE) is calculated using so-called

out-of-bag (OOB) data, which is a random subset of the data that is not used in the bagging approach. Then the same procedure is calculated again after permuting a variable. The differences are averaged and normalized by the standard error. A more important covariate associates with a larger difference (Liaw et al., 2002).

We used the RF implementation provided by the package *randomForest* in R (Liaw et al., 2002; R Core Team, 2016). Three parameters should be defined in RF model: the number of trees to grow (n_{tree}), the number of variables randomly sampled as candidates at each split (m_{try}), and the minimum size of terminal nodes (*nodesize*, Liaw et al., 2002). The default values were used for n_{tree} and *nodesize*, which were 500 and 5 respectively. The optimal value of m_{try} was tuned to 2 by the lowest OOB error estimate.

Defined as *e* in *scorpan* model, the residuals between the RF predictions and the measured values were spatially correlated and they could be predicted too. We treated the residuals (ϵ) as spatially correlated variables with a mean of zero and a variogram defined as follows (Cressie, 1993):

$$\gamma(\mathbf{h})=\frac{1}{2}\mathrm{var}[\epsilon(\mathbf{u})-\epsilon(\mathbf{u}+\mathbf{h})]=\frac{1}{2}\mathrm{E}[\langle\epsilon(\mathbf{u})-\epsilon(\mathbf{u}+\mathbf{h})\rangle^2] \quad (6.6)$$

where $\epsilon(\mathbf{u})$ and $\epsilon(\mathbf{u}+\mathbf{h})$ are random variables (residuals in our case) at positions $\mathbf{u}$ and $\mathbf{u}+\mathbf{h}$ separated by lag distance $\mathbf{h}$, and $\mathbf{E}$ refers to expectation. In this paper we assume that the function is isotropic and varies only according to the length of $\mathbf{h}$ which we denote h.

As suggested by Matheron (1971), an empirical variogram $\hat{\gamma}(h)$ was applied to estimate the theoretical semivariance $\gamma(h)$:

$$\hat{\gamma}(h)=\frac{1}{2n(h)}\sum_{i=1}^{n(h)}[\epsilon(\mathbf{u}_i)-\epsilon(\mathbf{u}_i+h)]^2 \quad (6.7)$$

where $\epsilon(\mathbf{u}_i)$ and $\epsilon(\mathbf{u}_i+h)$ are the residuals at positions $\mathbf{u}_i$ and $\mathbf{u}_i+h$, and $n(h)$ is the number of comparisons with a lag distance $\mathbf{h}$. Based on this empirical variogram, ordinary Kriging was used to predict residuals on mainland France. The final variograms were fitted by spherical variogram model after comparing spherical, circular and exponential models with cross-validation statistics (Saby et al., 2006). R package *gstat* (Pebesma and Graeler, 2013) was used to select variogram models and perform Kriging procedure. The final prediction summed the prediction results from RF predictions and

the Kriging outputs of residuals. This method was also named regression Kriging approach in geostatistics.

Model performance was evaluated by the 10-fold cross-validation results of RF models with three commonly suggested indices: the root mean square prediction error (RMSPE), the prediction coefficient of determination (R^2) and Lin's concordance coefficient (CC).

A good prediction usually has high R^2 and CC values, and low RMSPE value. All the modelling and computation were performed in R (R Core Team, 2016).

6.2.6 Map correction and calculation of C sequestration potential stocks

The negative value (oversaturated position) in the final maps of C sequestration potential density was replaced by 0 for we postulated that oversaturated regions had no additional sequestration potential. Besides, soil depth was taken into consideration as it is a determining factor in calculating stocks. The newest soil depth map (Figure S6.1) for France at 90 m resolution was used (Lacoste et al., 2016). For these locations with soil depths shallower than 50 cm, the C sequestration potential densities were adjusted by the ratio of actual soil thickness to the layer thickness (30 cm for topsoil and 20 cm for subsoil) within topsoil or subsoil layers.

Using corrected C sequestration potential density maps, we calculated C sequestration potential stocks for each land use by summing up the predicted stocks of all 90 m × 90 m grids corresponding to a given land use. These stocks are defined as model-based estimates. As suggested by Marchant et al. (2015), we also calculated the design-based estimator from observed RMQS sites, which provided an unbiased estimation of stocks. Design-based estimates of C sequestration potential stocks implied the use of the Horvitz-Thompson estimator (Brus and Saby, 2016, de Gruitjer et al., 2006). In this case, it corresponded to the multiplication of the total area by arithmetic mean C sequestration potential stocks (negative values were replaced by 0 as we did for mapping during the calculation of arithmetic mean values) of observed RMQS sites under a given land use.

6.3 Results

6.3.1 Observed C sequestration potential density

Figure 6.2 lists the designed-based estimates of the densities of the fine fraction C

stocks, C saturation, C sequestration potential and degree of C saturation deficit for topsoil and subsoil under different land covers according to Corine Land Cover 2006. The current topsoil C densities of the fine fraction were quite variable within different land covers while narrower inter-quantile ranges (IQR) were found in subsoil. In topsoil, forest and grassland showed high C density (mean and standard error at 6.62±0.11 and

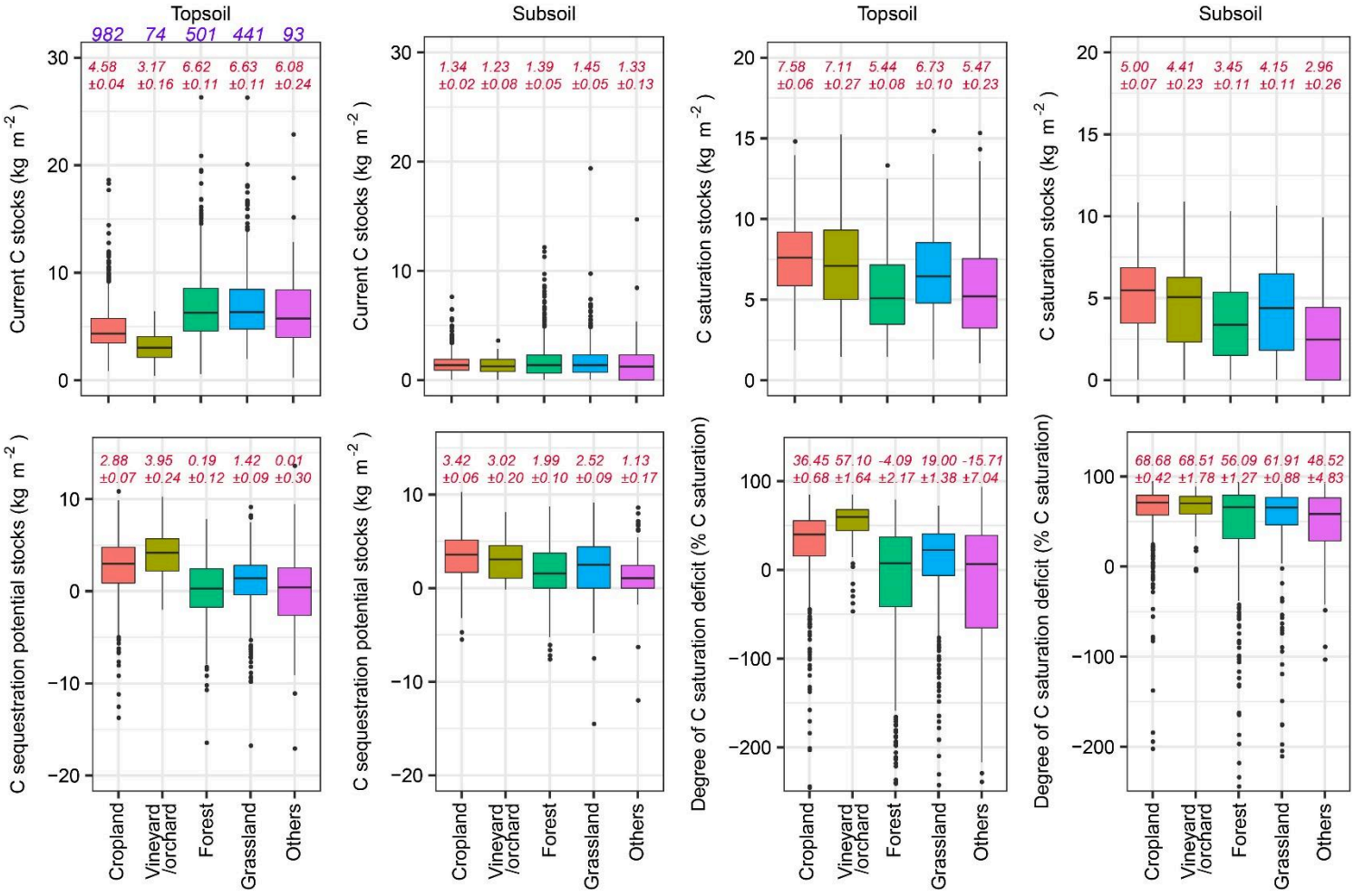


Figure 6.2 Density of current C of fine fraction, C saturation and C sequestration potential for topsoil and subsoil under different land covers. In histogram, the lower and upper hinges correspond to Q25 and Q75, the upper/lower whisker extends to the largest/smallest value no further than 1.5 times of inter-quartile range (Q75–Q25) from the upper/lower hinges. Blue italic numbers indicate the number of sites under different land covers. Red italic numbers show the mean and standard error of the mean (Geary's C variance approximation is used, Brus and Saby, 2016) which are calculated without the data beyond the end of the whiskers.

6.63±0.11kg m^{-2}, respectively) with high range varying between first (Q_{25}) and third quartile (Q_{75}), cropland and vineyard/orchard had lower C density (4.58±0.04 and 3.17±0.16 kg m^{-2}, respectively) with relatively low inter-quartile range. For other land uses, both high C density (6.08±0.24 kg m^{-2}) and high IQR were found. Compared with topsoil, the C density of subsoil was much lower and similar mean C densities (between 1.23 and 1.45 kg m^{-2}) were found under all the land uses. In topsoil, higher C sequestration potential densities were found in cropland and vineyard/orchard (2.88±0.07 and 3.95±0.24) than under forest and grassland (0.19±0.12 and 1.42±0.09). As shown in Figure 6.2, with median C sequestration potential density around 0 kg m^{-2}, nearly half of forest and other land cover soils were oversaturated in topsoil and they almost had no potential to sequester additional stable C. Compared with the C sequestration potential density in topsoil, subsoil showed a larger potential for C sequestration within 20 cm. Cropland and vineyard/orchard subsoils had high C sequestration potential density (3.42±0.06 and 3.02±0.20 kg m^{-2}) while forest and grassland also showed some potential to sequester C (1.99±0.10 and 2.52±0.09 kg m^{-2}). Vineyard topsoil had a highest degree of C saturation deficit (57.10±1.64%) and cropland ranked second (36.45±0.68%). Subsoil under all land uses showed greater mean degree of C saturation deficit (from 48.52% to 68.68%) under all land uses than topsoil, especially for forest and grassland.

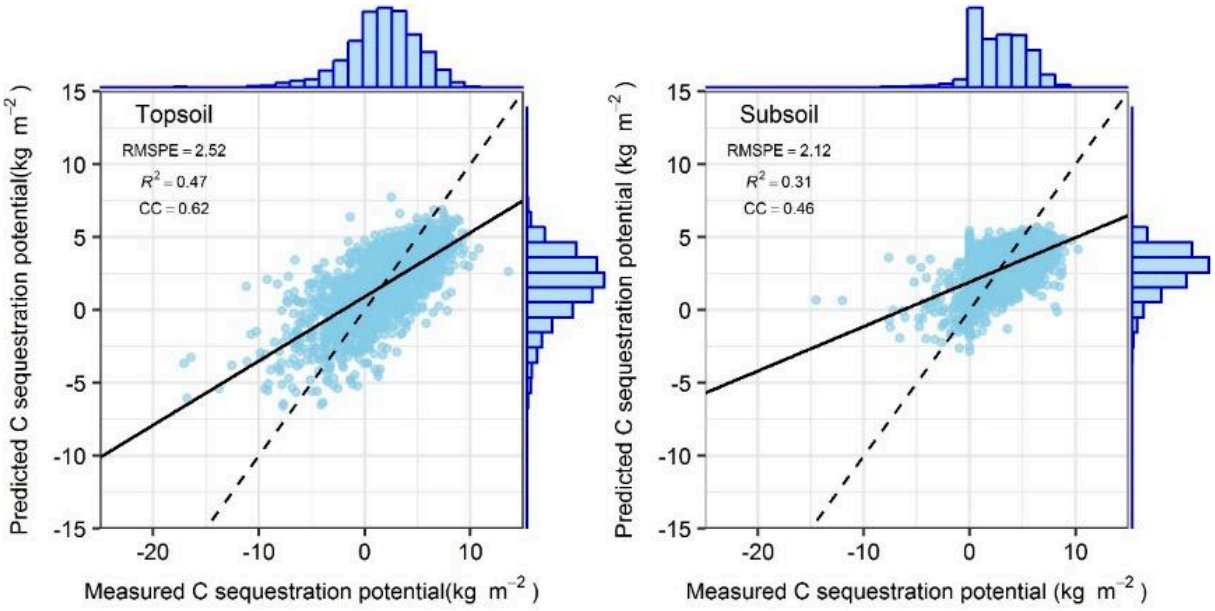


Figure 6.3 Scatter plots of C sequestration potential density for topsoil and subsoil using 10-fold cross-validation. Solid line is fitted line and dashed line is 1:1 line.

6.3.2 Performance of spatial predictive models

Predictive performances of RF models for topsoil and subsoil are given in Figure 6.3. RF models yielded good results in both topsoil and subsoil. With a higher R^2 (0.47) and CC (0.62), C sequestration potential density in topsoil was better predicted than for subsoil. The slopes of fitted line were <1 indicating that RF models slightly overestimated the low C sequestration potential density and underestimated high C sequestration potential density in both topsoil and subsoil.

Figure 6.4 presents variograms of residuals fitted by spherical variogram model. The semivariances reached their plateau (sill values) at around 114 km for topsoil and 97 km for subsoil. High nugget values in variograms suggested that most of the long range spatial structure of C sequestration potential had been captured by the RF models and that residuals were mainly characterized by variations at short distance, not captured by the RMQS sampling design with minimum distances of 16 km between points (Martin et al., 2014). The variance of the residuals was higher for topsoil than for subsoil.

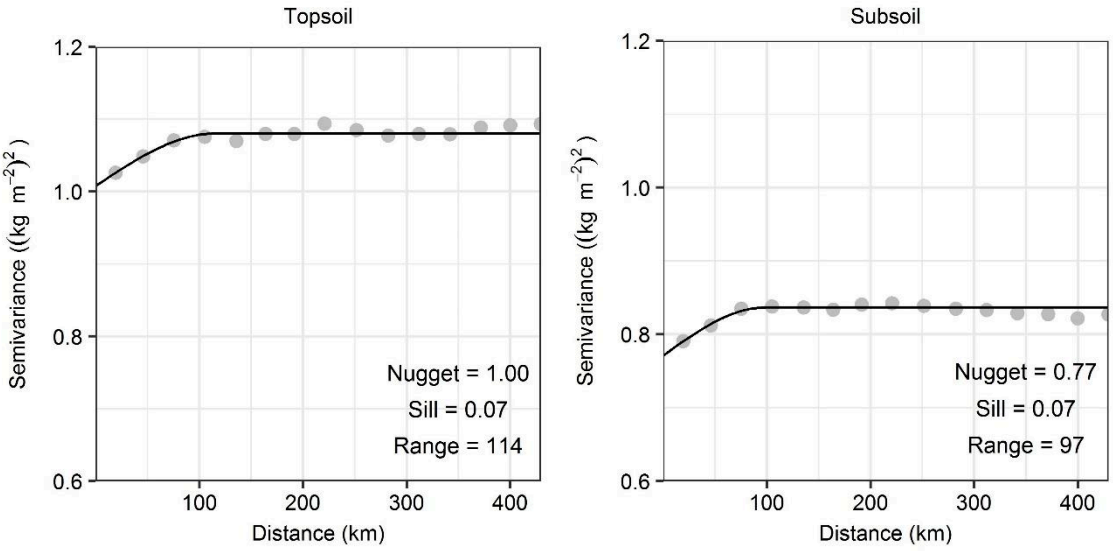


Figure 6.4 Variograms of residuals of C sequestration potential for topsoil and subsoil

6.3.3 Variable importance in predictive spatial models

The variable importance in RF models for topsoil and subsoil are listed in Figure 6.5. In topsoil, the most important controlling factor of C sequestration potential density was

land cover (36%). Parent material, NPP, elevation, MAT and MAP also had strong influence on C sequestration potential density (increased MSE between 20% and 30%). For subsoil, parent material ranked first in variable importance (increased MSE at 31%) in RF model. Besides, elevation, land cover, MAP, roughness and NPP were also important in C sequestration potential modelling with quite close importance (increased MSE between 18% and 22%). Curvature, exposition and aspect made little contribution in both topsoil and subsoil.

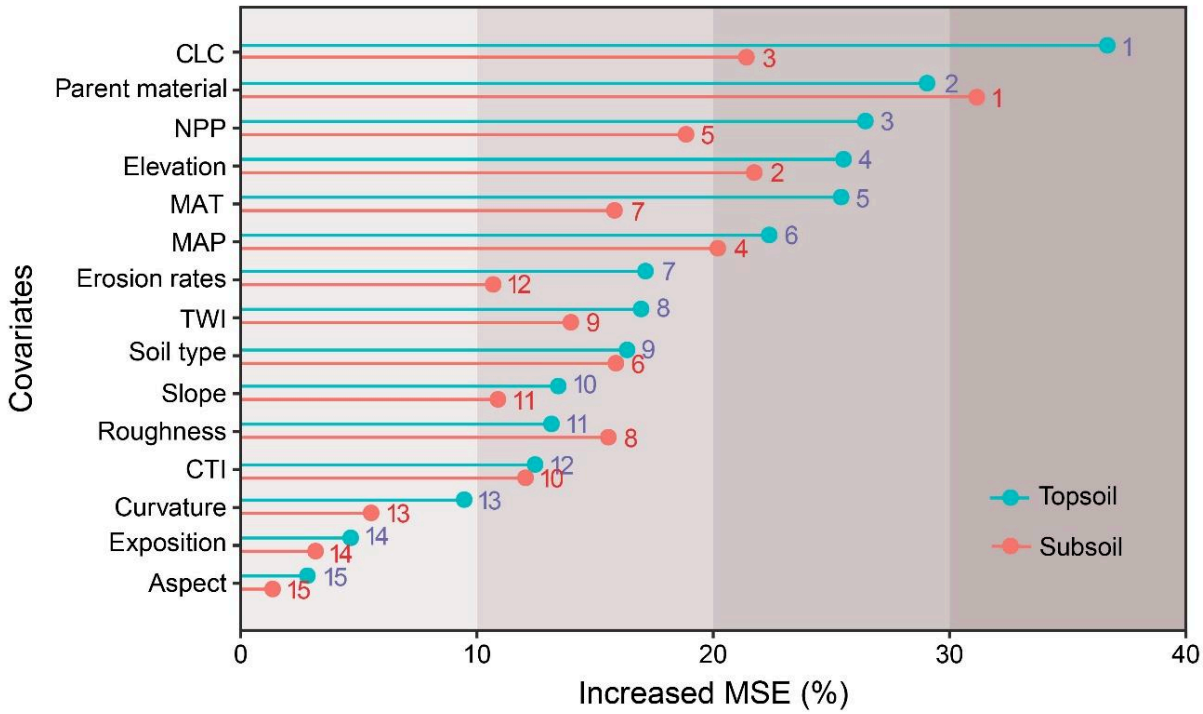


Figure 6.5 Variable importance of RF models for topsoil and subsoil. The ranks of variable importance for all the covariates are provided.

6.3.4 Spatial distribution of C sequestration potential

Figure 6.6 shows the spatial distribution of C sequestration potential density for topsoil. The highest C sequestration potential density (> 5 kg m^{-2}) was observed in intensively cultivated plains of the northern half and the southwestern part of France (Figure S6.2), in vineyards and orchards of the Mediterranean region and along the Rhône valley and in some regions dominated by clay-rich soils (Charentes and Lorraine). Central and western France had relatively low C sequestration potential

density (1-3 kg m^{-2}). There was no C sequestration potential in mountainous areas (Vosges, Jura, Massif Central, Alps and Pyrénées).

Figure 6.7 shows the C sequestration potential density map for subsoil. Different from topsoil, a large percentage of subsoil in mountainous areas and western Brittany had some potential to sequester C (0-2 kg m^{-2}). Higher C sequestration potential density (>5 kg m^{-2}) were observed in all intensively cultivated areas of France and for vineyards and orchards. A large area in Lorraine had low potential because of shallow soils.

Overall, subsoil had higher C sequestration potential density than topsoil in most regions (Figure 6.8). No sequestration potential was observed in topsoil for 21% of the country, whereas for subsoils this non sequestration potential area covered only 10% of the country. About 40% of topsoil and only 20% of subsoil had C sequestration potential density below 1 kg m^{-2}.

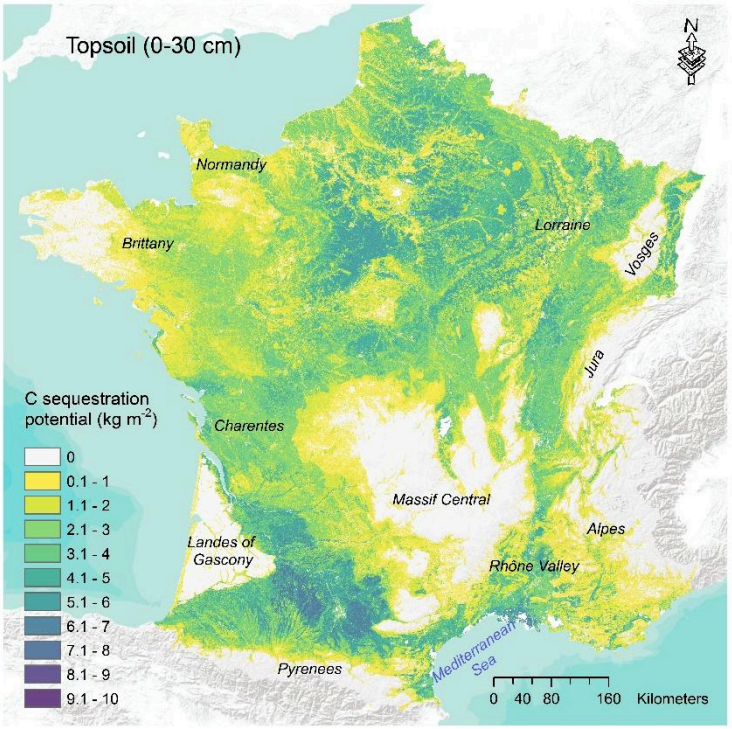


Figure 6.6 Map of C sequestration potential for topsoil (0–30 cm) in mainland France.

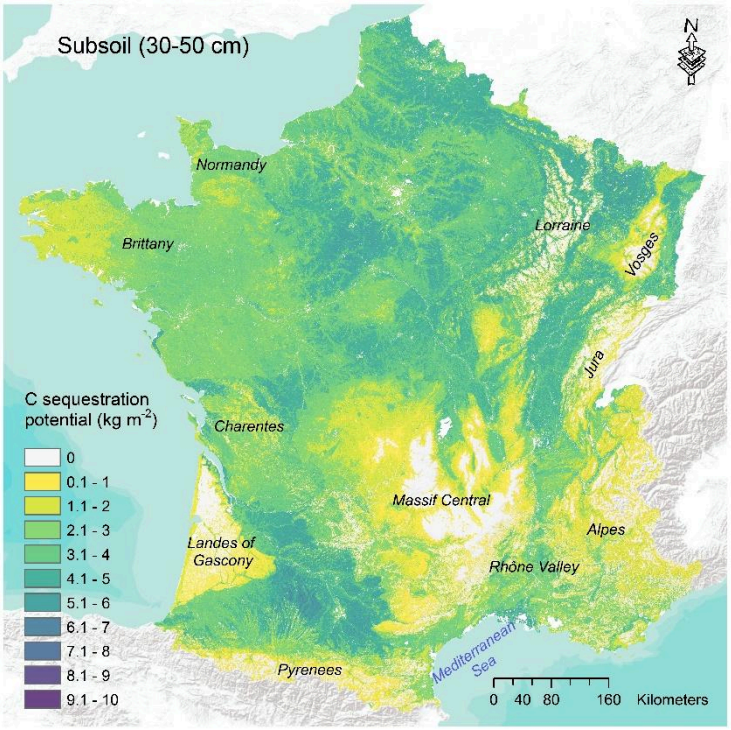


Figure 6.7 Map of C sequestration potential for subsoil (30–50 cm) in mainland France.

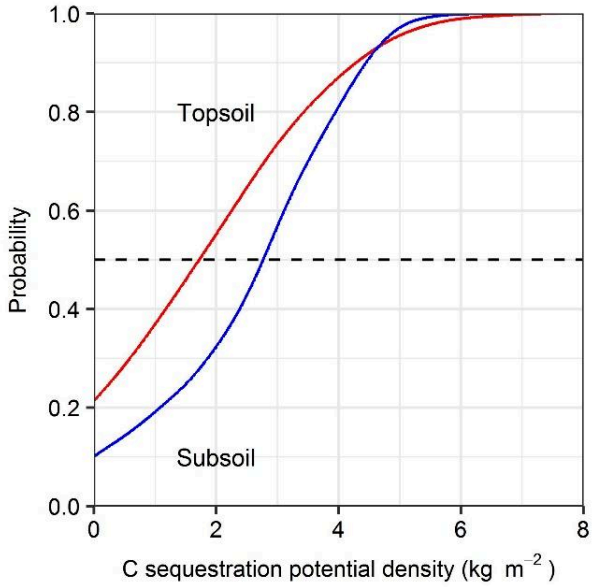


Figure 6.8 Cumulative probability distributions of C sequestration potential. Red line for topsoil and blue line for subsoil. Dashed line is at 50% probability.

6.3.5 Model-based and design-based estimates of C sequestration potential stocks

Table 6.3 lists model- and design-based estimates of C sequestration potential stocks for topsoil and subsoil in mainland France. Because of map artefacts, the total mapped area of mainland France was 527,159 km². Cropland accounted for about 43% of the total area, followed by forest (26%), grassland (19%), other land uses (10%) and vineyard/orchard (2%).

Carbon sequestration potentials were 1,008 Mt or 1,183 Mt C according to model or design-based estimates in French topsoils. Cultivated topsoil (cropland and vineyard/orchard) had represented a high proportion of this potential about 69% and 67% in model- and design-based estimates while proportions for forest and grassland topsoil were 25% and 30% in model-and design-based estimates. Despite their smaller layer thickness (20 cm), French subsoils showed high C sequestration potential stocks (1,360 Mt and 1,455 Mt) under almost all land uses. Forest and grassland subsoil showed a higher percentage (35% and 38% in model- and design-based estimates) than topsoil. Overall, based on model- or design-based estimates, about 2.4 Gt or 2.6 Gt C could be theoretically sequestered in the first 50 cm soils in mainland France.

Table 6.3 Model- and design-based estimates of C sequestration potential stocks.

Land cover	C sequestration potential stocks (Mt)						Area (km²)
	Topsoil (0-30 cm)		Subsoil (30-50 cm)		Total (0-50 cm)		
	Model-based	Design-based	Model-based	Design-based	Model-based	Design-based	
Cropland	646	692	752	774	1,398	1,466	225,506
Vineyard/orchard	49	53	41	40	90	93	13,116
Forest	120	180	237	297	357	477	138,572
Grassland	132	180	238	256	370	436	97,736
Others	61	78	92	88	153	166	52,049
Total	1,008	1,183	1,360	1,455	2,368	2,638	527,159

6.4 Discussion

6.4.1 The calculation of C sequestration potential density

The C sequestration potential density in soils of mainland France was calculated from an estimate of the current C content in the fine fraction, the C saturation equation of Hassink (1997), bulk density and coarse elements. Despite its wide usage, the concept of saturation deficit by Hassink (1997) has been criticized for the fact that it has never been truly validated and its relationship with land cover change and management practices remains not evaluated (Barré et al., 2017; O'Rourke et al., 2015). However, no other alternative indicator, which would be rapid and inexpensive to enable testing over a large range of soil properties, could be used to estimate C sequestration potential (Dignac et al., 2017). Therefore, using the saturation deficit as per Hassink (1997) is the best possible approach at this stage to estimate C sequestration potential at national scale before a better indicator is created.

The proportion of fine fraction C was estimated by weighted averaging values under each land uses for topsoil and subsoil. We acknowledge that the choice of these values can strongly influence the estimates of SOC_{sp} and bring uncertainties into the estimates.

6.4.2 The C sequestration potential density for topsoil and subsoil under different land covers

The observed differences in the C sequestration potential of topsoils reflect a well-known effect of land use, especially on the carbon depletion with cultivation which results in higher C sequestration potential densities (Figure 6.2). Similar results were found in Germany showing higher C sequestration potential density in cultivated soils than forest and grassland soils (Wiesmeier et al., 2014b). This C depletion of the fine fraction in cultivated soils is usually attributed to the breakdown of soil aggregates due to tillage and the consequent loss of physically-stabilized SOM, to lower C inputs from crops, including roots, and biomass exportation (Post and Kwon, 2000; Six et al., 2000; Wiesmeier et al., 2014b).=

In subsoil, a higher C sequestration potential density than in topsoil was found under all land covers. This result is consistent with the speculation that subsoil might be far from being saturated with C (Kell, 2012). The high potential in subsoil may be

due to the fact that C inputs are lower and come mainly from translocation from topsoil (Lorenz and Lal, 2005). Being less influenced by human activities, subsoil showed much less difference in C densities among land covers. However, cropland and vineyard/orchard still showed slightly higher C saturation density than forest and grassland because of the combined effects of soil depth, bulk density and saturation deficit.

The spatial pattern of C sequestration potential density in topsoil was similar to Angers et al. (2011) who mapped C sequestration potential of French arable topsoil at a coarser resolution (broader administrative level). The regions with high sequestration potential (10-20 g kg^{-1}) in topsoil were also mainly located in intensively cultivated cropland, vineyard/orchard in the Mediterranean region and clay-rich soils in northeastern France. Although the data used for our study were entirely independent from those of Angers et al. (2011), the range of C sequestration potential density in agricultural soils is almost the same. As the method for estimating the C sequestration potential is the same, this consistency reflects consistency in the input data used for the estimation (Soil testing database for Angers et al., 2011 and RMQS carbon and clay measurements for this study). However, the modelling approaches are totally different, mean estimates at canton levels were used for mapping in the study of Angers et al. (2011) while regression Kriging on unbiased sampling was applied in this study.

6.4.3 Controlling factors of C sequestration potential vary with depth

As shown in Figure 6.5, different controlling factors were responsible for determining the C sequestration potential in topsoil and subsoil at the national scale. Despite the differences between topsoil and subsoil, land use, parent material, elevation, climate data and NPP were identified as the most important controlling factors, which is consistent with other similar studies at the regional or national scale (Martin et al., 2011; Meersmans et al., 2012; Schillaci et al., 2017).

In topsoil, land cover ranked as the most important factor with more than 35% increased MSE. It is reasonable that human activities such as cultivation and tillage, and direct C inputs by plants have a large effect on C dynamics and accumulation (Post and Kwon, 2000; Wiesmeier et al., 2014b). Parent material, NPP, elevation, MAT and MAP were also highly contributive factors for topsoil with increased MSE from 20% to 30%.

Under temperate climate, parent material usually determines soil texture and mineralogy. This result was expected, as soil texture and mineralogy strongly influence C dynamics (Balesdent et al., 2017; Batjes and Sombroek, 1997; Mathieu et al., 2015; Torn et al., 1997) and as texture is explicitly taken into account in Hassink's equation. Photosynthesis (NPP) is the main source of C inputs in soil thus directly controls C sequestration. Temperature and precipitation influence C mineralization by controlling the activity of soil microorganisms, and also directly influence the distribution of the land uses and NPP (Delgado-Baquerizo et al., 2018; Del Grosso et al., 2008). The importance of elevation in C sequestration potential may originate from its correlation with temperature, parent material and land cover. Its finer resolution may account for the spatial pattern that are not revealed by coarser temperature and parent material information. Curvature, slope and aspect had lowest contributions in the modelling with increased MSE < 10%, which is similar to the results from Wiesmeier et al. (2014a).

Interestingly, parent material was the top controlling factor in subsoil with increased MSE > 30%, while land cover showed less importance than it did in topsoil. In addition, soil type gained higher importance in subsoil than it did in topsoil. Our results are in line with previous studies showing that subsoil C is less controlled by human activities, land cover and climate than topsoil, but more related to soil inherent properties such as parent material, soil type and soil texture (Mathieu et al., 2015; Mulder et al., 2015; Wiesmeier et al., 2012). With similar increased MSE around 20%, elevation, MAP and NPP also revealed their importance in driving C sequestration potential in subsoil. In this study, we modelled SOC sequestration potential in topsoil and subsoil separately. Further work may statistically consider the relationships between topsoil and subsoil layers in the SOC accumulation, fluxes and processes (Heinze et al., 2018; Heitkötter et al., 2017).

The contribution of covariates (increased MSE) should be taken with caution. Indeed, the contribution of covariates also depends on the mutual relationship among covariates, so that high contributing covariates can inadvertently bear part of the contribution of the less contributing covariates. In our case, the high contribution of elevation may partly mask the effect of MAT and MAP, which may indeed be the real biophysical controlling factors.

6.4.4 Estimation of the total soil C sequestration potential stocks in mainland France

We found that model-based estimates of total C sequestration potential stocks were larger in subsoil (1360 Mt) than in topsoil (about 1000 Mt). For forest and grassland, the C sequestration potential stock was almost 1.5 times greater in subsoil than in topsoil. Take design-based estimates as a reference, the model-based estimates of C sequestration potential stock are underestimated in both topsoil and subsoil under almost all the land covers (Table 6.3), which indicates that our maps underestimate high SOC_{sp} values. Larger underestimation was observed in topsoil than in subsoil of predicted maps when compared with design-based estimates, especially under forest and grassland.

In total, considering the upper 50 cm from design-based estimates, soils of mainland France could therefore sequester an additional 2,638 Mt of C. This amount equals to 9,673 Mt CO_2 equivalent, that is about 28 times higher than the French mean annual CO_2 emission from 2005 to 2014 (350 Mt) (World Bank, 2018). Nearly 64% of these CO_2 equivalents could be sequestered in cultivated soils. In order to reach the 4 per 1000 objective, mainland France should achieve a mean sequestration rate of 14.4 Mt C y^{-1} for topsoil (0-30 cm) or 18.5 Mt C y^{-1} for 0-50 cm (Mulder et al., 2016a, 2016b; Minasny et al., 2017). Given that the mid-point of first RMQS and the Corine Land Cover map date around 2006, we assumed that our estimation of C sequestration potential in mainland France was a baseline map for 2006. Maintaining a 4 per mille yearly increase means that a total of 1,354 Mt and 1,739 Mt C should be sequestered for 0-30 cm and 0-50 cm by the end of this century. It therefore appears that SOC sequestration potential in 0-50 cm exceeds the demand of the 4 per 1000 aspirational target. As subsoil has much larger C sequestration potential, more attention should be given to management practices with potential to raise the C content of deeper layers. In addition, new C inputs in subsoil may become more stable due to the absence of tillage effect (Haddaway et al., 2016).

As mentioned by Barré et al. (2017), one has to bear in mind that C sequestration potential refers to the C stored in the soil fine fraction, which is assumed as stable with relatively long residence time in soils, while the concept of C storage potential is referred to as the maximum gain in soil C stock at a given time by implementing

changes in land management and it is more relevant to total C stock (including the coarse C fraction). The amount of C stored in the coarse fraction (>20 μm), which more likely represents labile or intermediate C, may also represent a large proportion of total C pool. As reported by Wiesmeier et al. (2014b), nearly 60% and 40% of C were stored in the coarse fraction for topsoil in Bavarian forests and grasslands, respectively. For cultivated soils, several studies have shown that the coarse fraction may represent more than 20% of the total C in temperate and tropical regions (Balesdent et al., 1998; Barthès et al., 2008; Gelaw et al., 2015). Therefore, the C storage potential may in fact be much larger than the C sequestration potential.

6.4.5 Can this additional C sequestration be reached in France?

Dignac et al. (2017) summarized implementable management practices that could be adopted in France (e.g., crop rotation, cover crops, no-tillage, agroforestry, management of crop residues, grassland management, irrigation, fertilization, exogeneous SOC inputs such as composts of various origins). Despite the large theoretical C sequestration potential in French soils, there are many limitations to achieving it, including biomass and N availability as well as climatic and hydrologic constraints on NPP and C mineralization (Chow et al., 2006; Nemani et al., 2003; van Groenigen et al., 2017). It should also be noticed that some solutions (e.g., converting all cultivated lands to grassland or forest) are not realistic in practice. In addition, there may be technical, socioeconomic, political or cultural constraints to the feasibility of reaching the theoretical C potential, such as localization of suggested land management, tradeoffs with agricultural production and food security (Elbehri, 2015; Paustian, et al., 2016). Half of the forest topsoils in France presented SOC stocks larger than their C sequestration potential, which indicates that it is necessary to enhance our understanding of C sequestration potential or to refine the concept of the theoretical C potential (e.g., estimate a realistic potential of C storage rather than C sequestration, Barré et al., 2017).

Moreover, as raised by many authors (Don et al., 2011; Powlson et al., 2012; Smith, 2005), the permanence of the C storage is questionable as increases in C stocks are highly reversible. This permanence may be endangered if practices storing C are interrupted, or even for other less manageable issues such as climate change. From this point of view, under- and over-saturated soils are two sides of the same coin. Our

results show that overall, 176 Mt C exceed the C saturation in French topsoil and might thus be very sensitive to land use change. Therefore, it might be as important to preserve these sensitive stocks than to try to create new ones.

6.5 Conclusions

We estimated C sequestration potential for top- and subsoil and provided fine resolution maps at a national scale. Regression Kriging approach performed successfully in mapping C sequestration potential using environmental covariates. The controlling factors of SOC sequestration potential differed from topsoil and subsoil. The main controlling factors of SOC sequestration potential in topsoil and subsoil were land use and parent material, respectively. The regions with high sequestration potential in topsoil were mainly located in intensively cultivated cropland, vineyard/orchard in the Mediterranean region and clay-rich soils in northeastern France. In subsoil, a higher C sequestration potential than in topsoil was found under all land covers. Therefore, we should pay more attention to management practices with potential to raise the SOC in deeper layers, such as plant species or cultivars with deeper and thicker root systems, promoting soil faunal activities and manage subsoil microorganisms. Nearly half of forest and one third of grassland soils were over-saturated in topsoil. Although the overall C sequestration potential for French soils is very large, it might be as important to preserve the sensitive stocks in over-saturated topsoils than to try to create new ones.

6.S Supplementary materials

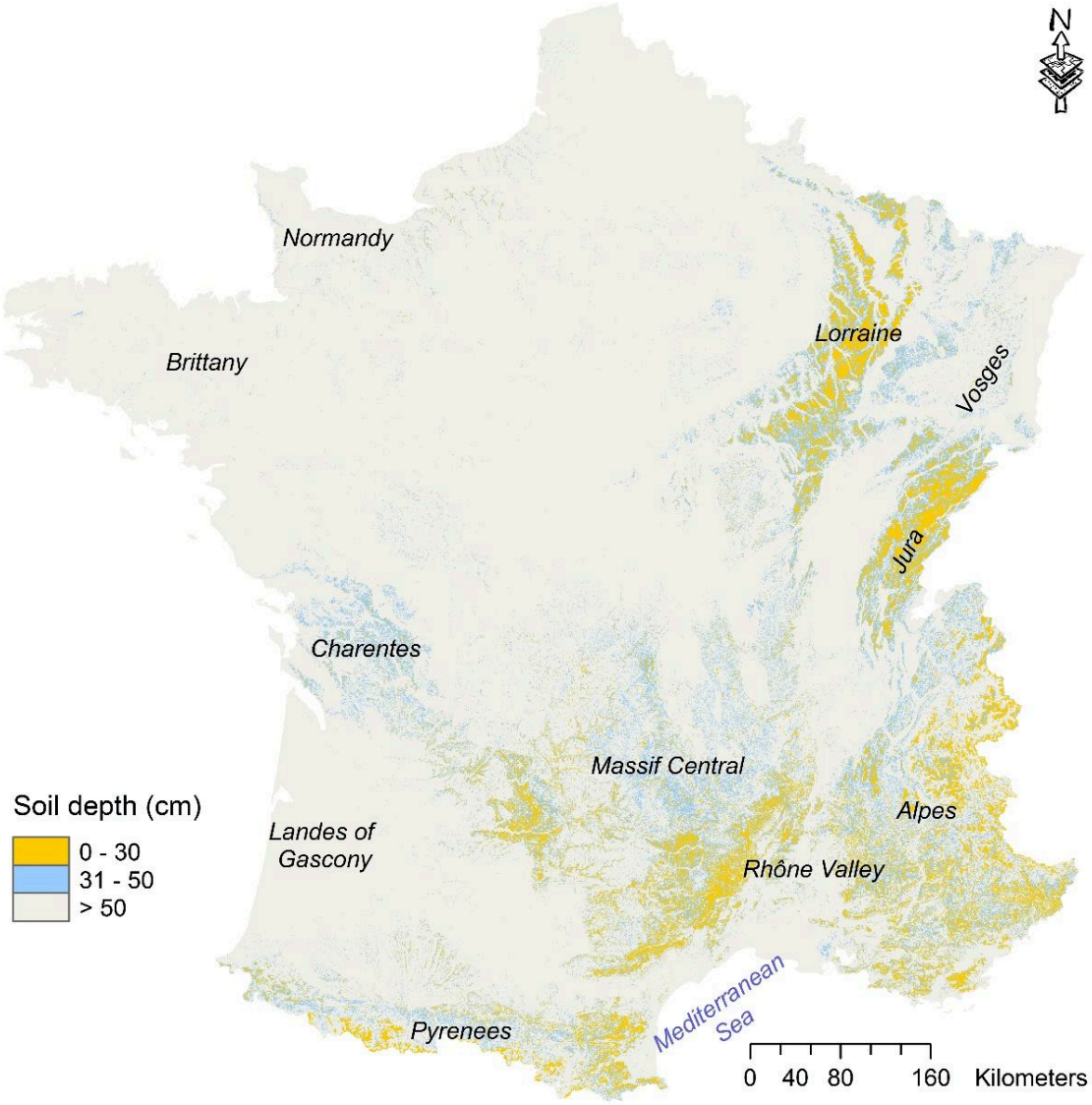


Figure S6.1 Soil depth map of mainland France in three depth intervals (0-30 cm, 31-50 cm and > 50 cm). The original soil depth map is from Marine et al. (2016), which is derived from gradient boosting modeling with quantile transformation.

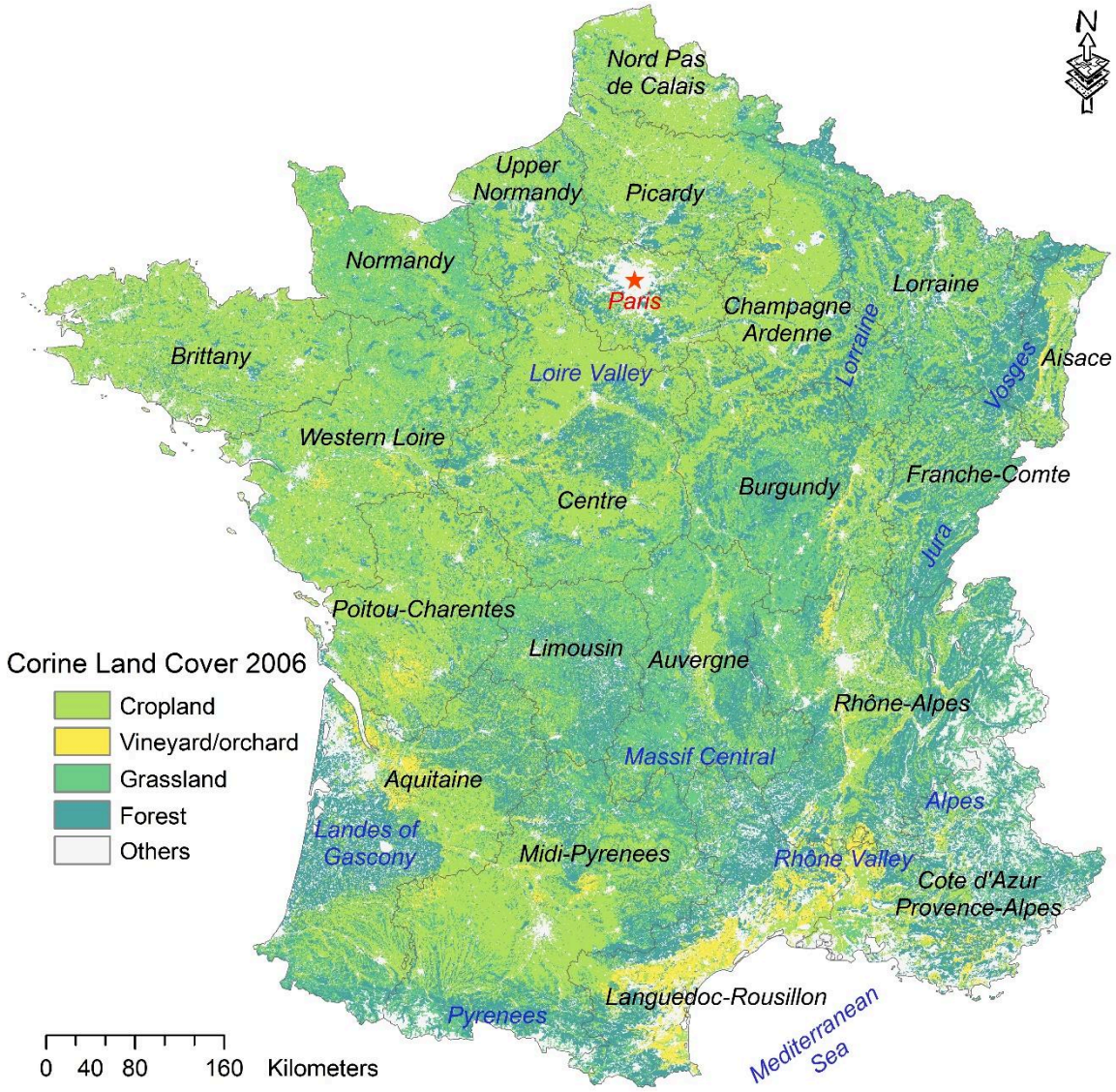
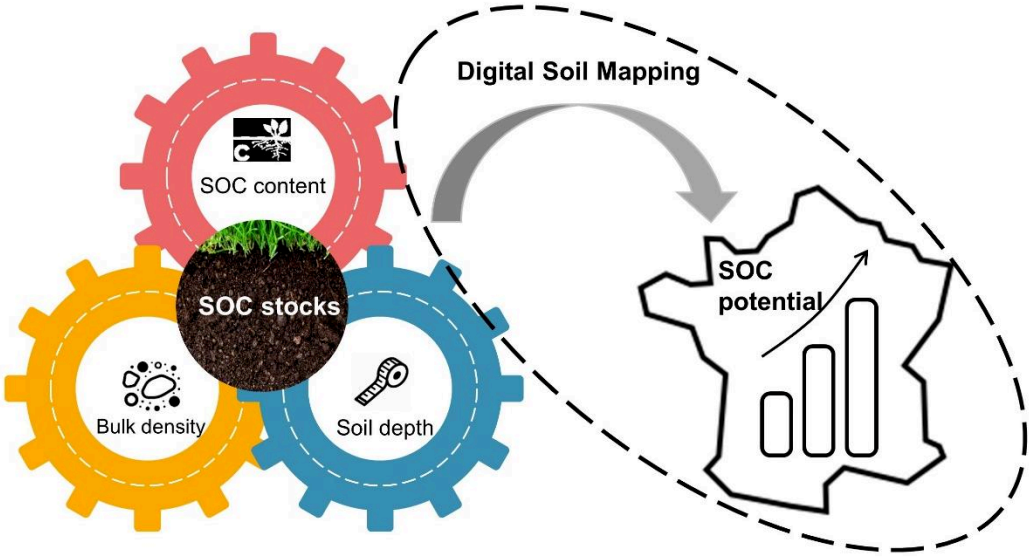


Figure S6.2. Corine Land Cover map of 2006 in mainland France with main regions and geography units marked

Chapter 7

National estimation of soil organic carbon storage potential for arable soils: a data-driven approach coupled with carbon-landscape zones



Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Saby, N.P.A., Walter, C., 2019. National estimation of soil organic carbon storage potential for arable soils: a data-driven approach coupled with carbon-landscape zones. Science of The Total Environment, 666, 355-367.

7.1 Introduction

Globally, the soil C pool (2500 Gt) is 3.3 times the size of the atmospheric pool (760 Gt) and 4.5 times the size of the above-ground vegetation pool (560 Gt). Variations in the SOC pool depend on the balance between C input and C output and on the soil intrinsic capacity to store or sequester SOC. Therefore, soils have the potential to partly offset anthropogenic greenhouse gas emissions by sequestering SOC (Lal, 2004; Paustian et al., 2016). Moreover, increasing soil organic carbon (SOC) generally improves soil quality and functioning, and thus can potentially contribute to enhance agricultural production and food security, restore degraded land, and promote ecosystem services such as erosion mitigation, soil water provision, nutrient availability for plants, and soil biodiversity (Lal, 2004; Stockmann et al., 2013). Recognizing the importance of increasing SOC at the global scale, a voluntary action initiative "4 per 1000 carbon sequestration in soils for food security and the climate" (http://4p1000.org/) was launched at the COP21. The 4 per 1000 initiative aims at promoting land management practices (e.g., conservation agriculture, cover cropping, agroforestry) leading to the protection of SOC stocks and to their increase, with a proposed aspirational annual growth rate of 0.4% of current SOC stocks in the 0 to 0.4m layer. This aspirational target of a 4 per 1000 rate of annual increase in global SOC stocks is still a matter of intense debate in the scientific community (e.g., Paustian et al., 2016; Chabbi et al., 2017; Minasny et al., 2017; Sanderman and Berhe, 2017; van Groenigen et al., 2017; Baveye et al., 2018; Minasny et al., 2018; Soussana et al., 2019). Most of the discussion evolves around the actual feasibility of reaching this goal due to: i) limitations linked to the availability of C inputs to soil and of other major elements (e.g. N, P), ii) the non-permanence of SOC stocks, which may be accentuated by climate change, iii) the limited capacity of SOC storage, both in terms of area and duration, iv) the difficulty to assess and verify changes in SOC which are both highly variable in space and time, and v) the

feasibility to implement massive changes in management practices and land use and the required commitments for a very long period of time. Nonetheless, the scientific community agrees on the urgent need to protect existing SOC stocks and to increase them where ever possible, acknowledging that beyond the biophysical limits and barriers for storing additional carbon in soils, socioeconomic limits may be even more constraining (Minasny et al., 2018; Soussana et al., 2019). However, this initiative urges the scientific community to provide biophysical estimates of the potential of soils to store additional carbon.

The SOC storage potential generally refers to the maximum gain in SOC stock attainable at a given timeline by implementing changes in land use or management, and will vary under different pedoclimatic conditions (Post and Kwon, 2000; Stockmann et al., 2013; Barré et al., 2017; Chenu et al., 2019). The concept of SOC saturation has been used to estimate the maximum amount of SOC that can be associated with the fine fraction (Hassink, 1997) and therefore considered as relatively stable. In the context of the 4 per 1000 initiative, the aspirational target of increasing SOC stocks at an annual growth rate of 0.4% relates to the total (whole-soil) SOC stocks in the 0-0.4 m layer (whole-soil, including the coarse fraction). Therefore, determining whole-soil SOC storage potential using the maximum SOC associated with the fine fraction is not appropriate because the SOC stored in the coarse fraction can represent a significant percentage of the total SOC stocks. As summarized by Chen et al. (2018, 2019b), under temperate climate, SOC in the coarse fraction could account, on average, for 15%, 34% and 31% of total SOC stocks under cropland, forest and grassland, respectively, in topsoil, and account for nearly 25%, 14% and 7% of SOC stocks for cropland, forest and grassland in subsoil.

For an improved quantification of SOC storage potential, Barré et al. (2017) proposed one avenue: i) First, establish the reference stocks with an estimate of the highest reachable SOC stock for a given soil; ii) second estimate possible SOC storage between the current SOC stock of a given soil and this reachable highest

SOC stock under a given land-use for different land management practices. Furthermore, Barré et al. (2017) suggested that this avenue can be achieved using either a data-driven approach (empirical observation of SOC stocks and storage) or mechanistic simulation models. The data-driven approach assumes that the highest reachable SOC stocks under a specific land use/cover or land management practices for each different pedoclimatic conditions could be empirically determined by the highest values (e.g., by the mean of using top quantiles) among the observed SOC stocks for these conditions. This hypothesis implicitly assumes that the values of the top quantiles reflect the optimal management practices for SOC storage and they are thus considered as 'proxies' of the maximum reachable SOC stocks under these different pedoclimatic conditions.

Based on the detailed and extensive French Soil Monitoring Soil Network data base, our objective was to test a data-driven approach for estimating SOC storage potential of arable soils in mainland France. We developed a procedure which consisted of: i) determining carbon-landscape zones by clustering the data from a combination of net primary production (C input), climatic decomposition index (C decomposition) and soil clay content (C protection from decomposition); ii) estimating the maximum SOC stocks of arable soils (topsoil and subsoil) for each carbon-landscape zone using four percentiles (0.80, 0.85, 0.90 and 0.95); iii) calculating by difference with the current SOC stocks, the SOC storage potential of arable topsoil and subsoil under these four percentiles.

7.2 Materials and methods

7.2.1 Soil data

Covering the entire mainland France under different soil, climate, relief and land cover conditions, 2,092 sites from the first campaign of the French Soil Monitoring Network (RMQS) were sampled from 2001 to 2009. The RMQS is based on a 16 km × 16 km square grid and all sites were selected at the centre of

each grid cell. Topsoil (0-30 cm) and subsoil (30-50 cm) were collected using a hand auger. For each site, on the basis of an unaligned sampling design with a 20 m × 20 m square, 25 samples were merged into a composite sample and then were air-dried (controlled at a temperature of 30 ºC and an air-moisture of 30%) and sieved to 2 mm before laboratory analysis. A soil pit was dug at 5 m from the south border of sampling sites, and the main soil characteristics were recorded and bulk density and percentage of coarse elements were measured (Martin et al., 2009). For some RMQS sites, subsoil did not exist as soils were thin at these locations. SOC was determined by dry combustion. Only these RMQS sites (n=1089) located on arable soils were used in this study (Figure 7.1).

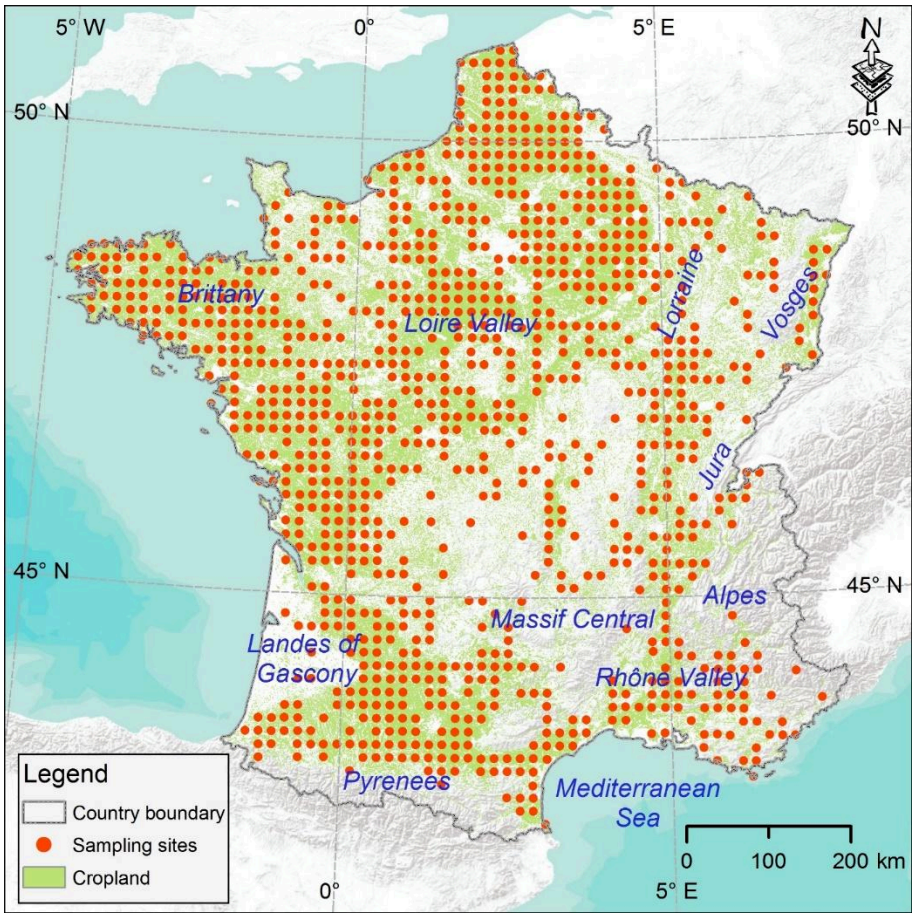


Figure 7.1 RMQS sites located in arable soils

The SOC stock was calculated as below:

$$SOC_{stock}=p\times SOC\times BD\times(100-ce)\times 10^{-2} \quad (7.1)$$

where SOC_{stock} is the SOC stock (kg m^{-2}), p is the actual thickness (cm) of topsoil or subsoil, *SOC*, *BD* and *ce* are the content of SOC (g kg^{-1} or ‰), bulk density (kg m^{-3}), and percentage of coarse elements (%).

7.2.2 Net primary production, climatic data, soil clay content and SOC stocks maps

Net primary production (NPP) was extracted from the MOD17A2H version 6 Gross Primary Production product (NASA LP DAAC, 2017) from 2000 to 2010. It is a cumulative 8-day composite of values with 500-meter original resolution. The 8-day NPP data is averaged into monthly data and resampled to 1 km resolution. Cities and water-covered regions have been masked in this product.

WorldClim Version 2 (Fick and Hijmans, 2017), which is spatially interpolated using between 9000 and 60000 weather stations globally, was used for climatic data: It has average monthly climate data for minimum, mean, and maximum temperature and for precipitation, solar radiation, wind speed and water vapour pressure for 1970-2000 at 1 km resolution.

Maps of soil clay content for topsoil (0-30 cm) and subsoil (30-50 cm) were derived from *GlobalSoilMap* France products (Mulder et al., 2016a). As these were produced at six standard depth intervals (e.g., 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm and 100-200 cm), soil clay content maps were harmonized using a depth-weighted method (Figure S7.1).

The Corine Land Cover 2006 (UE-SOeS, 2006) was used as the land cover/use classification map. It has an original resolution at 100 m and was resampled to 90 m in order to meet the requirement of the *GlobalSoilMap* project (Sanchez et al., 2009; Arrouays et al., 2014a). The Corine Land Cover map was reclassified as cropland, forest, grassland and others, and only cropland was presented in this study (Figure 7.1).

The current SOC stocks map for topsoil (0-30 cm) was produced using RMQS dataset by a hybrid model coupling the boosted regression trees (BRT) and robust geostatistical approaches described in Martin et al. (2014). The covariates used in

modelling were explicitly documented in Chen et al. (2018). To remove the interference of the positions without SOC stocks in subsoils (where subsoil does not exist), a three-stage approach was applied for SOC stocks modelling in the subsoil (30-50 cm): 1) produce a map to identify whether subsoils exist using BRT model; 2) produce a SOC stocks map by the hybrid model, where the RMQS sites without SOC stocks are excluded; 3) merge the two maps by keeping the SOC stock values where subsoils exist and setting the locations where subsoil do not exist as NA (not available). The SOC stocks maps for topsoil and subsoil have a spatial resolution of 90 m and they can be found in the Figure S7.2. The national SOC stocks were 3.65 Gt and 1.04 Gt for topsoil and subsoil, respectively. Cropland contained 1.37 Gt and 0.44 Gt SOC in the topsoil and subsoil.

All the datasets were reprojected to Lambert 93, which is an official projection for mainland France.

7.2.3 Calculation of climatic decomposition index

As carbon decomposition generally increases with temperature and moisture, a climatic decomposition index (CDI) was used to characterise the interaction between temperature and water stress as suggested by Carol Adair et al. (2008).

Before determining the CDI, potential evapotranspiration (PET) was calculated using Hargreaves model (Hargreaves et al., 1985), which performs well and requires less parameterization than the Penman-Monteith method (Hargreaves and Allen, 2003). Monthly PET (mm month^{-1}) is defined below:

$$PET=0.0023\times SR\times(T_{mean}+17.8)\times\sqrt{T_{range}} \quad (7.2)$$

where SR is monthly solar radiation (mm month^{-1}, transformed from KJ m^{-2} day^{-1}), T_{mean} is monthly mean temperature (ºC) and T_{range} is the difference between the monthly maximum and minimum temperature (ºC).

The CDI is calculated as a function of the mean monthly mean temperature (T), monthly precipitation (PPT) and monthly PET (Carol Adair et al., 2008):

$$CDI=F_T(T)\times F_W(PPT,PET) \quad (7.3)$$

$$F_T(T)=0.5766\times e^{308.56\times\left(\frac{1}{56.02}-\frac{1}{(273+T)-227.13}\right)} \quad (7.4)$$

$$F_W(PPT,PET)=\frac{1}{1+30\times e^{-8.5\times\frac{PPT}{PET}}} \quad (7.5)$$

where $F_T(T)$ and $F_W(PPT,PET)$ are the monthly effects of temperature and water stress on decomposition.

7.2.4 Delineation of carbon-landscape zones using Gaussian mixture models

Generally, SOC dynamics depend on the trade-off between the SOC input and SOC loss processes. When SOC input is greater than OC loss, the soil will accumulate C, and otherwise, soil C will decrease. Climatic decomposition index and NPP are here considered as proxies of C loss and input that control the SOC balance, and clay content considered as a controlling factor of SOC persistence. The underlying simplifying assumption is that decomposition mainly depends on both climate and soil characteristics. Therefore monthly CDI and NPP, and soil clay content were used to compute the carbon-landscape zones (CLZs) using Gaussian mixture model (GMM) which is a similar approach to that used by Mulder et al. (2015). To reduce multicollinearity and computing time, principal component analysis (PCA) was performed before the clustering step on monthly CDI and NPP data separately. We retained only the first three and four principal components that explained more than 95% of the variance for CDI and NPP, respectively. Therefore, after adding soil clay content for topsoil and subsoil, a total of nine variables were used for GMM clustering. Moreover, to reduce computing complexity, we also selected 20,000 pixels in France as calibration data set of the GMM clustering. The resulting clustering model was then used to predict to which CLZ each pixel of the entire territory belongs.

Gaussian mixture model was conducted to compute clusters that were considered as CLZs in this study. GMM is one of the model-based clustering techniques, which optimizes the fit between the measured data and

mathematical models using a probabilistic approach. GMM is based on the assumption that the data are generated by a mixture of Gaussian distributions. Then, the parameters of GMMs are estimated by maximisation of the likelihood using the Expectation Maximization (EM) algorithm. EM algorithm starts with a random initialization and then iteratively optimizes the clustering using two steps: (i) Expectation step determines the expected probability of assignment of data to clusters using current model parameters; (ii) Maximisation step updates the optimal model parameters of each mixture based on the new data assignment.

The number of clusters was tuned from 1 to 30 and their associated Bayesian information criterion (BIC) was calculated for the evaluation of clustering performance. The number of clusters was selected considering a trade-off between the BIC values and the available number of RMQS sites within each land use for each cluster. GMMs were performed using ClusterR package in R 3.3.2 (Mouselimis, 2016; R Core Team, 2016). The optimized CLZs map was resampled to 90 m resolution.

7.2.5 SOC storage potential and analysis of the sensitivity to the percentile setting

Empirical maximum SOC stock values were estimated for arable topsoil and subsoil under given CLZs using RMQS dataset (point observations). The underlying hypothesis is that the highest values correspond to a maximum SOC that is reachable under current management practices. We had to fix a given "percentile" from these highest values, because taking only the maximum value would have resulted in selecting only one extreme case (e.g., recently cleared forest, site with large external C input) that could lead to significant over-estimation. Four percentiles at 80%, 85%, 90% and 95% were tested to estimate the empirical maximum SOC stock values that could be reached under a given CLZ. A bootstrapping approach was applied to assess the uncertainty from data source both for each CLZ and tested percentiles. We repeated the bootstrapping procedure 100 times and thus obtained 100 estimates of the maximum SOC stock

values for each CLZ and percentile. The mean value obtained from these one hundred estimates was used as an estimate of the maximum SOC stock value for each CLZ and percentile. We then estimated the uncertainty (90% confidence intervals, 90% CIs) of these maximum SOC stock values by using the 5 and 95 percentile of the bootstrapping results.

The SOC storage potential was calculated as the difference between the empirically-determined maximum SOC stocks and current SOC stocks (Figure S7.2) under arable land use. Four SOC storage potential maps were produced using the four tested percentiles for both topsoil and subsoil, and their associated 90% CIs.

We evaluated the effect of percentile setting on the estimation of SOC storage potential by both comparing the differences in the SOC storage potential spatial distribution and national SOC storage potential estimates.

7.3 Results

7.3.1 Spatial distribution of CDI, NPP and their principal components

Figure 7.2 shows the spatial distribution of CDI and NPP in mainland France. CDI increased gradually from January to August and then decreased gradually to December. Different from CDI, NPP started to increase from January and reached the peak in June, and then decreased gradually to December.

Accounting for 98.3% and 97.0% of the total variances (95% was set as a threshold), the first three and four principal components (PCs) were kept for CDI and NPP, respectively. Figure 7.3 presents the final seven PCs used in clustering. The 3 PCs of CDI showed long range spatial patterns in mainland France while the spatial patterns for 4 PCs of NPP were mainly characterized by median and short ranges.

7.3.2 Carbon-landscape zones

The BIC value decreased quickly when the number of clusters was less than 10, and then it decreased slowly after 10 clusters (Figure 7.4). The result indicated

that more clusters were helpful for separating the differences within clusters. However,

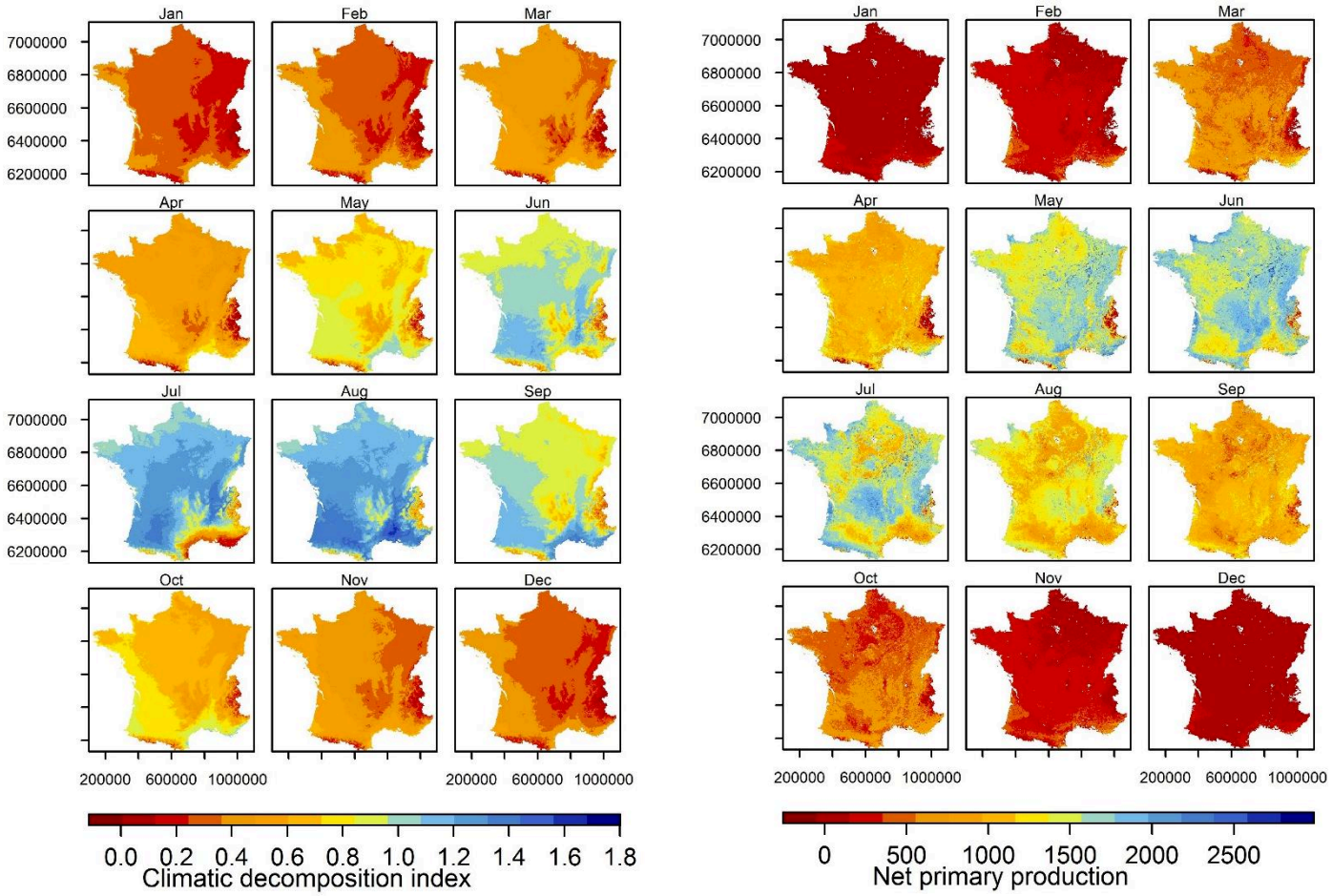


Figure 7.2 Spatial distribution of monthly climatic decomposition index and net primary production.

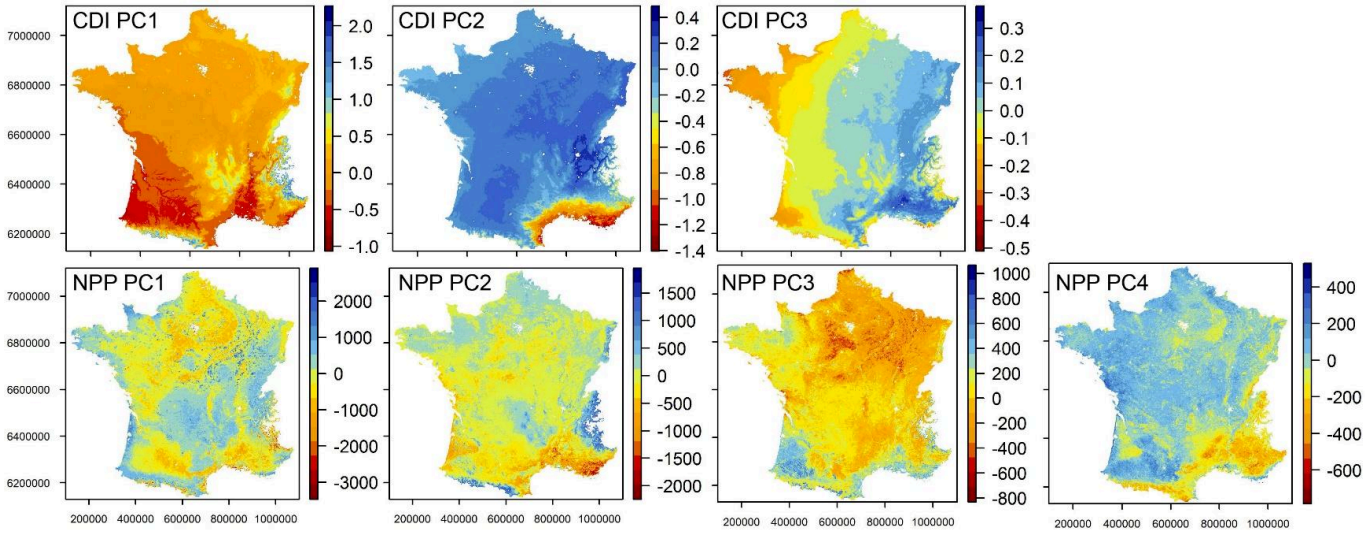


Figure 7.3 Spatial distribution of principal components for climatic

decomposition index and net primary production.

more clusters meant less available RMQS sites falling into each cluster. Figure 7.5 shows the number of RMQS sites located in each cluster. Our aim was to avoid clusters having a number of RMQS sites less than ten, which may not be enough to derive a robust estimate of the quantiles. Two clusters had less than ten RMQS sites when the number of clusters varied from 8 to 10. When the number of clusters increased from 11 to 13, three clusters were found with less than ten RMQS sites. We optimized the number of clusters at ten as it appeared to be the best compromise between separating the differences between clusters and keeping an acceptable number of clusters having less than ten RMQS sites.

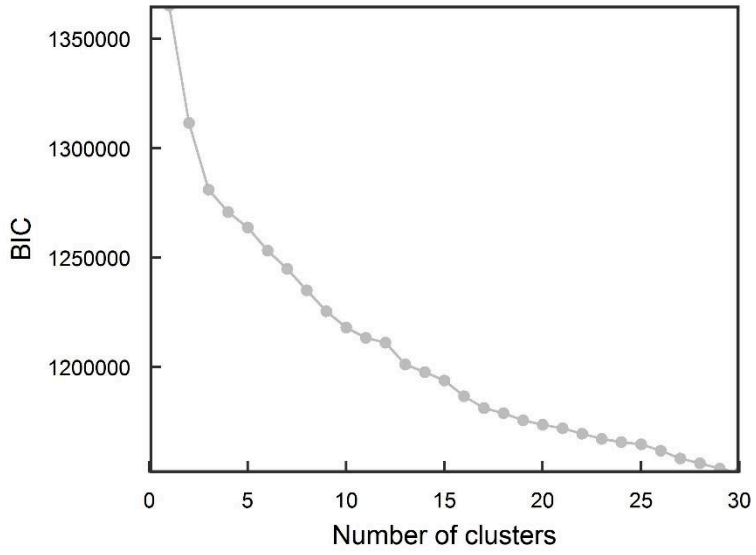


Figure 7.4 Relationship between the number of clusters and BIC.

Figure 7.6 illustrates the spatial distribution of CLZs in mainland France. CLZ 1 is mainly distributed in north-eastern France which is characterized by a rather continental climate and relatively high soil clay contents, mostly ranging from 22% to 35% in topsoil, and being even higher in subsoil (Figure S7.3). CLZ 2 represents most of western France characterized by a mild and wet oceanic climate and relatively homogeneous soil clay contents (mostly ranging from 15 to 20% both in top- and subsoil, Figure S7.3). CLZ 3 is located in northern France and mainly

corresponds to the maximal extension of deep loess deposits.

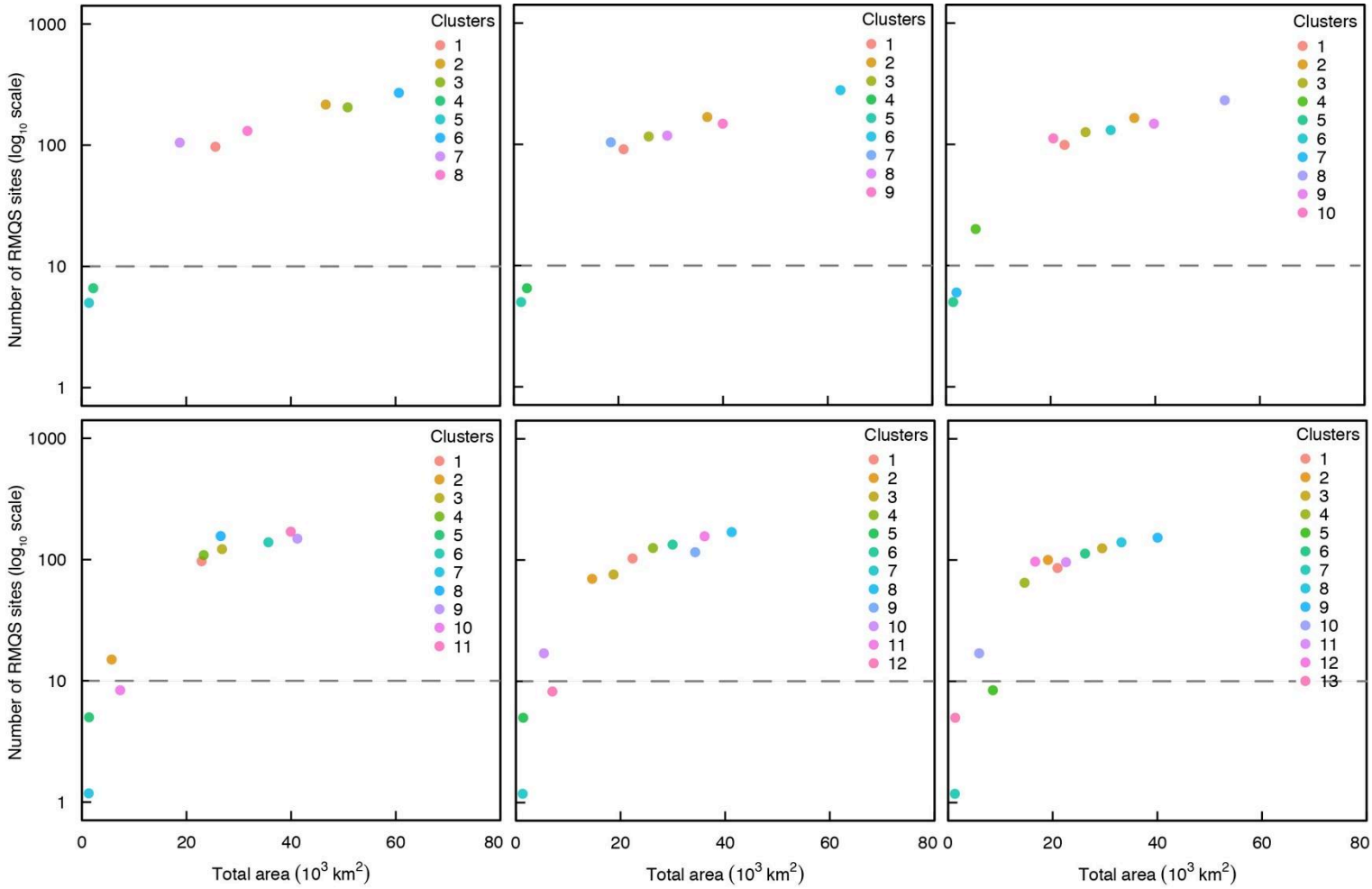


Figure 7.5 Number of RMQS sites located in each carbon-landscape zone.

It exhibits clay contents centred around 20% for topsoil and a bit higher for subsoil, both with a rather low statistical dispersion. CLZ 4 is located in the Massif Central and the Vosges mountains, and is characterized by a rather cold climate due to elevation and rather homogeneous clay contents, mostly ranging from 15% to 20% for both layers (Figure S7.3). CLZ 5 is located in southern France and strictly corresponds to the area of the 'Landes of Gascony' which is characterized by a mild climate and nearly pure sandy aeolian deposits having clay content nearly always less than 5% (Augusto et al., 2010). CLZ 6 is located in central France and corresponds to the foothills of the Massif Central, with a lower elevation than

its central part. Part of the CLZ 6 is also spread in various other locations, all of which corresponding to ancient alluvial deposits coming from these foothills. In

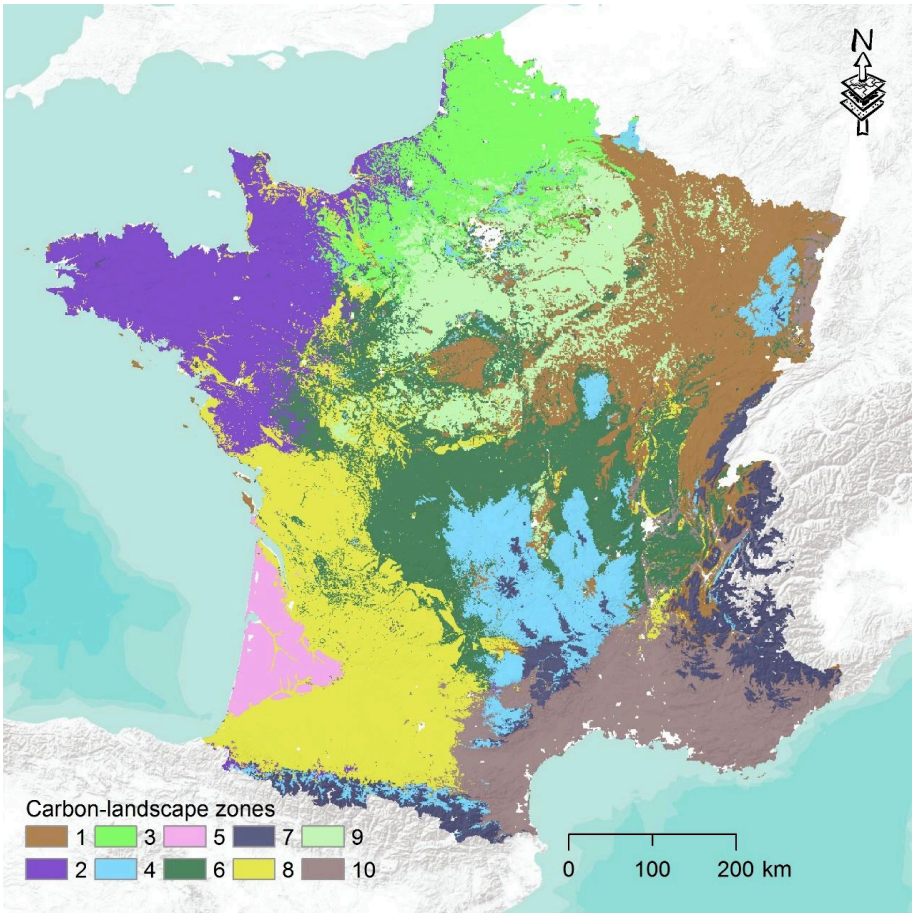


Figure 7.6 Optimal 10 carbon-landscape zones in France.

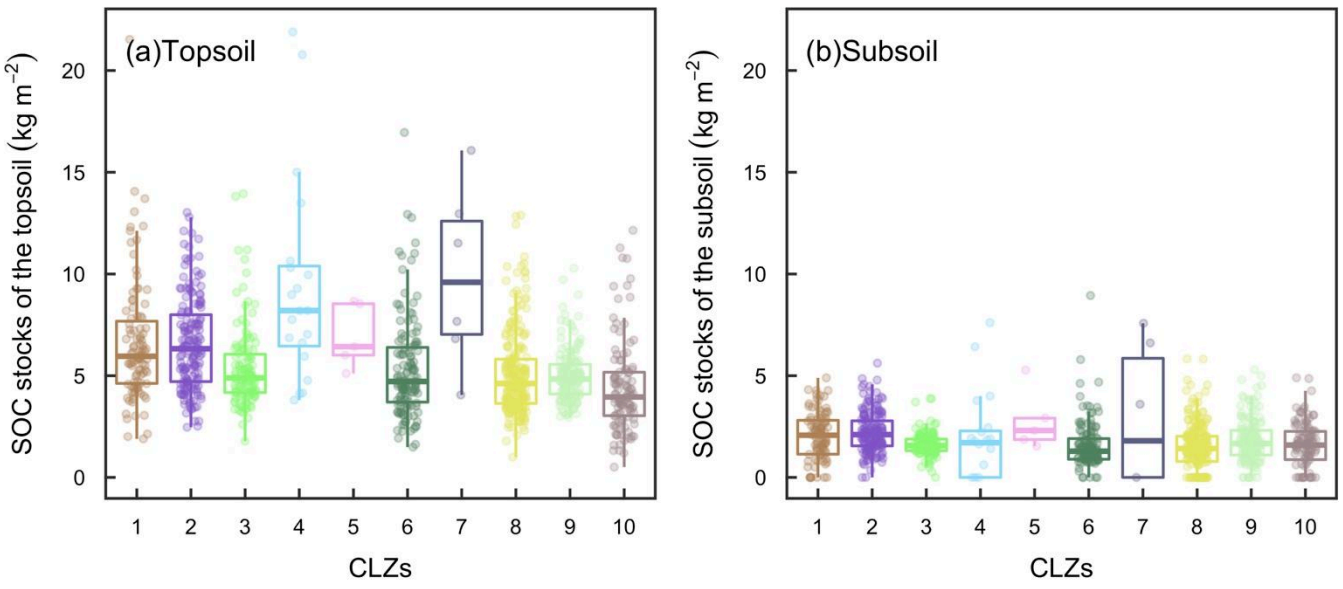


Figure 7.7 Boxplots of SOC content in topsoil and subsoil under 10 carbon-landscape zones.

topsoil, most clay contents range from 15% to 20% and slightly higher in subsoil. CLZ 7 is exclusively located in the highest elevations located at the top of the main mountain ranges (Pyrenees, Alps, Jura and Massif Central), with soil texture being rather clayey (around 30%). This CLZ also includes many shallow soils (Lacoste et al., 2016), and thus the information on clay content of the 0.3 to 0.5 m layer is often missing. CLZ 8 occupies most of south-western France characterized by mild winters and hot summers. It is characterized by a very large range and dispersion of clay content in both layers, although a large part (interquartile range) ranges from 20% to 30%. CLZ 9 is mainly distributed in central and northern France. Its clay content in topsoil and subsoil is centred around 25% (Figure S7.3) and showing a small increase in subsoil. Lastly, CLZ 10 shows low NPP values in autumn, because of land use consisting mainly of vineyards and wheat crops. It is clearly located in the Mediterranean region with very hot temperatures and very low NPP in summer. The clay content is centred around 20%, with a statistical dispersion similar to the other CLZs.

Figure 7.7 presents the design-based estimates of SOC stocks for arable soils for the ten CLZs in topsoil and subsoil. In order to get unbiased estimates, these estimates were computed using the values obtained from the RMQS grid values within each CLZ. The median SOC stocks of topsoil ranged from 4.89 to 9.67 kg m^{-2} under the 10 CLZs. Fewer differences of SOC stocks were found in subsoil, and subsoil had much lower SOC stocks than topsoil with a range of median SOC stocks from 1.31 to 2.08 kg m^{-2}.

7.3.3 Empirical maximum SOC stocks under four percentile settings

As expected, there was a clear trend that the maximum SOC stocks for topsoil and subsoil increased when percentile became higher, however, the magnitude of these increases varied among different CLZs (Figure 7.8). In topsoil, large differences (>4 kg m^{-2}) in maximum SOC stocks between percentile of 0.95 and percentile of 0.8 were observed in CLZ 1 and CLZ 4, and the differences ranged from 0.28 to 3.65 kg m^{-2} for other CLZs. In subsoil, differences in maximum SOC

stocks between percentile of 0.95 and percentile of 0.8 were below 1.5 kg m^{-2} for almost all the CLZs, except for CLZ 4 with a value of 2.71 kg m^{-2}.

The 90% CIs also differed between CLZs as well as between percentiles. A large percentage of high 90% CIs (upper limit minus lower limit > 10 kg m^{-2} for topsoil or > 5 kg m^{-2} for subsoil) of maximum SOC stocks were found in CLZ 4 and CLZ 7, which indicated large variability for these two mountainous CLZs having a rather low number of sites. Besides, subsoil in arable soils had lower 90% CIs than topsoil.

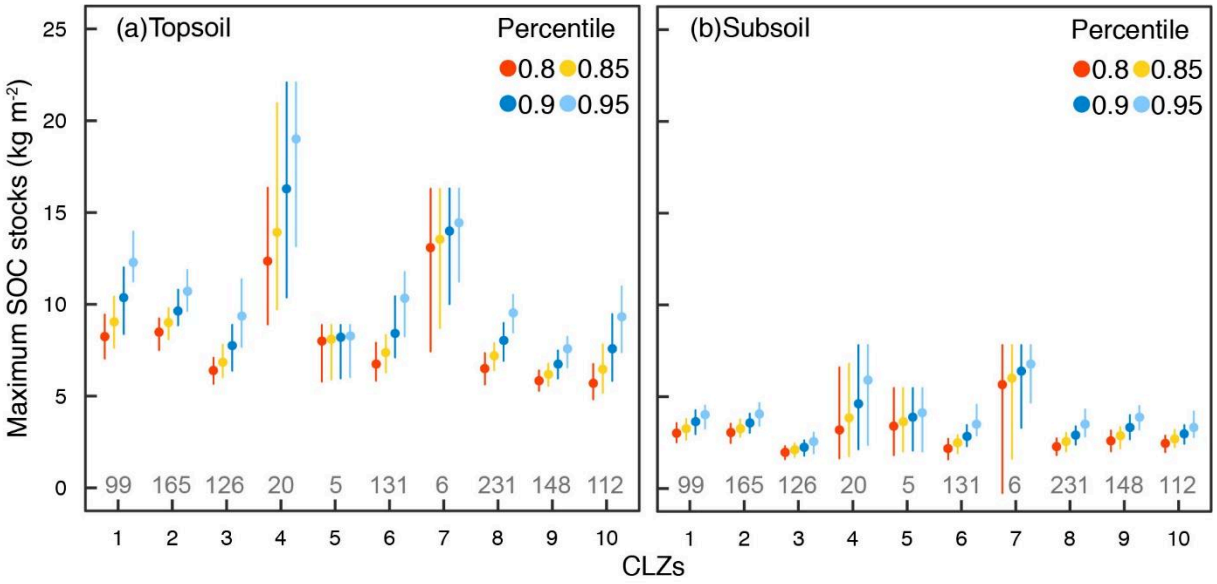


Figure 7.8 Empirically maximum SOC stocks in topsoil and subsoil under four percentile settings. The four colours are related to four percentiles. For each percentile, bar shows the interval between upper limit and lower limit of 90% CIs. Number of samples is shown in grey.

7.3.4 Spatial distributions of SOC storage potential

Figure 7.9 and Figure 7.10 show the spatial distributions of SOC storage potential and 90% CIs under four percentile settings for topsoil and subsoil, respectively. When percentile was set at 0.8, French arable topsoil had a SOC storage potential less than 2 kg m^{-2} except for a part of Brittany and south-eastern

France near the Mediterranean Sea. With the increasing percentile, intensively cultivated plains of the central, the northern half and the southwestern part of France showed a large potential to store more SOC. Cropland located around mountainous regions including the Pyrenees, the Alps, the Jura and the Vosges generally had a relatively low SOC storage potential across all percentiles. Large differences were observed for total SOC storage potential under different percentile settings (Table 7.1). The French national SOC storage potential and 90% CIs for arable topsoil were 336 (203, 501) Mt when percentile was 0.8. Larger increases were observed for total SOC storage potential and 90% CIs with the increasing percentiles, which reached at 470 (308, 662) Mt, 674 (434, 950) Mt and 1020 (740, 1,283) Mt for a percentile of 0.85, 0.9 and 0.95, respectively.

The subsoil showed much lower SOC storage potential than topsoil. Most regions of mainland France had low SOC storage potential (< 1 kg m^{-2}) at percentiles of 0.8 and 0.85, and relative high SOC storage potential (1-3 kg m^{-2}) were observed in central France. Similar with topsoil, increasing percentiles resulted in higher SOC storage potential across mainland France, and fewer differences of SOC storage potential were found between cropland located around mountainous regions and other regions under four percentile settings. At percentile of 0.8, subsoil had the potential to sequester 165 Mt additional SOC with a 90% CI between 91 Mt and 250 Mt. Total SOC storage potential and their 90% CIs were 228 (150, 306) Mt, 309 (226, 404) Mt and 433 (331, 560) Mt for percentiles of 0.85, 0.9 and 0.95, respectively.

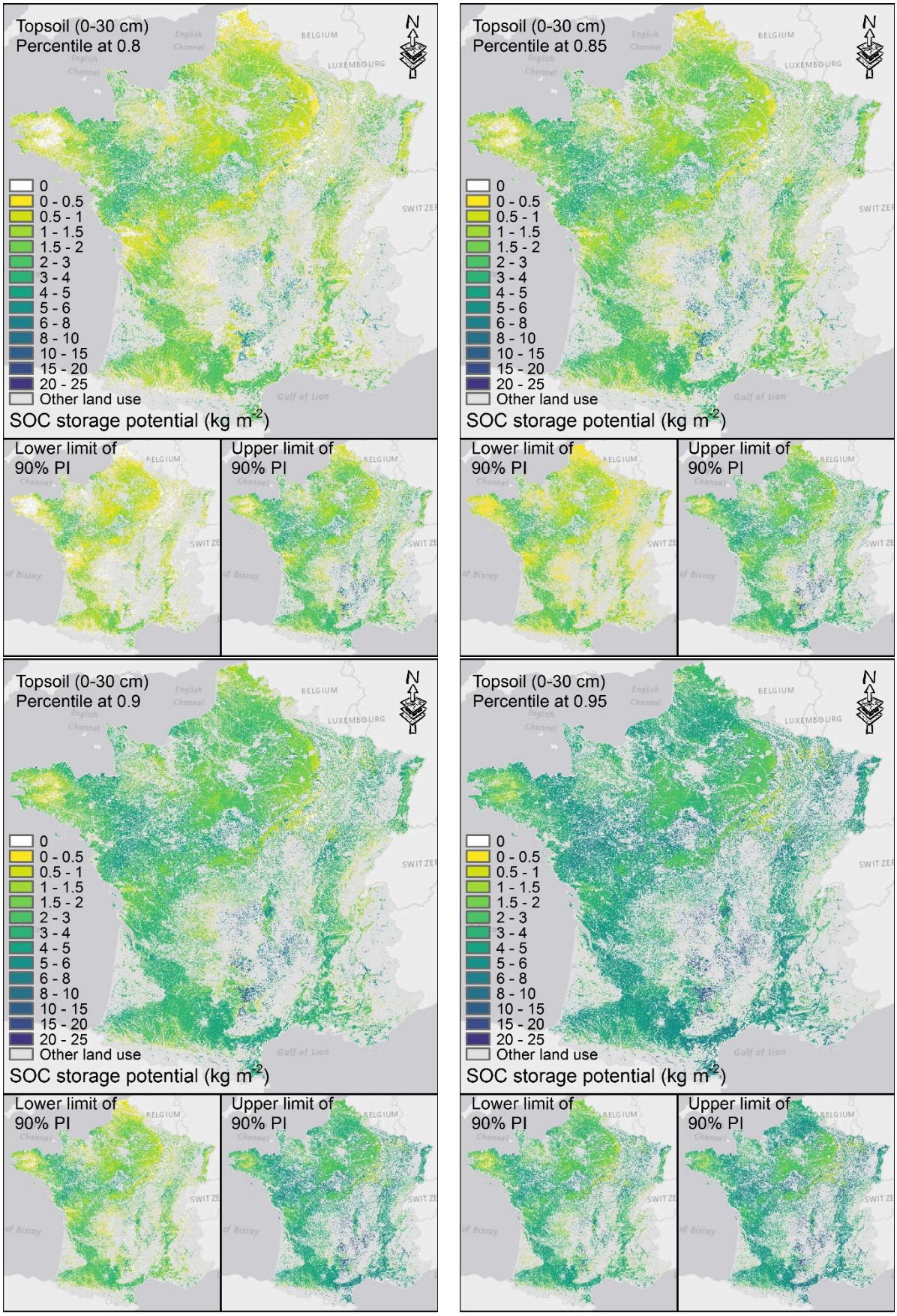


Figure 7.9 SOC storage potential for arable topsoil under four percentile settings.

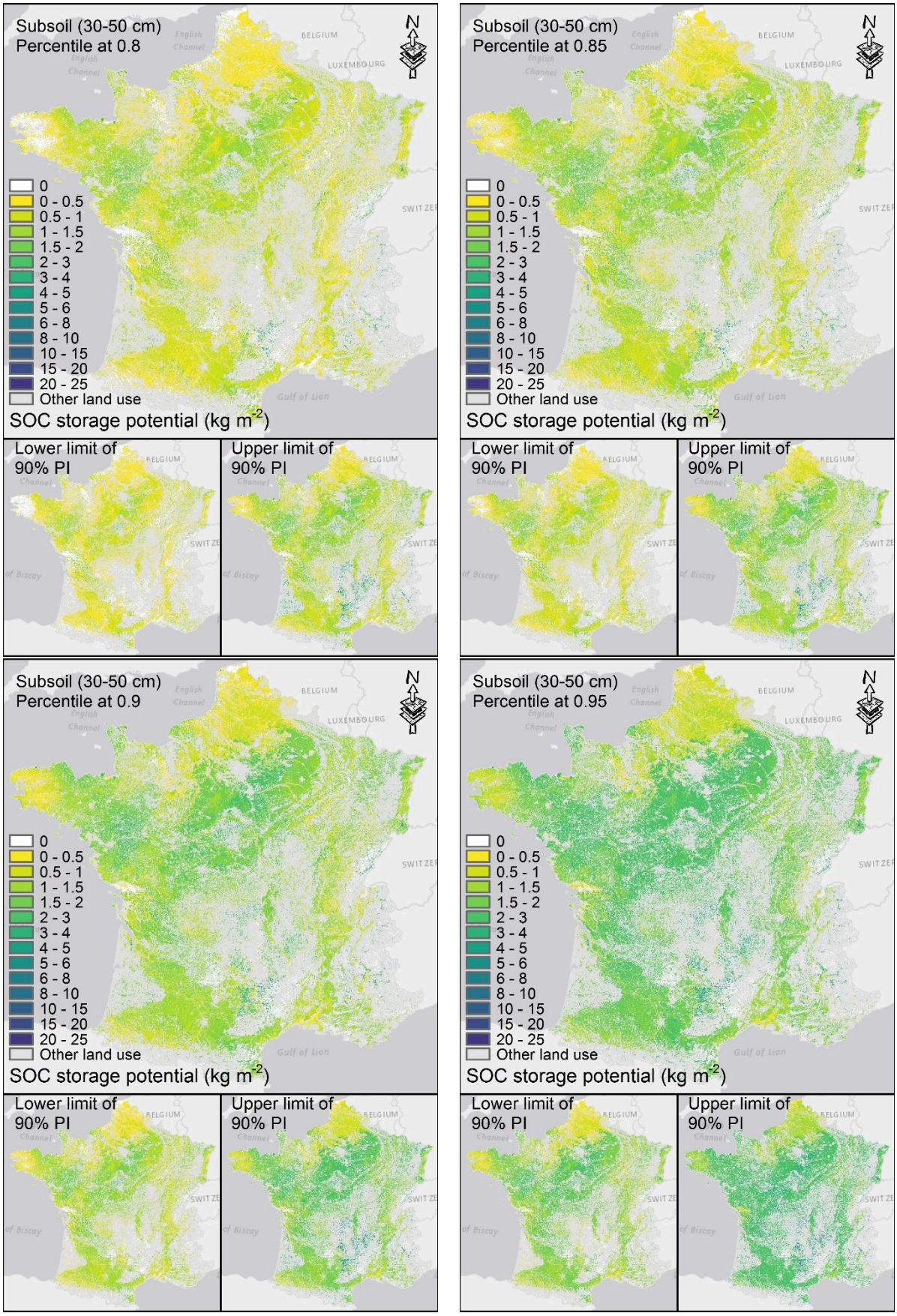


Figure 7.10 SOC storage potential for arable subsoil under four percentile settings.

Table 7.1 National SOC storage potential stocks of French arable soils under different percentile settings. Lower limit and upper limit of 90% CIs are also provided.

Soil horizon	Area (km²)	Total SOC storage potential under four percentile settings (Mt)			
		0.8	0.85	0.9	0.95
Topsoil	239,395	336(203, 501)	470(308, 662)	674(434, 950)	1020(740, 1,283)
Subsoil	228,467	165(91, 250)	228(150, 306)	309(226, 404)	433(331, 560)

7.4 Discussion

7.4.1 Optimizing and mapping Carbon Landscape Zones

The estimates of SOC storage potential using a data-driven approach were based on a stratification of the study area using the CLZs, therefore a procedure for optimizing the number of CLZs was necessary. We observed a negative trend between the number of clusters and BIC, which indicated that using more clusters allowed to explain more variance of our covariates. However, as soil data was finite, creating too many clusters would have resulted in fewer soil data available for each CLZ. We assumed in this study that performing a statistical analysis with less than 10 samples was not robust; therefore we decided to optimize the number of clusters by considering a trade-off between the BIC value and the number of RMQS sites located within each cluster. For building the CLZs we used empirical functions from the literature. Thus, the final delineation of these CLZs might be sensitive to the coefficients we used for these functions which have not been directly validated for France. Interestingly, though using a very different set of covariates and soil point data (different covariates, and a much larger number of soil point data), Mulder et al. (2015) found that the same number of clusters (10) was optimal to partition points data into soil-landscape systems relevant to SOC. Moreover, their maps showed rather similar spatial patterns (e.g. in the

Mediterranean region, mountains, and western France). These findings suggest that our CLZ delineations are relevant for SOC in France.

7.4.2 National SOC storage potential

As expected, the percentile setting had a strong influence on the estimation of SOC storage potential (Table 7.1). If we use the national SOC storage potential at a percentile of 0.8 as a benchmark, the total SOC storage potential at percentiles of 0.85, 0.9 and 0.95 were 1.40, 2.01 and 3.04 times larger, respectively, in topsoil and were 1.38, 1.87 and 2.62 times larger, respectively, in the subsoil. Clearly, the estimates of SOC storage potential are very sensitive to the percentile chosen, especially at high values setting (e.g., 0.95).

7.4.3 Limitations of the data-driven approach

The data-driven approach has previously been implemented in a few pedoclimatic regions to estimate SOC storage potential. Stolbovoy and Montanarella (2008) used data from the European Soil Portal database to determine the maximum observed SOC stocks for a given soil type under a given climate, from which they subtracted the observed SOC stocks under cultivated land. Lilly and Baggaley (2013) determined for each typological soil unit the observed maximum SOC stocks, from which they subtracted the observed median SOC stock under cultivated topsoils. One main difference between these studies and the present one is that they did not calculate percentiles but used only as reference the maximum observed values which are obviously much more sensitive to the presence of very high values. Another difference is that they used coarse resolution data, some of which may not always be directly related to controlling factors of SOC (e.g., soil type, highly aggregated data for delineating large bioclimatic regions).

We show here that this approach has some limitations. It is very sensitive to percentile setting. This is partly attributable to the fact that the SOC distributions are highly skewed with long tails at high SOC values (e.g., CLZs 1, 3, 4, 6, 8 and 10,

see Figure 7.7 and Figure 7.8). This approach could be also considered as data 'hungry'. This sensitivity is also linked to the fact that we have a rather limited number of observations for some CLZs, especially those with a small crop land area (e.g. CLZs, 4 and 7, see Figure 7.8), which hampers the robustness of the data-driven approach. Another limitation may come from the fact that some cultivated soils may have been recently converted from other land uses (e.g., grassland, forest) and may not have yet reached an equilibrium level, which could partly explain the long tails that we observed. One alternative approach would consist in performing dedicated sampling in the CLZs following a probability sampling (number of samples are proportional to the area of clusters) as suggested by De Gruijter et al. (2015). In this approach, the number of sites is selected with a minimum number in order to get precise estimates of the quantiles.

In addition, Barré et al. (2017) already mentioned two other limitations. Firstly, this approach provides an estimate of soil storage potential under present management practices, therefore this estimate could be largely underestimated when new SOC aggrading techniques are adopted. As discussed by Sparling et al. (2003), current management practices may strongly affect the outcomes of a data driven approach when deriving desirable soil organic carbon contents from the median of observed SOC contents. Secondly, another limit of data-driven approaches would be that, for most available databases, management practices are not documented, and thus make it difficult to determine their influence (Barré et al., 2017). Indeed, in some cases there is still a large diversity of soils within a same CLZ and also very different land use histories which are not considered in this approach. The influence of these two factors on the potential storage maps can be easily seen for instance for western France (CLZ2, characterized by a gradient linked to the date of grasslands conversion to croplands). Similarly, the gradients observed in piedmont areas may be linked to the fact that large parts of them have been more or less recently converted from

forest or grassland to cropland (e.g., Arrouays et al., 1994, 1995a, 1995b; Saby et al., 2008) and thus still have quite large SOC stocks reflecting their past land use. Finally, a CLZ may include very different agricultural production systems and in some cases reaching the storage potential would not only require to change the management practices, but the whole production system. The estimates we provide may be refined in the future by taking into account the different agricultural production systems (for CLZ with enough sites).

Despite these limitations, we consider that this first national approximation of SOC storage potential is valuable in making use of a detailed and robust nation-scale database. We further point out some operational advantages of the data driven approach in section 7.4.6.

7.4.4 Complementarity with other approaches

Using a method based on the carbon saturation equation of Hassink (1997), Chen et al. (2018) estimated the SOC sequestration potential in mainland France using the same RMQS data. In their work, the concept of SOC sequestration potential referred to the additional SOC associated with soil fine fraction (< 20 μm), assumed to have pluri-decadal residence times. Their results showed that arable topsoil and subsoil could theoretically sequester 646 Mt and 752 Mt SOC, respectively. Though SOC associated with the soil fine fraction does not represent the total SOC, their estimate of SOC sequestration potential in arable topsoil was close to the percentile of 0.9 derived SOC storage potential (674 Mt), suggesting that SOC sequestration potential can hardly be reached under current management practices. The maps of SOC sequestration potential obtained applying Hassink's equation (Chen et al., 2018) and the maps of SOC storage potential obtained through the data driven approach show rather good qualitative agreements in the western part of France. However, noticeable differences are observed in mountain areas and in the most clayey CLZs for which the data driven approach predicts a much lower additional storage potential than the theoretical SOC sequestration potential. Apart from the fact that the two maps rely on

different concepts (sequestration and storage, e.g., Barré et al., 2017; Chenu et al., 2019) and different modes of calculation, this may also suggest that the pedoclimatic conditions in rather cold or clayey situations do not allow to reach the theoretical SOC sequestration potential because of insufficient plant biomass inputs. In arable subsoil, SOC storage potentials derived from a data-driven approach (under all percentiles) were much lower than C-saturation theoretical SOC sequestration potential. This may be attributed to the fact that the present data-driven estimate of SOC storage potential is based on current land management practices, while reaching the estimated SOC sequestration potential for subsoil may need more advanced land management practices with more potential to raise SOC in both topsoil and deeper layers (Chenu et al., 2019). This may be also simply due to the fact that the French pedoclimatic conditions do not allow to reach the theoretical SOC sequestration potential as assessed by the C saturation concept. These two approaches (i.e. C-saturation theoretical SOC sequestration potential and SOC storage potential) are complementary. We are aware that using Hassink's approach, these values may not be reached in cropland soils, even when managed in an optimal way. However, Hassink's approach may be useful to identify potential biophysical limitations to sequester additional SOC. Both approaches are meaningful and can be used complementarily.

As suggested by Barré et al. (2017), the model-driven approach would be another way of estimating SOC storage potential. In a model-driven approach, process-based models are used for determining highest reachable SOC stocks by simulating different management scenarios. Such an approach has been applied to EU by Lugato et al. (2014). Compared to a data-driven approach, this process-based model may be more suitable as it is able to monitor SOC stock dynamics. However, there are also some limitations to this model-driven approach: i) a lot of input data is required for modelling, for instance, a CENTURY model needs site-specific precipitation, temperature, soil texture, bulk density, initial SOC, land use and corresponding management practice; ii) the initialization for C

dynamic models is still very problematic and the simulation for large dataset is time-consuming; iii) the accuracy of C dynamics model prediction needs to be validated by resampled soil data and (iv) the soil management options considered are limited to those accounted for in current SOC dynamics models (e.g. agroforestry may not be considered in most models).

7.4.5 SOC storage potential and the 4 per 1000 goal

Based on our current SOC stock maps shown in Figure S7.2, the total SOC stocks are estimated at 1.37 Gt and 1.81 Gt for French arable soils for the 0-30 cm layer and the 0-50 cm layer, respectively. If we base these estimates on the total area of French arable soils, reaching the 4 per 1000 aspirational target would require a storage rate of 5.48 Mt C $year^{-1}$ for 0-30 cm, or 7.24 Mt C $year^{-1}$ for 0-50 cm. According to the C storage rate for 0-30 (0-50) cm, it would take 61 (69), 85 (96), 122 (135) and 186 (200) years to reach the SOC storage potential under percentiles of 0.8, 0.85, 0.9 and 0.95, respectively. Thus our data-driven estimates of C storage potential suggest that achieving an annual rate of increase of 0.4% would have to be maintained for decades before reaching the SOC storage potential of these soils, provided that relevant management options can be implemented for such an annual SOC storage, and keeping in mind that an equilibrium level may be reached after a few decades.

7.4.6 The data driven approach, a potentially operational tool

We observed that that SOC storage potential is very sensitive to the percentile used in the calculation. We suggest that this approach offers potential for operational purposes as it enables to set targets of SOC carbon storage for both policy makers and farmers. For instance, decision-makers may decide to implement policies aiming at reaching a minimal objective (for instance, all sites should reach the 0.6 percentile), an intermediate objective (0.8 percentile) or an ambitious objective (0.9 percentile). It could therefore be a very suitable tool to determine to which extent soils can contribute to Intended Nationally

Determined Contributions (INDCs). As an additional step, more emphasis should be put both on policy and recommendations to reach these objectives for different soils, agricultural productions systems and land use histories within each CLZ, and ultimately on developing methods to verify that the targeted objectives are reached. This approach could then be further used to improve the data-driven approach and to design future objectives. Similarly, at a local scale, farmers may compare their present SOC stocks to the theoretically reachable ones within their CLZ, and decide which goal may be reachable by implementing more or less drastic or costly changes to their management practices. They may even find out that the SOC stocks at their farm level are already close to the maximal reachable value, and thus concentrate on not losing SOC rather than on trying to increase the current stocks.

7.5 Conclusions

We tested a data-driven approach to estimate SOC storage potential based on carbon-landscape zones for arable soils using the French National Soil Monitoring Network. Under the trade-off between the BIC index and available data for robust statistics, the optimized number of carbon-landscape zones was determined at 10, using monthly net primary production, climatic decomposition index, and clay content data. The national SOC storage potential varied from 336 Mt to 1020 Mt for topsoil and from 165 Mt to 433 Mt for subsoil under four percentile settings (0.8, 0.85, 0.9 and 0.95), which shows that the data-driven approach is very sensitive to the selected percentile. This sensitivity was partly attributable to a rather low number of observations in some carbon-landscape zones and mainly to skewed distributions with long tails of high SOC contents. However, we argue that this data driven approach offers meaningful advantages from an operational point of view, as it enables to adapt targets of SOC carbon storage by taking into account both policy makers' and farmers' considerations. We also argue that the data driven approach is also a convenient way to provide

quantitative estimates of the SOC storage potential over large areas having widely distributed soil data. Dedicated surveys and research on management practices effects are still necessary in order to better estimate the reachable SOC stocks and the feasibility of their implementation.

Further work will focus on estimating SOC storage potential by the model-driven approach in mainland France. Producing model-driven estimates may enable to determine a more reliable percentile setting for the data-driven approach and thus provide references for the regions where exhaustive data for applying process-based models is not available.

7.S Supplementary materials

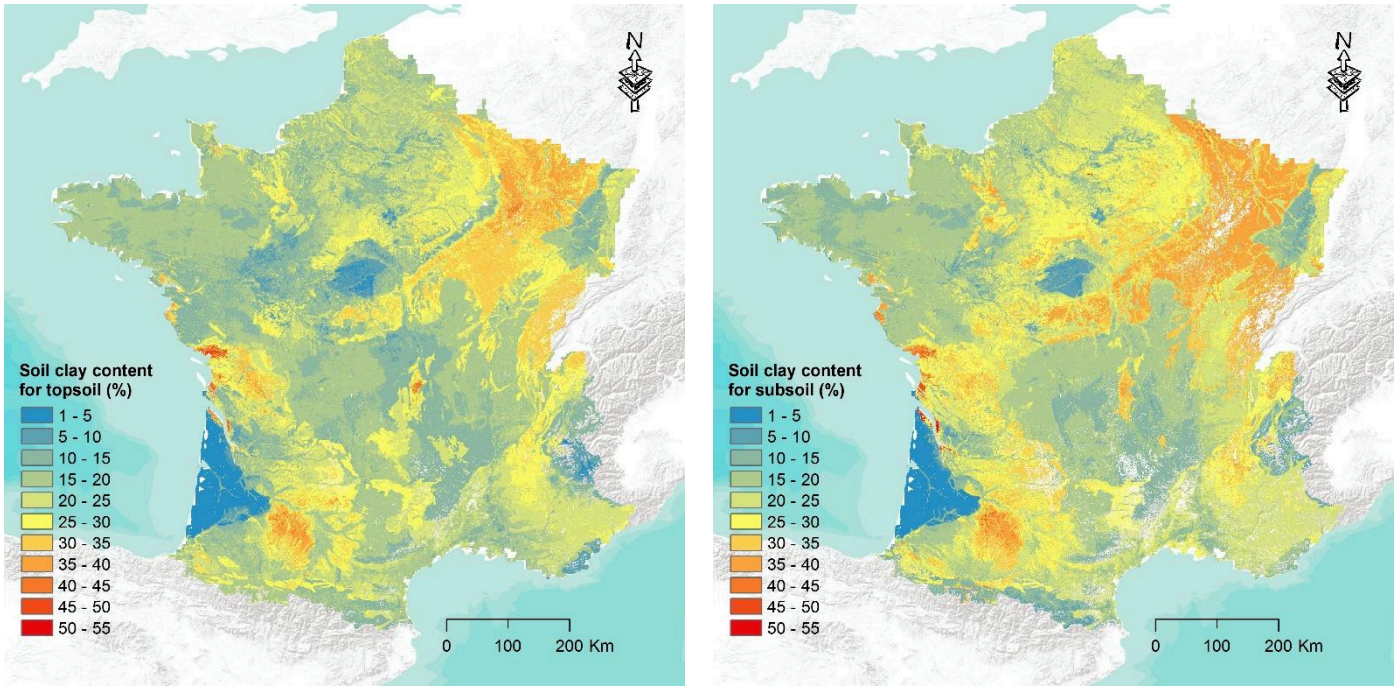


Figure S7.1 Soil clay content for topsoil and subsoil

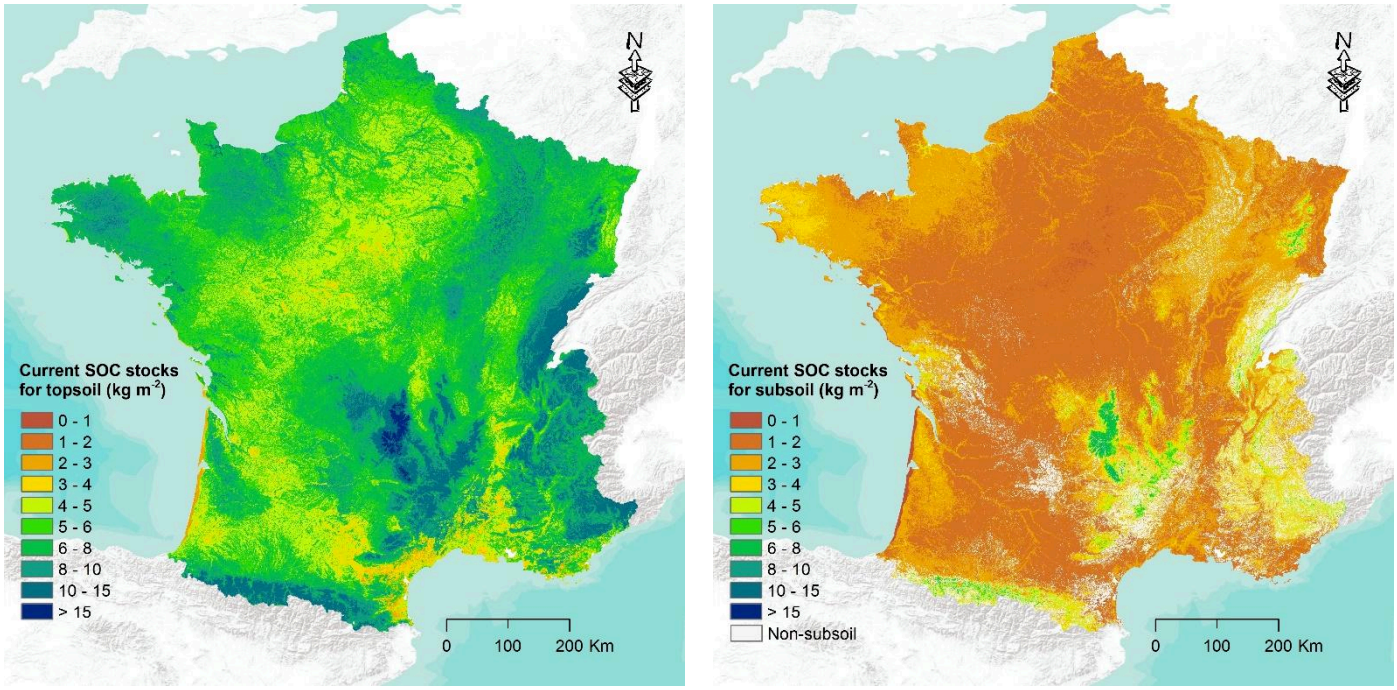


Figure S7.2 Current SOC stocks for topsoil and subsoil

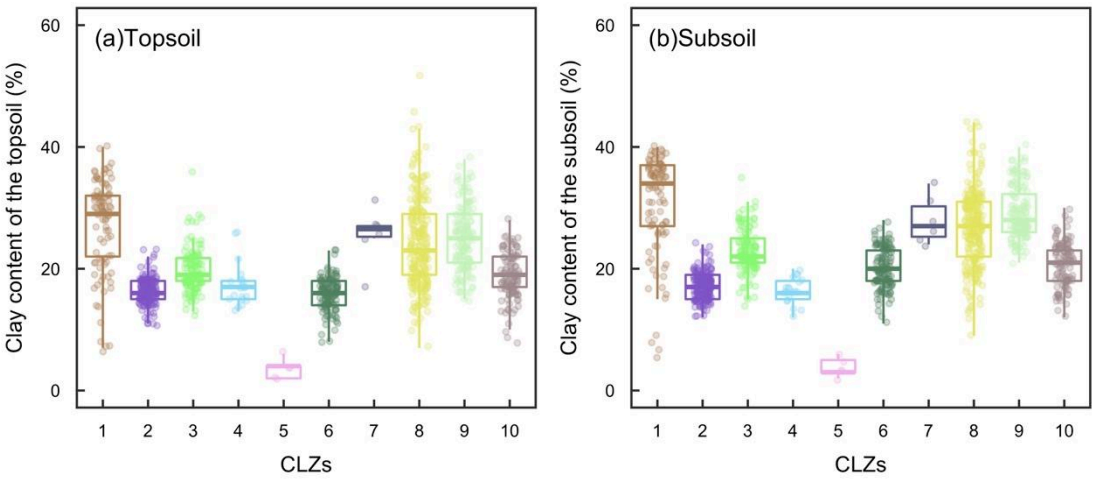
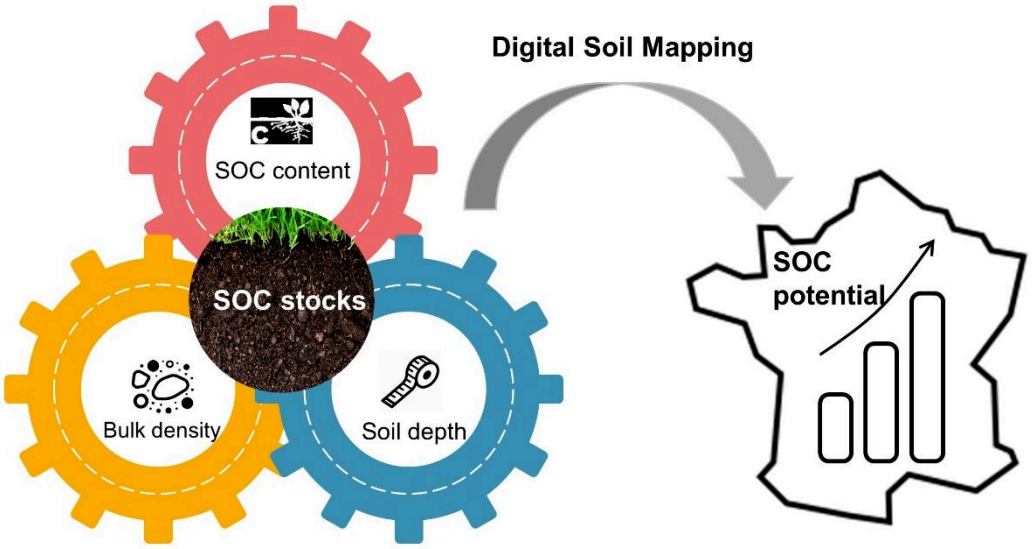


Figure S7.3 Boxplots of clay content in topsoil and subsoil under the 10 carbon-landscape zones.

Chapter 8

Conclusions and perspectives



8.1 Introduction

In **Chapter 1**, I discussed the main drivers for the rise and development of digital soil mapping (DSM), and introduced the core of DSM, the *scorpan* model. I also gave a brief history about DSM and discussed the achievements and challenges for the DSM community, which was then more detailed in the review of **Chapter 2**. By addressing soil property of interest (SOC) and the related challenges, I briefly introduced the main objectives and the structure of this thesis.

In this final **Chapter**, the major findings and remaining issues from **Chapter 3** to **Chapter 7** are discussed in Section 8.2. Section 8.3 presents the directions for future studies related to broad-scale DSM, and Section 8.4 gives final considerations about this thesis.

8.2 Overview of findings and remaining issues

The overall aim of this thesis was to improve national SOC map and to assess the potential of soil to store or sequester additional SOC using DSM and statistical models. The objectives were addressed in five aspects (Figure 1.1) from **Chapter 3** to **Chapter 7**, which were stated in **Chapter 1**.

In **Chapter 3**, I tested the potential of model averaging for improving French topsoil (0-20 cm) SOC content map by merging existing SOC maps produced at national, continental, and global scales. All five model averaging approaches improved the national SOC map when using more than 100 soil samples for calibrating the model averaging approaches. Using 200 calibration data uniformly spread over France was efficient for model averaging approaches. The results also suggested that merging SOC maps using model averaging is also applicable to data-poor situations and might thus be attractive to data-poor countries to define an affordable sampling strategy in order to build country-based predictions of SOC.

In **Chapter 4**, easy-to-measure soil properties, including SOC, pH, particle size fractions and gravel content, were used to build machine leaning based pedotransfer functions (PTFs) for predicting bulk density (BD) in Region Centre of France. Compared to six groups of revised PTFs, machined leaning based PTFs performed better with a reasonable predictive accuracy. The validity domain of PTFs determined by Standardized Euclidean distance was tested over mainland France, and the results showed that those samples beyond the validity domain should avoid to be predicted by PTFs. The results also showed that integrating additional sampling soil samples exceeding validity domain into PTFs improved predictive ability. The use of validity domain of PTFs was able to avoid invalid prediction of BD and thus can be used to choose additional point measurements in order to reduce the uncertainty later incorporated to the calculation of SOC stocks. This approach is already implemented in France for defining a new sampling strategy to improve PTFs in predicting available water capacity (Román Dobarco et al, 2019b).

Chapter 5 introduced the use of random survival forest (RSF) in soil thickness (ST) (or soil depth) probability modelling to deal with right censored data for DSM. RSF produced a probability function of ST for each soil observation as well as each unvisited location. This function allowed the estimation of a probability of exceeding a given ST, indicating each soil location was right censored or not. The model evaluation indicated an overall good performance of RSF to predict the probability of exceeding the six *GlobalSoilMap* standard depths (5, 15, 30, 60, 100 and 200 cm). The RSF proved suitable for using right censored soil data for DSM study, and the produced probability map (i.e., 200 cm) can be used as a knowledge for future sampling design to improve our knowledge in deep soil. One remaining question is how to use these probability of exceeding a given depth to predict the ST (or soil depth) in all locations. There are three possible solutions: (1) use the ST extracted from the median probability in the predicted function; (2) use the ST extracted from a fixed probability, allowing the

classification of censored and actual ST at high accuracy among RSF calibration datasets; 3) perform a derivative analysis on the probability curve.

In **Chapter 6**, I estimated SOC sequestration potential using Hassink's equation and produced fine resolution maps using DSM for topsoil and subsoil in mainland France. The 10-fold cross-validation results indicated good performances for two SOC sequestration potential maps. The regions with high sequestration potential in topsoil were mainly located in intensively cultivated cropland, vineyard/orchard in the Mediterranean region and clay-rich soils in northeastern France. Subsoil had higher SOC sequestration potential than topsoil under all land covers, and thus more attention can be paid to the management practices with potential to raise the SOC in deeper layers.

In **Chapter 7**, I tested a data-driven approach to estimate SOC storage potential based on carbon-landscape zones in French arable soils. The carbon-landscape zones were determined by net primary production, climatic decomposition index, and clay content. Though the data-driven approach is sensitive to the selected percentile, it offers meaningful advantages from an operational point of view, as it enables to adapt targets of SOC carbon storage by taking into account both policy makers' and farmers' considerations. Besides, the data driven approach is also a convenient way to provide quantitative estimates of the SOC storage potential over large areas having widely distributed soil data. Model-driven approach can be further explored to estimate SOC storage potential under different management practices, and enable to determine a more reliable percentile setting for the data-driven approach. When comparing results obtained by mapping SOC sequestration potential (**Chapter 6**) and SOC storage potential (**Chapter 7**) based on present practices, the differences between the two maps were very large and suggested that under present land use and management, the theoretical SOC sequestration potential cannot be reached, especially in subsoil. This result also advocates for more research on soil management practices enabling to provide more C inputs in deep layers.

Finally, I think that this thesis addresses some different issues which together contribute to significant advances in DSM and in digital soil assessment (DSA). In **Chapter 2**, I made a preliminary review of broad-scale DSM in various situations. Although it gives and overview of the data and methods used, it should be considered as a preliminary synthesis. One of my next objectives is to transform this first attempt to a more in-depth review paper. In **Chapter 3**, I tested a way to merge map predictions from several data sources and scales. This constitutes a first step towards harmonization and taking advantage of global and country-based predictions that capture different controlling factors depending on their scale, on the point data and on covariates used. This study paves the way for integrating different predictions that capture different controlling factors. Moreover, it gives a practical example on how data-poor countries could efficiently optimize rather cheap sampling campaigns to significantly improve country-based predictions. In **Chapter 4,** I addressed the issue of the validity domain of PTFs. This may be considered as a methodological contribution for areas where data for building such PTFs is rather sparse, or localized only in part of a country. The example I choose is purely methodological, as BD measurements are available for each of the RMQS grid in France, but it brings some insights on how to check the validity domain of a PTF and how to design new campaigns to improve it. In **Chapter 5**, I tested a new method to deal with censored data about soil thickness. This is important because a large proportion of legacy data are likely censored due to digging depth limitations. Further works should explore how to use these probability of exceeding a given depth to predict the ST (or soil depth) in all locations.

Chapter 6 and Chapter 7 are first steps to move from DSM to DSA. Instead of predicting the present status of a soil property (SOC), I tried to map its theoretical potential to increase. More efforts are needed to assess which changes in land use and/or management practices could help to reach increases in SOC and to which level, where and under which conditions these changes are feasible.

A recent published scientific report on assessing the feasibility of 4p1000 in France is a good example for addressing these challenges (Pellerin et al., 2019). Another challenge is to assess how the effects of these changes could be incorporated in dynamic, process-oriented models (e.g., Century, RothC). Models at this level have higher soil, climatic and management data requirements which may make them difficult to apply for national-scale or global-scale digital soil assessment. Due to different limitations of data-driven and model-driven approaches, we should consider how both approaches could be complementary and interact. For example, we can think how to conciliate data-driven approach with the best available proxies of the data required to implement model-driven approach.

8.3 General perspectives about DSM

I start this section by discussing the "Pedometrics Challenges" proposed by Heuvelink (2019). Though these 10 challenges were initially suggested for the general Pedometrics community, they inspire me a lot in the context of DSM and DSA. Hereafter, I list the challenges that relevant to my thesis and give my thoughts and suggestions for each challenge from the scope of a digital soil mapper. After this, I will also discuss other perspectives related to DSM and DSA.

8.3.1 Relevance to some of the "Pedometrics Challenges"

Challenge 2: ***Can we develop communicable measures of uncertainty?***

Recognizing the importance of uncertainty, more and more studies have provided the confidence intervals (CIs) or prediction intervals (PIs) for predicted digital soil maps. Taking the examples from my thesis, 90% CIs have been reported in **Chapter 5** (ST), **Chapter 6** (SOC sequestration potential) and **Chapter 7** (SOC storage potential). However, many end users do not care or understand our measures of uncertainty. Therefore, we have the responsibility to better communicate on uncertainty, on why quantified uncertainty is important, and how it can be used. From another side, the demands and feedbacks from the

end users can continually simulate the development of DSM community and lead DSM to a more operational tool in decision making and risk analysis. For this, CIs or PIs may not be the best tools to incorporate uncertainty in moving from DSM to DSA. We have to develop further methods to be able to predict complete probability distribution functions which should be a better input to run models dealing with risk analysis. Then, the consequences of uncertainty could also be communicated using the outputs of the DSA models, rather than communicating on the uncertainty of input data. This will certainly lead to more understandable consequences of uncertainty by end users.

Challenge 3: ***Can we develop sound scaling methodologies?***

We are still confused about the concept of scale mainly due to its poor definition. In the meantime, most biogeochemical models (e.g., Century, RothC) are developed at a plot scale, so how these mechanical models can be applied to regional, national or even global scale by up-scaling model parameters remains challenging. This issue is highly related to the research topics on projecting SOC changes under climate scenarios and predicting SOC storage potential by model-driven approach (perspectives mentioned in **Chapter 7**). Two directions can be considered for the future studies: (1) explore the potential of using easy-to-measure soil information, i.e., Rock-Eval thermal analysis (Cécillon et al., 2018), to quantify the model parameters, and thereby to initialize model parameters at a broad scale; (2) build more agro-pedo-climate comprehensive long-term experimental fields to develop novel biogeochemical models that take spatial variations and processes into account.

Challenge 4: ***Can we incorporate mechanistic pedological knowledge in digital soil mapping?***

We heavily rely on machine learning for DSM in current studies, while pedological knowledge is often only used to help us to identify relevant environmental covariates. We will completely lose ourselves in computer science and statistics and also lose scientific credit if we continue to ignore the

pedological knowledge in DSM. Fortunately, the DSM community realizes this issue and several recent studies on Structural equation modelling (SEM) and mechanistic soil evolution model provide good examples to incorporate pedological knowledge in DSM though the model accuracy may be lower than machine learning based models (Angelini et al., 2017; Ma et al., 2019). In the future studies, more attempts should be focused on this direction,

Challenge 5: *Can we make sufficiently accurate global soil maps?*

Large achievement has been made in *GlobalSoilMap* initiative for delivering global soil information at 90 m resolution. As stated by Heuvelink (2019), the map spatial resolution is easier reached than accuracy, and we should aim to make global soil maps that both satisfy the resolution requirements and pre-defined accuracy standards. We should also notice that the pre-defined accuracy standards may change a lot for different purposes and various scales, for example, a R^2 of 0.6 would be enough for global C modelling while it is still a little bit low for decision making of management practice at a local scale. Considering the fact that the map accuracy remains low even for some crucial soil properties (i.e., SOC), several solutions may be helpful to improve map accuracy: (1) collect more soil data using optimal soil sampling designs; (2) explore the potential environmental covariates from remote sensing and airborne data (e.g., Sentinel 2, Sentinel 3, RISAT-2B Earth-Observation Satellite and gamma-ray); (3) test novel spatial predictive models (e.g., deep learning, SEM).

Another issue related to this challenge is that the accuracy of national or global soil maps are assessed at a broad scale while the extracted soil maps for local usage may have totally different accuracy. Therefore, there is a need to assess the accuracy of extracted soil map locally, which needs additional soil sampling with affordable cost. If the local or global soil maps do not satisfy the accuracy requirement, the model averaging approach mentioned in **Chapter 3** may help to improve the national and/or global soil maps with a better accuracy. Moreover, this approach may be considered as a first step to integrate predictions at various

scales and to progress in products harmonization.

Challenge 7: ***Can we quantify uncertainty in soil observations and analyze how this affects soil mapping?***

It is well known that soil observations and laboratory soil analysis inherently contain some measurement errors, which vary among different observations and analytical methods as well as different soil properties. In current studies, field observations and laboratory measurements are used as "true" soil observation, and the measurement error is often ignored. In order to deliver high quality soil information, characterizing, quantifying and reporting soil measurement error are necessary. The rapid extension of soil predictions derived from proximal soil sensing, which have substantial uncertainty, makes this issue ever more important. The soil measurement or prediction error should further be taken into account in DSM studies as this uncertainty will be certainly propagated to the final digital soil map.

This challenge is highly relevant to my thesis. In **Chapter 4**, the proposed validity domain is able to exclude the soil samples that are not applicable for PTFs, which may reduce the uncertainty of BD prediction that is later propagated in the calculation of SOC stocks. In **Chapter 5**, the random survival forest proves its ability in modelling and mapping uncertain soil data (censored ST) and the resulted probability map can be used for additional sampling design to reduce the map uncertainty of ST in France. In **Chapter 6**, the average proportion of SOC in fine fraction is used to calculate SOC sequestration potential using Hassink's equation. The usage of the average of SOC in fine fraction certainly brings uncertainty in modelling and mapping SOC sequestration potential, of which I also simulated its consequences in the Chen et al. (2019b) that is not included in this thesis. In **Chapter 7**, the historical land use change (i.e., forest or grassland to arable soil) brings a large uncertainty in the estimates of national SOC storage potential. This uncertainty may be reduced by integrating historical land use change from remote sensing data on the condition of being able to estimate the

consequence of these changes on the SOC dynamics. Besides, the model-driven approach (e.g., Century, RothC) that simulates SOC changes under different land management practice, will help to choose a more realistic percentile for data-driven approach, and thus reduce the uncertainty related to percentile setting.

This challenge also calls for the discussion on the issue that should we (1) compute SOC stocks using each component (i.e., SOC content, BD, ST, gravel) first and then produce the SOC stock map or (2) produce the maps for each component first and then merge them into the SOC stock map. Two approaches have been both used in DSM practice, however there is no clear answer yet. The outputs from this discussion will provide a guideline on how to minimize the uncertainty of SOC stock map as well as for the digital maps of other soil functions.

Challenge 8: ***How to map soil functions?***

Most focuses are still on modelling and mapping soil class and primary soil properties in the DSM community and less attention has been paid on soil functional modelling and mapping (e.g., crop yield, additional SOC storage potential, and water purification and regulation). In fact, many end users require maps of soil functions for modelling and decision making in practice, and thus we have to put greater efforts in mapping soil functions with quantified uncertainty. In order to deliver better evidence-based soil functional maps, more links should be established with other disciplines for assessing and mapping soil functions. Similarly, the moving from DSM to DSA is seen as a trend in the next decade. The **Chapter 6** on SOC sequestration and the **Chapter 7** on storage potential are good examples on how to transform predictions of a soil property in predictions of a potential soil function. However, they remain estimates of a theoretical potential and still miss a link with references indicating how, where, and to which level this potential can be reached.

Challenge 10: ***What can we learn about soil processes from calibrated machine learning models?***

Machine learning has been widely used in modelling and mapping soil classes and soil properties while we only do it for prediction presently. Though many machine learning algorithms provide a variable importance index to help us identify the main controlling factors, they still have not been often used to help us improve the understanding of soil variations, confirm pedological knowledge or reveal new insights. In some cases of classification and regression trees, mapping the rules of splitting the covariates may help to improve our understanding (e.g., Viscarra Rossel et al., 2014, 2019). Ultimately, mapping the residuals of predictions (adopted in **Chapter 6** and **Chapter 7**) may help to capture spatial structures or gradients suggesting the existence of controlling factors that were not included in the modelling process. More thoughts are needed to benefit more about pedological knowledge from machine learning.

8.3.2 Other ways forward

1. Data privacy

In the context of modelling and mapping soil information at a global scale, the issues related to data privacy become more crucial than ever before (Arrouays et al., 2019, submitted). These issues mainly results from the different legislations between countries, and fears by data holders about losing the intellectual property of their data. Apart from trying to sign specific data agreement between different partners, merging predicted soil maps without sharing the original soil point data (i.e., *GlobalSoilMap*) may be a solution (see model averaging approach in **Chapter 3**). Another recently proposed solution is merging models without sharing data using a block-chain approach (Padarian and McBratney, 2019). These solutions will definitely make a great process in completing the *GlobalSoilMap* products as well as improving SoilGrids products for data-sparse regions.

2. Capacity building and training

As mentioned in **Chapter 1**, DSM has progressively moved from academic and research activity to more operational activity in delivering soil information to both scientific community and decision and policy makers (Minasny and

McBratney, 2016). Being an evidence-based technique that is part of the everyday job (operational), there is a crucial need for training DSM techniques for young scientists and staffs involved in land management. Besides, the capacity building and training is a solution to help local decision and policy makers to better understand digital soil maps with their associated uncertainty (Challenge 2), and even allow them to produce local soil maps without sharing their soil data (data privacy issue). Large efforts have been made by some institutions (e.g., FAO-GSP, ISRIC-World Soil Information, University of Sydney, and NRCS-USDA) to disseminate DSM knowledge and tools, and these efforts should be pursued and more actors should be involved in capacity building and training. A relevant strategy could be to involve in same teams proficient traditional soil surveyors and new generation of soil-data scientists skilled in DSM methodology. Traditional soil surveyors could improve DSM by performing detailed field observations, controlling and harmonizing legacy soil observations, helping to choose the relevant covariates for DSM, and finally checking the consistency of final predictions and/or identifying controlling factors of soil properties that were not included in the covariates.

3. Multi-sensor fusion

Large efforts have been made in using remote sensing data, airborne hyperspectral data and proximal soil sensing data for DSM and digital soil monitoring. These data are often used separately due to the data availability, specific study purpose, and scale of variability. Instead of choosing the best one, there is a new direction in DSM using multi-sensor fusion. Progress in remote sensing, airborne imaging and proximal soil sensing, along with the development of national, continental and global soil spectral libraries (Viscarra Rossel. et al, 2016), it is promising for the use of multi-sensor fusion techniques in improving DSM and DSA. A nice example for integrating new remote sensing data (Sentinel 2) in national predictions of clay content has been recently demonstrated by Loiseau et al. (2019).

4. Principle of parsimony for covariates selection

More and more environmental covariates have been used for DSM in the recent years, resulting from the improved data availability. In this context, the principle of parsimony is becoming more crucial for the covariates selection in DSM than ever before. I admit that using more environmental covariates may sometimes improve the model accuracy, especially in machine learning. However, we should also notice that this practice may tend to introduce more uncertainty from input data and make the results less interpretable from a soil science point of view. Therefore, based on the principle of parsimony, covariates selection is necessary in DSM, not only based on a pure statistical selection, but also on their pedological relevance.

8.4 Final considerations

In this thesis, after a general review about broad-scale DSM, I first brought some substantial advances on modeling some soil properties that are relevant to general DSM objectives and to the calculation of SOC stocks at a national level. I showed an example on how various predictions of a soil property can be merged using ensemble methods and provided inputs on how to take advantage of global predictions in 'data-poor' countries. I then focused on the validity domain of PTFs used for BD predictions and on developing a novel approach to deal with censored ST and produce probability map of exceeding a given ST. Then I moved from DSM of soil properties to DSA of soil functions, exemplified by SOC sequestration and storage potentials. All these works contribute to improving some aspects related to DSM and *GlobalSoilMap* and illustrate how DSM can evolve into DSA of soil functional properties. Then I finish this thesis by discussing the most important Pedometrics challenges that relate to my work. I outline the inputs that my work provided to reaching these challenges and highlight the remaining issues to be solved in the near future.

References

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Scientific Data, 5, 170191.

Achache, J., Debeglia, N., Grandjean, G., Guillen, A., Le Bel, L., Ledru., P., Renaud, X., Autran,A., Bonijoly, D., Calcagno, P., Pluchery, E., Guennoc, P., Truffert, C., Rossi, P., Vairon, J.,Avouac, J.P., Poli, E., Senechal, G., Brun, J.P., Galdeano, A., Diament, M., Tarits, P., Mervier, J., Paul, A., Poupinet, G., Marquis, G., Bayer, R., Chautra, J.M., 1997. GEOFRANCE 3D: l'imagerie geologique et geophysique 3D du sous-sol de la France. Mémoires de la Société Géologique de France, 172, 53–71.

Adams, W.A., 1973. The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. Journal of Soil Science, 24, 10–17.

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services—A global review. Geoderma, 262, 101–111.

AFNOR, 1992. Qualité des sols – Méthodes physiques – Mesure de la masse volumique apparente d'un échantillon de sol non remanié — Méthode du cylindre., NF X31-501. AFNOR.

AFNOR, 1994. Qualité du sol – détermination du pH, ISO 10390:1994. AFNOR.

AFNOR, 2003. Qualité du sol – Détermination de la distribution granulométrique des particules du sol – Méthode à la pipette, NF X31–107. AFNOR.

Akpa, S.I.C., Ugbaje, S.U., Bishop, T.F.A., Odeh, I.O.A., 2016. Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation. Soil Use and Management, 32, 644–658.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital mapping of soil particle-size fractions for Nigeria. Soil Science Society of America Journal, 78(6), 1953-1966.

Aksoy, E., Yigini, Y., Montanarella, L., 2016. Combining soil databases for topsoil organic carbon mapping in Europe. PloS One, 11, e0152098.

Alexander, E.B., 1980. Bulk densities of California soils in relation to other soil properties. Soil Science Society of America Journal, 44, 689–692.

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. The Annals of Statistics, 10(4), 1100–1120.

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., 2017. Multivariate mapping of soil with structural equation modelling. European Journal of Soil Science, 68(5), 575-591.

Angers, D., Arrouays, D., Saby, N., Walter, C., 2011. Estimating and mapping the carbon saturation deficit of French agricultural topsoils. Soil Use and Management, 27, 448–452.

Arrouays, D,. Pélissier, P., 1994. Changes in carbon storage in temperate humic loamy soils after forest clearing and continuous corn cropping in France. Plant Soil 160, 215–223.

Arrouays, D., Balesdent, J., Mariotti, A., Girardin, C., 1995a. Modelling organic carbon turnover in cleared temperate forest soils converted to maize cropping by using 13C natural abundance measurements. Plant and Soil, 173(2), 191–196.

Arrouays, D., Deslais, W., Badeau, V., 2001. The carbon content of topsoil and its geographical distribution in France. Soil Use and Management, 17, 7–11.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014a. Chapter Three — GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. Advances in Agronomy 125, 93–134.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J.W., Hengl, T., Heuvelink, G.B.M., ..., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ, 14, 1–19.

Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Martin, M.P., Bellamy, P.H., Lark, R.M., Kibblewhite, M., 2012. Generic Issues on Broad-Scale Soil Monitoring Schemes: A Review. Pedosphere, 22, 456–469.

Arrouays, D., McKenzie, N.J., Hempel, J., Richer de Forges, A.C., McBratney, A.B., 2014b. GlobalSoilMap: Basis of the Global Spatial Soil Information System. 1st ed. CRC Press Taylor & Francis Group, pp. 9–12.

Arrouays, D., Piggio, L., Salazar, O., Mulder, V.L., 2019. Digital Soil Mapping and GlobalSoilMap. Main advances and ways forward. Geoderma Regional, submitted.

Arrouays, D., Saby, N., Walter, C., Lemercier, B., Schvartz, C., 2006. Relationships between particle-size distribution and organic carbon in French arable topsoils. Soil Use and Management, 22, 48–51.

Arrouays, D., Vion, I., Kicin, J.L., 1995b. Spatial analysis and modeling of topsoil carbon storage in forest humic loamy soils of France. Soil Science, 159, 191–198.

Augusto, L., Bakker, M.R., Morel, C., Meredieu, C., Trichet, P., Badeau, V., ... Ranger, J., 2010. Is grey literature a reliable source of data to characterize soils at the scale of the region? A case study in a maritime pine forest of south-western France. European Journal of Soil Science, 61, 807–822.

Baldock, J., Skjemstad, J., 2000. Role of the soil matrix and minerals in protecting natural organic materials against biological attack. Organic Geochemistry, 31, 697–710.

Balesdent, J., 1996. The significance of organic separates to carbon dynamics and its modelling in some cultivated soils. European Journal of Soil Science, 47, 485–493.

Balesdent, J., Basile-Doelsch, I., Chadoeuf, J., Cornu, S., Fekiacova, Z., Fontaine, S., ... Hatté, C., 2017. Renouvellement du carbone profond des sols cultivés: Une estimation par compilation de données isotopiques. Biotechnology, Agronomy, Society and Environment, 21, 1–9.

Balesdent, J., Besnard, E., Arrouays, D., Chenu, C., 1998. The dynamics of carbon in particle-size fractions of soil in a forest-cultivation sequence. Plant and Soil, 201, 49–57.

Balesdent, J., Chenu, C., Balabane, M., 2000. Relationship of soil organic matter dynamics to physical protection and tillage. Soil and Tillage Research, 53, 215–230.

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

Bargaoui, Y.E., Walter, C., Michot, D., Saby, N.P., Vincent, S., Lemercier, B., 2019. Validation of digital maps derived from spatial disaggregation of legacy soil maps. Geoderma, 356, 113907.

Barré, P., Angers, D. A., Basile-Doelsch, I., Bispo, A., Cécillon, L., Chenu, C., ... Pellerin, S., 2017. Ideas and perspectives: Can we use the soil carbon saturation deficit to quantitatively assess

the soil carbon storage potential, or should we explore other strategies? Biogeosciences Discuss. https://doi.org/10.5194/bg-2017-395.

Barthès, B.G., Kouakoua, E., Larré-Larrouy, M.-C., Razafimbelo, T.M., de Luca, E.F., Azontonde, A., ... Feller, C.L., 2008. Texture and sesquioxide effects on water-stable aggregates and organic matter in some tropical soils. Geoderma, 143, 14–25.

Bates, J.M., Granger, C.W., 1969. The combination of forecasts. Journal of the Operational Research Society, 20, 451–468.

Batjes, N., Sombroek, W., 1997. Possibilities for carbon sequestration in tropical and subtropical soils. Global Change Biology, 3, 161–173.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science, 47(2), 151–163.

Baveye, P., Berthelin, J., Tessier, D., Lemaire, G., 2018. The "4 per 1000" initiative: A credibility issue for the soil science community? Geoderma, 309, 118–123.

Bell, J.C., Cunningham, R.L., Havens, M.W., 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Science Society of America Journal, 58(2), 464–470.

Bellamy, P.H., Loveland, P.J., Bradley, R.I., Lark, R.M., Kirk, G.J., 2005. Carbon losses from all soils across England and Wales 1978–2003. Nature, 437(7056), 245.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A.B., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry, 29(9), 1073-1081.

Benites, V.M., Machado, P., Fidalgo, E.C., Coelho, M.R., Madari, B.E., 2007. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. Geoderma, 139, 90–97.

Bernoux, M., Arrouays, D., Cerri, C., Volkoff, B., Jolivet, C., 1998. Bulk densities of Brazilian Amazon soils related to other soil properties. Soil Science Society of America Journal, 62, 743–749.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma, 91, 27–45.

Bonfatti, B.R., Hartemink, A.E., Vanwalleghem, T., Minasny, B., Giasson, E., 2018. A mechanistic model to predict soil thickness in a valley area of Rio Grande do Sul, Brazil. Geoderma, 309, 17–31.

Botula, Y.-D., Nemes, A., Ranst, E., Mafuka, P., Pue, J., Cornelis, W., 2015. Hierarchical Pedotransfer Functions to Predict Bulk Density of Highly Weathered Soils in Central Africa. Soil Science Society of America Journal, 79, 476-486.

Bouma, J., 2014. Soil science contributions towards sustainable development goals and their implementation: linking soil functions with ecosystem services. Journal of Plant Nutrition and Soil Science, 177(2), 111-120.

Bouma, J., 2018. The challenge of soil science meeting society's demands in a "post-truth","fact free" world. Geoderma, 310, 22-28.

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.

Weil, R.R., Brady, N.C., 2016. The nature and properties of soils. Pearson.

Breiman, L., 2001. Random forests. Machine Learning, 45(1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Chapman and Hall, Boca Raton, FL.

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma, 338, 464–480.

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

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

Brus, D.J., Saby, N.P.A., 2016. Approximating the variance of estimated means for systematic random sampling, illustrated with data of the French Soil Monitoring Network. Geoderma, 279, 77–86.

Calaway, R., Analytics, R., Weston, S., Tenenbaum, D., 2015. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.10. (https://cran.r-project.org/web/packages/doParallel)

Carol Adair, E., Parton, W.J., Del Grosso, S.J., Silver, W.L., Harmon, M.E., Hall, S.A., ... Hart, S.C., 2008. Simple three-pool model accurately describes patterns of long-term litter decomposition in diverse climates. Global Change Biology, 14(11), 2636–2660.

Castellano, M., Poffenbarger, H., Cambardella, C., Liebman, M., Mallarino, A., Olk, D., ... Six, J., 2017. Evaluation of carbon saturation across gradients of cropping systems diversity and soil depth. In EGU General Assembly Conference Abstracts, 19, 10357.

Cattle, J.A., McBratney, A., Minasny, B., 2002. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. Journal of Environmental Quality, 31(5), 1576–1588.

Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., Saby, N.P.A., 2019. Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. Geoderma, 337, 99–110.

Cécillon, L., Baudin, F., Chenu, C., Houot, S., Jolivet, R., Kätterer, T., ..., Savignac, F., 2018. A model based on Rock-Eval thermal analysis to quantify the size of the centennially persistent organic carbon pool in temperate soils. Biogeosciences, 15(9), 2835-2849.

Cerdan, O., Govers, G., Le Bissonnais, Y., Van Oost, K., Poesen, J., Saby, N., Gobin, A., Vacca, A., Quinton, J., Auerswald, K., Klik, A., Kwaad, F.J.P.M., Raclot, D., Ionita, I., Rejman, J., Rousseva, S., Muxart, T., Roxo, M.J., Dostal, T., 2010. Rates and spatial variations of soil erosion in Europe: A study based on erosion plot data. Geomorphology, 122, 167–177.

Chabbi, A., Lehmann, J., Ciais, P., Loescher, H.W., Cotrufo, M.F., Don, A., ... Rumpel, C., 2017. Aligning agriculture and climate policy. Nature Climate Change, 7(5), 307.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., ..., Zhang, W., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. ISPRS Journal of Photogrammetry and Remote Sensing, 103, 7-27.

Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Saby, N.P.A., Walter, C., 2019a. National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones. Science of the Total Environment, 666, 355–367.

Chen, S., Arrouays, D., Angers, D.A., Martin, M.P., Walter, C., 2019b. Soil carbon stocks under

different land uses and the applicability of the soil carbon saturation concept. Soil & Tillage Research, 188, 53–58.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Aeeouays, D., Shi, Z., 2019c. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. Science of the Total Environment, 655, 273-283.

Chen, S., Martin, M.P., Saby, N.P.A., 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., Heuvelink, G., Piggio, L., Caubet, M., Román Dobarco, M., Walter, C., Arrouays, D., 2019d. Model averaging for mapping topsoil organic carbon in France. Geoderma, submitted.

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

Chenu, C., Angers, D.A., Barré, P., Derrien, D., Arrouays, D., Balesdent, J., 2019. Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations. Soil & Tillage Research, 188, 41–52.

Chenu, C., Balanabe, M. Pétraud, J.P., Champdavoine, V., Arrouays, D., 2004. Compartimentation et fractionnement du carbone organique des sols Test de mise en "routine" de la détermination des matières organiques particulaires sur le RMQS.

Chow, A. T., Tanji, K. K., Gao, S., Dahlgren, R.A., 2006. Temperature, water content and wet–dry cycle effects on DOC production and carbon mineralization in agricultural peat soils. Soil Biology and Biochemistry, 38, 477–488.

Ciampalini, R., Martin, M.P., Saby, N.P.A., Richer-de-Forges, A.C., Arrouays, D., Nehlig, P., Martelet, G., 2014. Modelling soil particle-size distribution in the region “Centre” (France). In: Arrouays, D., McKenzie, N., Hempel, J., Richer-de-Forges, A.C., McBratney, A. (Eds.), Globalsoilmap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 121–126.

Clifford, D., Guo, Y., 2015. Combining two soil property rasters using an adaptive gating approach. Soil Research,53, 907–912.

Clothier, B.E., Hall, A.J., Deurer, M., Green, S.R., Mackay, A.D., 2011. Soil ecosystem services: sustaining returns on investment into natural capital. In: Sauer, T.J., Norman, J.M., Sivakumar, M.V.K. (Eds.), Sustaining Soil Productivity in Response to Global Climate Change: Science, Policy, and Ethics. Wiley-Blackwell, Oxford, pp. 117–139.

Cressie, N., 1993. Statistics for spatial data. John Wiley, New York, pp. 29–104.

Dam, R.F., Mehdi, B.B., Burgess, Madramootoo, C.A., Mehuys, G.R., Callum, I.R., 2005. Soil bulk density and crop yield under eleven consecutive years of corn with different tillage and residue practices in a sandy loam soil in central Canada. Soil and Tillage Research 84, 41–53.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F., Knotters, M., 2006. Sampling for natural resource monitoring. Springer, Berlin.

De Gruijter, J.J., Minasny, B., McBratney, A.B., 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. Journal of Survey Statistics and Methodology, 3(1), 19–42.

De Oliveira, V., 2005. Bayesian inference and prediction of Gaussian random fields based on censored data. Journal of Computational and Graphical Statistics, 14(1), 95–115.

De Souza, E., Filho, E., Schaefer, C., Batjes, N., dos Santos, G., Pontes, L., 2016. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Scientia Agricola, 73, 525–534.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Science Society of America Journal, 69, 500–510.

Del Grosso, S., Parton, W., Stohlgren, T., Zheng, D., Bachelet, D., Prince, S., ..., Olson, R., 2008. Global potential net primary production predicted from vegetation class, precipitation, and temperature. Ecology, 89, 2117–2126.

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D., ..., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. Science, 359, 320–325.

Dietrich, W.E., Reiss, R., Hsu, M.L., Montgomery, D.R., 1995. A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. Hydrological Processes, 9(3–4), 383–400.

Dignac, M.F., Derrien, D., Barré, P., Barot, S., Cécillon, L., Chenu, C., ..., Basile-Doelsch, I., 2017. Increasing soil carbon storage: Mechanisms, effects of agricultural practices and proxies: a review. Agronomy for Sustainable Development, 37, 14.

Dokuchaev, V. V. 1883. Russian chernozem. In: Selected Works of V.V. Dokuchaev, Volume 1. Moscow, 1948 (ed. S. Monson), pp. 14–419. (Transl. into English by N.Kaner). Israel Program for Scientific Translations Ltd. (for USDA-NSF), Jerusalem, Israel.

Don, A., Schumacher, J., Freibauer, A., 2011. Impact of tropical land-use change on soil organic carbon stocks–a meta-analysis. Global Change Biology, 17, 1658–1670.

Elbehri, A., 2015. Climate change and food systems: global assessments and implications for food security and trade. Food and Agriculture Organization of the United Nations (FAO).

Elzhov, T.V., Mullen, K.M., Spiess, A.-N., Bolker, B., 2013. minpack.lm: R Interface to the Levenberg–Marquardt Nonlinear Least-squares Algorithm Found in MINPACK, Plus Support for Bounds. R Package Version 1.2-1. https://cran.r-project.org/web/packages/minpack.lm/index.html.

FAO, 2011. FAO in the 21st century.

Faroux, S., Tchuenté, A.K., Roujean, J.L., Masson, V., Martin, E., Le Moigne, P., 2013. ECOCLIMAP-II/Europe: A twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models. Geoscientific Model Development, 6(2), 563–582.

Federer, 1983. Nitrogen mineralization and nitrification: depth variation in four New England forest soils. Soil Science Society of America Journal 47, 1008–1014.

Feranec, J., Jaffrain, G., Soukup, T., Hazeu, G., 2010. Determining changes and flows in European landscapes 1990–2000 using Corine land cover data. Applied Geography, 30(1), 19–35.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37(12), 4302-4315.

Filzmoser, P., Hron, K., 2008. Dissimilar samples detection for compositional data using robust

methods. Mathematical Geosciences, 40, 233–248.

Finke, P.A., 2012. On digital soil assessment with models and the Pedometrics agenda. Geoderma, 171, 14-30.

Fridley, B.L., Dixon, P., 2007. Data augmentation for a Bayesian spatial model involving censored observations. Environmetrics, 18(2), 107–123.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29, 1189–1232.

Friedman, J.H., Meulman J.J., 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine, 22: 1365–1381.

Ge, Y., Avitabile, V., Heuvelink, G.B.M., Wang, J., Herold, M., 2014. Fusion of pan-tropical biomass maps using weighted averaging and regional calibration data. International Journal of Applied Earth Observation and Geoinformation, 31, 13–24.

Gelaw, A.M., Singh, B., Lal, R., 2015. Organic carbon and nitrogen associated with soil aggregates and particle sizes under different land uses in Tigray, northern Ethiopia. Land Degradation and Development, 26, 690–700.

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

Granger, C.W., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting, 3, 197–204.

Gray, J.M., Bishop, T.F., 2016. Change in soil organic carbon stocks under 12 climate change projections over New South Wales, Australia. Soil Science Society of America Journal, 80(5), 1296-1307.

Groshans, G.R., Mikhailova, E.A., Post, C.J., Schlautman, M.A., 2018. Accounting for soil inorganic carbon in the ecosystem services framework for United Nations sustainable development goals. Geoderma, 324, 37–46.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma, 152(3-4), 195-207.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. Soil Science Society of America Journal 75, 1201–1213.

Haddaway, N.R., Hedlund, K., Jackson, L.E., Kätterer, T., Lugato, E., Thomsen, I.K., ... Isberg, P.E., 2016. How does tillage intensity affect soil organic carbon? A systematic review protocol. Environmental Evidence, 5, 1.

Hargreaves, G.H., Allen, R.G., 2003. History and evaluation of Hargreaves evapotranspiration equation. Journal of Irrigation and Drainage Engineering, 129, 53–63.

Hargreaves, G.L., Hargreaves, G.H., Riley, J.P., 1985. Irrigation water requirements for Senegal River Basin. Journal of Irrigation and Drainage Engineering, 111, 265–275.

Harrell Jr, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the yield of medical tests. Jama, 247(18), 2543–2546.

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma, 207, 256-267.

Hassink, J., 1997. The capacity of soils to preserve organic C and N by their association with clay and silt particles. Plant and Soil, 191(1), 77–87.

Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1997. The soil production function

and landscape equilibrium. Nature, 388, 358–361.

Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1999. Cosmogenic nuclides, topography, and the spatial variation of soil depth. Geomorphology, 27(1-2), 151–172.

Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 2001. Stochastic processes of soil production and transport: Erosion rates, topographic variation and cosmogenic nuclides in the Oregon Coast Range. Earth Surface Processes and Landforms 26(5), 531–552.

Heinze, S., Ludwig, B., Piepho, H-P., Mikutta, R., Don, A., Wordell-Dietrich, P., ..., Marschner, B., 2018. Factors controlling the variability of organic matter in the top- and subsoil of a sandy Dystric Cambisol under beech forest. Geoderma, 311, 37–44.

Heitkötter, J., Niebuhr, J., Heinze, S., Marschner, B., 2017. Patterns of nitrogen and citric acid induced changes in C-turnover and enzyme activities are different in topsoil and subsoils of a sandy Cambisol. Geoderma, 292, 111–117.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 122, e0169748.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., ..., Gonzalez, M.R., 2014. SoilGrids1km—global soil information based on automated mapping. PLoS One, 9(8), e105992.

Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1–2), 75–93.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One, 10, e0125814.

Heuscher, S.A., Brandt, C.C., Jardine, P.M., 2005. Using soil physical and chemical properties to estimate bulk density. Soil Science Society of America Journal, 69, 51–56.

Heuvelink, G.B.M., 2019. The '10PM Challenges'. Pedometron, 43, 9-13.

Heuvelink, G.B.M., Brus, D.J., de Gruijter, J.J., 2006. Optimization of sample configurations for digital mapping of soil properties with universal kriging. Developments in Soil Science, 31, 137-151.

Heuvelink, G.B.M., 2018. Uncertainty and uncertainty propagation in soil mapping and modelling. Pedometrics, pp.439-461.

Heuvelink, G.B.M., Bierkens, M.F.P., 1992. Combining soil maps with interpolations from point observations to predict quantitative soil properties. Geoderma, 55, 1–15.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25, 1965–1978.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–401.

Info Terre – Site cartographique de référence sur les géosciences, 2014. Indice de développement et de persistance des réseaux (IDPR), edited, BRGM – Centre scientifique et technique,

Orléans, France.

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.

Inventaire Forestier National, 2006. BD Forêt® In: Service de l'inventaire forestier et statistique - Institut national de l'information géographique et forestière (IGN) (Ed.), Nogent-sur-Vernisson, 613 France.

Ishwaran H., Kogalur U.B., 2017. Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.5.1.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. The Annals of Applied Statistics, 2(3), 841–860.

IUCN, 2015. Land degradation neutrality: implications and opportunities for conservation.

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

Jalabert, S.S.M., Martin, M.P., Renaud, J.P., Boulonne, L., Jolivet, C., Montanarella, L., Arrouays, D., 2010. Estimating forest soil bulk density using boosted regression modelling. Soil Use and Management, 26, 516–528.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled srtmfor the globe version 4. available from the CGIAR-CSI SRTM 90m Database. (http://srtm.csi.cgiar.org).

Jeffrey, D.W., 1970. A Note on the use of Ignition Loss as a Means for the Approximate Estimation of Soil Bulk Density. The Journal of Ecology, 58, 297–299.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGrawHill, New York, pp.1–20.

Jolivet, C. Arrouays, D. Bernoux, M. 1998. Comparizon of soil organic carbon and organic matter determinations in sandy soils of France. Communications in Soil Science and Plant Analysis 29, 2227-2233.

Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N.P.A., 2006. Le réseau de mesures de la qualité des sols de France (RMQS). Etat d'avancement et premiers résultats. Etude et Gestion des Sols, 13, 149–164.

Jolivet, C., Arrouays, D., Leveque, J., Andreux, F., Chenu, C., 2003. Organic carbon dynamics in soil particle-size separates of sandy Spodosols when forest is cleared for maize cropping. European Journal of Soil Science, 54, 257–268.

Joly, D., Brossard, T., Cardot, H., Cavailhes, J., Hilal, M., Wavresky, P., 2010. Les types de climats en France, une construction spatiale. Cybergeo: European Journal of Geography.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., ..., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Scientific Data, 4, 170122.

Kaur, R., Kumar, S., Gurung, H.P., 2002. A pedo-transfer function (PTF) for estimating soil bulk density from basic soil data and its comparison with existing PTFs. Soil Research, 40, 847–857.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., 2016. The significance of soils and

soil science towards realization of the United Nations Sustainable Development Goals. SOIL, 2(2), 111–128.

Kell, D.B., 2012. Large-scale sequestration of atmospheric carbon via plant roots in natural and agricultural ecosystems: Why and how. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 1589–1597.

Keller, T., Håkansson, I., 2010. Estimation of reference bulk density from soil particle size distribution and soil organic matter content. Geoderma, 154, 398–406.

Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1: 50,000 soil map of the Netherlands. Geoderma, 241, 313–329.

Kempen, B., Dalsgaard, S., Kaaya, A.K., Chamuya, N., Ruipérez-González, M., Pekkarinen, A., Walsh, M.G., 2019. Mapping topsoil organic carbon concentrations and stocks for Tanzania. Geoderma, 337, 164-180.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma, 339, 40-58.

King, D., Jones, R., Thomasson, A., 1995. European Land Information Systems for Agro-Environmental Monitoring.

Knotters, M., Brus, D.J., Voshaar, J.O., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma, 67(3–4), 227–246.

Koch, A., McBratney, A., Lal, R., 2012. Global soil week: Put soil security on the global agenda. Nature, 492(7428), 186-186.

Koch, A., McBratney, A.B., Adams, M., Field, D., Hill, R., Crawford, J., Minasny, B., Lal, R., Abbott, L., O'Donnell, A., Angers, D., 2013. Soil security: solving the global soil crisis. Global Policy, 4, 434–441.

Koven, C.D., Lawrence, D.M., Riley, W.J., 2015. Permafrost carbon– climate feedback is sensitive to deep soil carbon decomposability but not deep soil nitrogen dynamics. Proceedings of the National Academy of Sciences, 112(12), 3752-3757.

Kuhn M., 2008. Caret package. Journal of Statistical Software, 28, 1–26.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2012. Cubist models for regression. R package Vignette R package version 0.0, 18.

Kuriakose, S.L., Devkota, S., Rossiter, D.G., Jetten, V.G., 2009. Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the Western Ghats of Kerala, India. Catena, 79(1), 27–38.

Laborczi, A., Szatmári, G., Kaposi, A.D., Pásztor, L., 2019. Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma, 352, 360-372.

Lacarce, E., Saby, N., Martin, M., Marchant, B., Boulonne, L., Meersmans, J., Jolivet, C., Bispo, A., Arrouays, D., 2012. Mapping soil Pb stocks and availability in mainland France combining regression trees with robust geostatistics. Geoderma, 170, 359–368.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology, 133(1-2), 90-99.

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.

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

Lagacherie, P., McBratney, A.B., 2006. Chapter 1 spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Developments in Soil Science, 3–22.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science, 304, 1623–1627.

Lal, R., 2018. Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. Global Change Biology, 24(8), 3285-3301.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science, 51, 137–157.

Laroche, B., Richer-de-Forges, A.C., Leménager, S., Arrouays, D., Schnebelen, N., Eimberck, M., Toutain, B., Lehmann, S., Tientcheu, E., Héliès, F., Chenu, J.-P., Parot, S., Desbourdes, S., Girot, G., Voltz, M., Bardy, M., 2014. Le programme Inventaire Gestion et Conservation des Sols. Volet Référentiel Régional Pédologique. Etude Gest Sols, 21, 125–140.

Leenaars, J.G., Claessens, L., Heuvelink, G.B., Hengl, T., González, M.R., van Bussel, L.G., Guilpart, N., Yang, H., 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.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news, 2, 18–22.

Libohova, Z., Wills, S., Odgers, N.P., Ferguson, R., Nesser, R., Thompson, J.A., Larry, T.W., Hempel, J.W., 2014. Converting pH 1:1 H2O and 1:2 CaCl2 to 1:5 H2O to contribute to a harmonized global soil database. Geoderma, 213, 544–550.

Lilly, A., Baggaley, N.J., 2013. The potential for Scottish cultivated topsoils to lose or gain soil organic carbon. Soil Use and Management, 29, 39–47.

Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Bourennane, H., Saby, N.P.A., Martin, M.P., Vaudour, E., Gomez, C., Lagacherie, P., 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.

Lorenz, K., Lal, R., 2005. The depth distribution of soil organic carbon in relation to land use and management and the potential of carbon sequestration in subsoil horizons. Advances in Agronomy, 88, 35–66.

Lugato, E., Bampa, F., Panagos, P., Montanarella, L., Jones, A., 2014. Potential carbon sequestration of European arable soils estimated by modelling a comprehensive set of management practices. Global Change Biology, 20(11), 3557–3567.

Luo, Y., Ahlström, A., Allison, S.D., Batjes, N.H., Brovkin, V., Carvalhais, N., ..., Georgiou, K., 2016. Toward more realistic projections of soil carbon dynamics by Earth system models. Global Biogeochemical Cycles, 30(1), 40-56.

Ma, Y., Minasny, B., Welivitiya, W.D.P., Malone, B.P., Willgoose, G.R., McBratney, A.B., 2019. The feasibility of predicting the spatial pattern of soil particle-size distribution using a pedogenesis model. Geoderma, 341, 195-205.

Malone, B., McBratney, A., Minasny, B., 2011. Empirical estimates of uncertainty for mapping

continuous depth functions of soil attributes. Geoderma, 160, 614–626.

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, 138–152.

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for digital soil mapping. Basel, Switzerland: Springer International Publishing.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, 232, 34–44.

Manrique, L.A., Jones, C.A., 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Science Society of America Journal, 55, 476–481.

Marchant, B.P., Lark, R.M., 2007. The Matérn variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys. Geoderma, 140(4), 337-345.

Marchant, B.P., Villanneau, E.J., Arrouays, D., Saby, N.P.A., Rawlins, B.G., 2015. Quantifying and mapping topsoil inorganic carbon concentrations and stocks: approaches tested in France. Soil Use and Management, 31, 29-38.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., Arrouays, D., 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Science Society of America Journal, 73, 485–493.

Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P.A., Paroissien, J.B., Jolivet, C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma, 223, 97–107.

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences, 8, 1053–1065.

Marx, A., Erhard, M., Thober, S., Kumar, R., Schäfer, D., Samaniego, L., Zink, M., 2019. Climate Change as Driver for Ecosystem Services Risk and Opportunities. In Atlas of Ecosystem Services. Springer, Cham, pp. 173–178.

Matheron, G., 1971. La théorie des variables régionalisées et ses applications. Vol. 5. Les cahiers du centre de morphologie mathématique. Fontainebleau.

Mathieu, J.A., Hatté, C., Balesdent, J., Parent, É., 2015. Deep soil carbon dynamics are driven more by soil type than by climate: A worldwide meta-analysis of radiocarbon profiles. Global Change Biology, 21, 4278–4292.

May, M., Royston, P., Egger, M., Justice, A.C., Sterne, J.A., 2004. Development and validation of a prognostic model for survival time data: application to prognosis of HIV positive patients treated with antiretroviral therapy. Statistics in Medicine, 23(15), 2375–2398.

McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma, 213, 203-213.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma, 109(1-2), 41-73.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1–2), 3–52.

McNally, S.R., Beare, M.H., Curtin, D., Meenken, E.D., Kelliher, F., Calvelo Pereira, R., ..., Baldock, J., 2017. Soil carbon sequestration potential of permanent pasture and continuous cropping soils in New Zealand. Global Change Biology, 23, 4544–4555.

Meersmans, J., Arrouays, D., Van Rompaey, A.J., Pagé, C., De Baets, S., Quine, T.A., 2016. Future C loss in mid-latitude mineral soils: Climate change exceeds land use mitigation potential in

France. Scientific Reports, 6, 35798.

Meersmans, J., Martin, M.P., De Ridder, F., Lacarce, E., Wetterlind, J., De Baets, S., … Arrouays, D., 2012. A novel soil organic C model using climate, soil type and management data at the national scale in France. Agronomy for Sustainable Development, 32, 873–888.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S.C.B., Frieler, K., Knutti, R., … Allen, M.R., 2009. Greenhouse-gas emission targets for limiting global warming to 2°C. Nature, 458, 1158–1162.

Meyer, M.D., North, M.P., Gray, A.N., Zald, H.S., 2007. Influence of soil thickness on stand characteristics in a Sierra Nevada mixed-conifer forest. Plant and Soil, 294(1–2), 113–123.

Millennium Ecosystems Assessment, 2005. Ecosystems and human well-being: policy responses. Island Press, Washington, DC.

Minasny B., McBratney A.B., Bristow K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma, 93, 225-253.

Minasny, B., Arrouays, D., McBratney, A.B., Angers, D.A., Chambers, A., Chaplot, V., …, Gimona, A., 2018. Rejoinder to Comments on Minasny et al., 2017 Soil carbon 4 per mille Geoderma 292, 59–86. Geoderma, 309, 124–129.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., …, Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma, 292, 59–86.

Minasny, B., McBrateney, A.B., 2013. Why you don't need to use RPD. Pedometron, 33, 13.

Minasny, B., McBratney, A.B., 1999. A rudimentary mechanistic model for soil production and landscape development. Geoderma, 90(1–2), 3–21.

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

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

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance and the World Reference Base. Geoderma, 155, 132–139.

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Advances in Agronomy, 118, 1-47.

Minasny, B., McBratney, A.B., Mendonça-Santos, M., Odeh, I., Guyon. B., 2006. Prediction and digital mapping of soil carbon storage in the lower Namoi valley. Soil Research, 44, 233–244.

Mogensen, U.B., Ishwaran, H., Gerds, T.A., 2012. Evaluating random forests for survival analysis using prediction error curves. Journal of Statistical Software, 50(11), 1–23.

Monlong, J., 2018. Hippocamplus, Github repository, https://github.com/jmonlong/Hippocamplus/blob/master/content/post/2018-06-09-ClusterEqualSize.Rmd

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, 57(2), 443–452.

Mouselimis, L., 2016. Clustering using the ClusterR package. http://mlampros.github.io/2016/09/12/clusterR_package/

Mulder, V.L., de Bruin, S., Weyermann, J., Kokaly, R.F., Schaepman, M.E., 2013. Characterizing regional soil mineral composition using spectroscopy and geostatistics. Remote Sensing of Environment, 139, 415-429.

Mulder, V.L., Lacoste, M., Martin, M.P., Richer-de-Forges, A., Arrouays, D., 2015. Understanding large-extent controls of soil organic carbon storage in relation to soil depth and soil-landscape systems. Global Biogeochemical Cycles, 29, 1210–1229.

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

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016b. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma, 263, 16–34.

Nanko, K., Ugawa, S., Hashimoto, S., Imaya, A., Kobayashi, M., Sakai, H., Ishizuka, S., Miura, S., Tanaka, N., Takahashi, M., Kaneko, S., 2014. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. Geoderma, 213, 36–45.

NASA LD, 2001. NASA Land Processes Distributed Active Archive Center (LP DAAC) USGS/Earth Resources Observation and Science (EROS) Center.

NASA LP DAAC, 2017. MOD17A2H: MODIS/TERRA Gross Primary Production. Version 6. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov), accessed October 30, 2017, at http://dx.doi.org/ 10.5067/MODIS/MOD17A2H.006.

Nemani, R.R., Keeling, C.D., Hashimoto, H., Jolly, W.M., Piper, S.C., Tucker, C.J., ... Running, S.W., 2003. Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science, 300, 1560–1563.

Nemes, A., Quebedeaux, B., Timlin, D.J., 2010. Ensemble Approach to Provide Uncertainty Estimates of Soil Bulk Density. Soil Science Society of America Journal, 74, 1938–1945.

O'Rourke, S.M., Angers, D.A., Holden, N.M., McBratney, A.B., 2015. Soil organic carbon across scales. Global Change Biology, 21, 3561–3574.

Odeh, I.O., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma, 67(3–4), 215–226.

Odeh, I.O.A., Chittleborough, D.J., McBratney, A.B., 1991. Elucidation of soil-landform interrelationships by canonical ordination analysis. Geoderma, 49(1–2), 1–32.

Omuto, C.T., Vargas, R.R., 2015. Re-tooling of regression kriging in R for improved digital mapping of soil properties. Geosciences Journal, 19(1), 157-165.

Orton, T.G., Rawlins, B.G., Lark, R.M., 2009. Using measurements close to a detection limit in a geostatistical case study to predict selenium concentration in topsoil. Geoderma, 152(3–4), 269–282.

Orton, T.G., Saby, N., Arrouays, D., Jolivet, C.C., Villanneau, E.J., Paroissien, J.B., Marchant, B.P., Caria, G., Barriuso, E., Bispo, A., Briand, O., 2012. Analyzing the spatial distribution of PCB concentrations in soils Using below–quantification limit data. Journal of Environmental Quality, 41(6), 1893–1905.

Padarian, J., McBratney, A.B., 2019. A new model for intra- and inter-institutional soil data sharing. SOIL, under review. https://doi.org/10.5194/soil-2019-65

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Regional, 9, 17–28.

Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning for digital soil mapping. SOIL, 5(1), 79-89.

Padarian, J., Minasny, B., McBratney, A.B., Dalgliesh, N., 2014. Predicting and mapping the soil available water capacity of Australian wheatbelt. Geoderma Regional, 2, 110–118.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G.P., Smith, P., 2016. Climate-smart soils. Nature, 532, 49.

Pebesma, E., Graeler, B., 2013. The gstat package: spatial and spatio-temporal geostatistical modelling, prediction and simulation. R Package Version, 1-0, R Fondation for Statistical Computing, Vienna, Austria.

Pellerin, S., Bamière, L., Launay, C., Martin, R., Schiavo, M., Angers, D., ..., Cardinael, R., 2019. Stocker du carbone dans les sols français, quel potentiel au regard de l'objectif 4 pour 1000 et à quel coût?. (In French)

Pelletier, J.D., Rasmussen, C., 2009. Geomorphically based predictive mapping of soil thickness in upland watersheds. Water Resources Research, 45(9), 417.

Penížek, V., Borůvka, L., 2006. Soil depth prediction supported by primary terrain attributes: a comparison of methods. Plant, Soil and Environment, 52(9), 424–430.

Poeplau, C., Vos, C., Don, A., 2017. Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. SOIL, 3, 61–66.

Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation—an example from Scotland. Geoderma, 232, 284–299.

Poggio, L., Lassauce, A., Gimona, A., 2019. Modelling the extent of northern peat soil and its uncertainty with Sentinel: Scotland as example of highly cloudy region. Geoderma, 346, 63–74.

Post, W.M., Kwon, K.C., 2000. Soil carbon sequestration and land-use change: processes and potential. Global Change Biology, 6(3), 317–327.

Powlson, D., Bhogal, A., Chambers, B., Coleman, K., Macdonald, A., Goulding, K., Whitmore, A., 2012. The potential to increase soil carbon stocks through reduced tillage or organic material additions in England and Wales: A case study. Agriculture, Ecosystems and Environment, 146, 23–33.

Quinlan, J.R., 1992. Learning with continuous classes. In 5th Australian Joint Conference on Artificial Intelligence, 92, 343–348.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabot, E., Wiesmeier, M., Schlüter, S., Vogel, H.J., 2018. Soil structure as an indicator of soil functions: a review. Geoderma, 314, 122–137.

Rad, M.R.P., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma, 232, 97-106.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133, 1155–1174.

Reis, C.E.S.D., Dick, D.P., Caldas, J.D.S., Bayer, C., 2014. Carbon sequestration in clay and silt fractions of Brazilian soils under conventional and no-tillage systems. Scientia Agricola, 71, 292–301.

Reyes Rojas, L.A., Adhikari, K., Ventura, S.J., 2018. Projecting soil organic carbon distribution in central Chile under future climate scenarios. Journal of Environmental Quality, 47(4), 735-745.

Rial, M., Cortizas, A. M., Rodríguez-Lado, L., 2016. Mapping soil organic carbon content using spectroscopic and environmental data: A case study in acidic soils from NW Spain. Science of the Total Environment, 539, 26-35.

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

Ridgeway, G., 2012. Generalized Boosted Models: A guide to the gbm package. R package vignette.

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., Saby, N.P.A, 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma, 298, 67–77.

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

Román Dobarco, M., Cousin, I., Le Bas, C., Martin, M.P., 2019b. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma, 336, 81-95.

Román, J.R., Roncero-Ramos, B., Chamizo, S., Rodríguez-Caballero, E., Cantón, Y., 2018. Restoring soil functions by means of cyanobacteria inoculation: importance of soil conditions and species selection. Land Degradation & Development, 29(9), 3184–3193.

Rousseeuw, P.J., Zomeren, B.C., 1990. Unmasking multivariate dissimilar samples and leverage points. Journal of the American Statistical Association, 85, 633–639.

Rumpel, C., Amiraslani, F., Koutika, L., Smith, P., Whitehead, D., Wollenberg, D., 2018. Put more carbon in soils to meet Paris climate pledges. Nature, 564, 32–34.

Rumpel, C., Eusterhues, K., Kögel-Knabner, I., 2004. Location and chemical composition of stabilized organic carbon in topsoil and subsoil horizons of two acid forest soils. Soil Biology and Biochemistry, 36(1), 177-190.

Saby, N.P.A., Arrouays, D., Antoni, V., Foucaud-lemercier, B., Follain, S., Walter, C., Schvartz, C., 2008. Changes in soil organic carbon content in a French mountainous region, 1990-2004. Soil Use and Management, 24, 254–262.

Saby, N.P.A., Arrouays, D., Boulonne, L., Jolivet, C., Pochot, A., 2006. Geostatistical assessment of Pb in soil around Paris, France. Science of the Total Environment, 367, 212–221.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., de Lourdes Mendonça-Santos, M., Minasny, B., 2009. Digital soil map of the world. Science, 325, 680–681.

Sanderman, J., Berhe, A.A., 2017. Biogeochemistry: The soil carbon erosion paradox. Nature Climate Change, 7(5), 317.

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M. F., Benson, L., Bukoski, J.J., Carnell, P., Cifuentes-Jara, M., Donato, D., Duncan, C., Eid, E.M., zu Ermgassen, P., Lewis, C.J.E., Macreadie, P.I., Glass, L., Gress, S., Jardine, S.L., Jones, T.G., Nsombo, E.N., Rahman, M.M., Sanders, C.J., Spalding, M., Landis, E., 2018. A global map of mangrove forest soil carbon at

30 m spatial resolution. Environmental Research Letters, 13, 055002.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappie, M., Märker, M., Saia, S., 2017. Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. Science of the Total Environment, 601, 821-832.

Sequeira, C., Wills, S., Seybold, C., West, L., 2014. Predicting soil bulk density for incomplete databases. Geoderma, 213, 64–73.

Shangguan, W., Hengl, T., Mendes de Jesus, J., 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.

Six, J., Conant, R., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms of soil organic matter: Implications for c-saturation of soils. Plant and Soil, 241, 155–176.

Six, J., Elliott, E., Paustian, K., 2000. Soil macroaggregate turnover and microaggregate formation: A mechanism for c sequestration under no-tillage agriculture. Soil Biology and Biochemistry, 32, 2099–2103.

Smith, P., 2005. An overview of the permanence of soil organic carbon stocks: Influence of direct human-induced, indirect and natural effects. European Journal of Soil Science, 56, 673–680.

Soil Survey Division Staff, 1993. Soil survey manual. Soil conservation service U.S. Department of Agriculture Handbook, pp.18.

Song, X.D., Yang, F., Ju, B., Li, D.C., Zhao, Y.G., Yang, J.L., Zhang, G.L., 2018. The influence of the conversion of grassland to cropland on changes in soil organic carbon and total nitrogen stocks in the Songnen Plain of Northeast China. Catena, 171, 588-601.

Soussana, J.F., Lutfalla, S., Ehrhardt, F., Rosenstock, T., Lamanna, C., Havlik, P., ..., Lal, R., 2019. Matching policy and science: Rationale for the '4 per 1000 - soils for food security and climate' initiative. Soil & Tillage Research, 188, 3–15.

Soussana, J.F., Saint-Macary, H., Chotte, J.L., 2015. Carbon sequestration in soils: the 4 per mil concept. Agriculture and agricultural soils facing climate change and food security challenges: public policies and practices conference. Paris, Sept. 16, 2015. http://www.ag4climate.org/programme/ag4climate-session-2-3-soussana.pdf

Sparling, G., Parfitt, R.L., Hewitt, A.E., Schipper, L.A., 2003. Three approaches to define desired soil organic matter content. Journal of Environmental Quality, 32, 760-766.

Stewart, C.E., Paustian, K., Conant, R.T., Plante, A.F., Six, J., 2007. Soil carbon saturation: Concept, evidence and evaluation. Biogeochemistry, 86, 19–31.

Stockmann, U., Adams, M., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., ..., Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agriculture, Ecosystems & Environment, 164, 80–99.

Stockmann, U., Padarian, J., McBratney, A., Minasny, B., de Brogniez, D., Montanarella, L., ... & Field, D.J., 2015. Global soil organic carbon assessment. Global Food Security, 6, 9-16.

Stolbovoy, V., Montanarella, L., 2008. Application of soil organic carbon status indicators for policy-decision making in the EU, In: Toth, G., Montanarella, L., Rusco, E. (Eds.), Threats to soil quality in Europe, 87–99.

Styc, Q., Lagacherie, P., 2016. Predicting soil depth using a survival analysis model. Oral presentation at the 7th Global Workshop on Digital Soil Mapping, Aarhus, Denmark, 27 June

to 1 July.

Sun, X.L., Zhao, Y.G., Wu, Y.J., Zhao, M.S., Wang, H.L., Zhang, G.L., 2012. Spatio-temporal change of soil organic matter content of Jiangsu Province, China, based on digital soil maps. Soil Use and Management, 28(3), 318-328.

Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X., Zhang, X., 2018. Non-stationary bias correction of monthly CMIP5 temperature projections over China using a residual-based bagging tree model. International Journal of Climatology, 38, 467–482.

Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil depth from topographic and land cover attributes. Water Resources Research, 45(10), 438.

Thomas, M., Clifford, D., Bartley, R., Philip, S., Brough, D., Gregory, L., ..., Glover, M., 2015. Putting regional digital soil mapping into practice in Tropical Northern Australia. Geoderma, 241, 145-157.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Science, 163, 190–202.

Torn, M.S., Trumbore, S.E., Chadwick, O.A., Vitousek, P.M., Hendricks, D.M., 1997. Mineral control of soil organic carbon storage and turnover. Nature, 389, 170.

Tóth, G., Jones, A. Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental Monitoring and Assessment, 185, 7409–7425.

Tranter, G., McBratney, A.B., Minasny, B., 2009. Using distance metrics to determine the appropriate domain of pedotransfer function predictions. Geoderma, 149, 421–425.

Tranter, G., Minasny, B., McBratney, A., Murphy, B., Mckenzie, N., Grundy, M., Brough, D., 2007. Building and testing conceptual and empirical models for predicting soil bulk density. Soil Use and Management, 23, 437–443.

UE-SOeS, 2006. Corine Land Cover. Service de l'Observation et des Statistiques (SOeS) du Ministère de l'Environnement. de l'Énergie et de la Mer. Tech. rep.

van Groenigen, J.W., Van Kessel, C., Hungate, B.A., Oenema, O., Powlson, D.S., Van Groenigen, K.J., 2017. Sequestering soil organic carbon: a nitrogen dilemma. Environmental Science and Technology, 51, 4738–4739.

van Wesemael, B., Paustian, K., Andrén, O., Cerri, C.E., Dodd, M., Etchevers, J., ..., Ogle, S., 2011. How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO_2 fluxes in agricultural soils?. Plant and Soil, 338(1-2), 247-259.

Vanwalleghem, T., Poesen, J., McBratney, A., Deckers, J., 2010. Spatial variability of soil horizon depth in natural loess-derived soils. Geoderma, 157(1–2), 37–45.

Vaysse, K., Heuvelink, G.B., Lagacherie, P., 2017. Spatial aggregation of soil property predictions in support of local land management. Soil Use and Management, 33(2), 299-310.

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.

Villanneau, E.J., Saby, N.P., Marchant, B.P., Jolivet, C.C., Boulonne, L., Caria, G., ..., Arrouays, D., 2011. Which persistent organic pollutants can we map in soil using a large spacing systematic soil monitoring design? A case study in Northern France. Science of the Total Environment, 409(19), 3719-3731.

Viscarra Rossel, R.A, Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nature Geoscience, 12, 547-552.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158, 46–54.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., ..., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Reviews, 155, 198-230.

Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Research, 53(8), 845-864.

Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A, 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology, 20, 2953–2970.

Vogel, H.J., Bartke, S., Daedlow, K., Helming, K., Kögel-Knabner, I., Lang, B., Rabot, E., Russell, D., Stößel, B., Weller, U., Wiesmeier, M., 2018. A systemic approach for modeling soil functions. SOIL, 4(1), 83–92.

Von Steiger, B., Webster, R., Schulin, R., Lehmann, R., 1996. Mapping heavy metals in polluted soil by disjunctive kriging. Environmental Pollution, 94(2), 205–215.

Wadoux, A.M.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma, 351, 59-70.

Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B., 2019a. Sampling design optimization for soil mapping with random forest. Geoderma, 355, 113913.

Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma, 324, 138–147.

Wadoux, A.M.C., Marchant, B.P., Lark, R.M., 2019b. Efficient sampling for geostatistical surveys. European Journal of Soil Science, 70, 975-989.

Wadoux, A.M.C., Padarian, J., Minasny, B., 2019c. Multi-source data integration for soil mapping using deep learning. SOIL, 5(1), 107-119.

Walter, C., Lagacherie, P., Follain, S., 2006. Integrating pedological knowledge into digital soil mapping. Developments in Soil Science, 31, 281-615.

Walvoort, D.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 & Geosciences, 36(10), 1261-1267.

Wang, J., Endreny, T.A., Hassett, J.M., 2006. Power function decay of hydraulic conductivity for a TOPMODEL-based infiltration routine. Hydrological Processes, 20(18), 3825–3834.

Wiesmeier, M., Barthold, F., Spörlein, P., Geuß, U., Hangen, E., Reischl, A., ..., Kögel-Knabner, I., 2014a. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). Geoderma Regional, 1, 67–78.

Wiesmeier, M., Hübner, R., Spörlein, P., Hangen, E., Reischl, A., Schilling, B., ..., Kögel-Knabner, I., 2014b. Carbon sequestration potential of soils in southeast Germany derived from stable soil organic carbon saturation. Global Change Biology, 20, 653–665.

Wiesmeier, M., Schad, P., von Lützow, M., Poeplau, C., Spörlein, P., Geuß, U., ..., Kögel-Knabner,

I., 2014c. Quantification of functional soil organic carbon pools for major soil units and land uses in southeast Germany (Bavaria). Agriculture, Ecosystems & Environment, 185, 208–220.

Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., ..., Kögel-Knabner, I., 2012. Soil organic carbon stocks in southeast Germany (Bavaria) as affected by land use, soil type and sampling depth. Global Change Biology, 18, 2233–2245.

World Bank, World Development Indicators., 2018. CO2 emissions (kt). Retrieved from https://data.worldbank.org/indicator/EN.ATM.CO2E.KT?end=2014&locations=FR&start=1960&view=chart

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., ..., Bates, P.D., 2017. A high-accuracy map of global terrain elevations. Geophysical Research Letters, 44(11), 5844-5853.

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Science of the Total Environment, 557, 838-850.

Zhang, G.L., Liu, F., Song, X.D., 2017. Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture, 16(12), 2871-2885.

Zhang, T., Frauenfeld, O.W., Serreze, M.C., Etringer, A., Oelke, C., McCreight, J., Barry, R.G., Gilichinsky, D., Yang, D., Ye, H., Ling, F., 2005. Spatial and temporal variability in active layer thickness over the Russian Arctic drainage basin. Journal of Geophysical Research: Atmospheres 110, D16101.

Zhou, Y., Hartemink, A.E., Shi, Z., Liang, Z., Lu,Y., 2019. Land use and climate change effects on soil organic carbon in North and Northeast China. Science of the Total Environment, 647, 1230-1238.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal, 65(5), 1463–1472.

Zomer, R.J., Bossio, D.A., Sommer, R., Verchot, L.V., 2017. Global Sequestration Potential of Increased Organic Carbon in Cropland Soils. Scientific Reports, 7, 15554.

Acknowledgements

I was born in the countryside of southern China and had been live there in the first 13 years. Due to the lack of technical and financial support, the farmers from my hometown managed their lands only based on their experiences at that time (the 1990s). To keep or increase the crop yields, a large number of chemical fertilizers and pesticides had been applied in farmland, which later induced land degradation and water pollution (i.e., eutrophication). I can still remember that it was common practice to burn straw after the harvest to clean the land, which definitely removed major input of organic matter and also led to nutrient loss and atmospheric pollution. My childhood memory motivated me to obtain a bachelor's degree and master's degree in Agriculture at Zhejiang University, from which I thought I can make a substantial contribution to this top agricultural country. I have witnessed the great progress on three issues of agriculture, the countryside and farmers in China during the last decade, but it remains many challenges related to precision and sustainable agriculture. Owning to the utmost importance of soil quality and health in agriculture, up-to-date and high-resolution soil information are urgently needed to support the decision making in land management. This knowledge gap further motivated me to learn more from other developed countries as a PhD student. Now, three years have passed since I stepped in France and I almost touch the finishing line of my PhD journey. Here, I would like to acknowledge all the help and support that I have received during the last three years.

First of all, I highly appreciate the financial support from China Scholarship Council that allows me to fulfil my scientific adventure in France. This funding is gradually boosting the scientific breakthroughs and advances, and making China and Earth better and better.

I would like to express sincere gratitude to my supervisors Dr Dominique

Arrouays and Prof. Christian Walter for their kind guidance, strong support, great patience, high motivation, and immense knowledge on soil science. I spent the most time in Orléans with Dr Dominique Arrouays, his deep love in soil science along with diligence, optimism, kindness and French humour, made my PhD journal fruitful and enjoyable. Though Prof. Christian Walter is far away from me in Rennes, he is always ready to help me and taking good care of my PhD process through email exchanges and videoconferences. His broad knowledge of different sub-disciplines in soil science deeply widens my perspectives during my PhD journey.

Besides my supervisors, I would like to thank all the members in my PhD committee: Prof. Philippe Lagacherie, Dr Nicolas Saby, Dr Blandine Lemercier, and Dr Jeroen Meersmans. Their insightful comments and suggestions greatly improved my researches, and encouraged me to widen my horizons on other disciplines.

My sincere thanks go to Dr Titia Mulder, Prof. Gerard Heuvelink, and Dr Laura Poggio, for their warm host in Wageningen University and ISRIC for 40 days, and also for their great help and encouragement during the collaboration. I also would like to thank Dr Denis Angers, Prof. Claire Chenu, and Dr Pierre Barré, for their significant contribution in my publications.

I am deeply grateful to two directors of INRA Unité InfoSol, Dr Marion Bardy and Dr Antonio Bispo, for supporting and encouraging me to attend various training courses and international conferences. I also owe a sincere thank to my INRA colleagues, Dr Manuel Martin, Richer-de-Forges Anne, Dr Mercedes Román Dobarco, Eugénie Tientcheu, Stephanie Guerrier, Sébastien Lehmann, Christine Le Bas, Dr Bassem Dimassi, and others are not mentioned here, for their great help both in scientific research and daily life.

I would like to extend thanks to Renlei Wang for his great help during our stay at Orléans in the last three years. Special thanks are reversed for Bifeng Hu for his accompany, help and encouragement in both study and daily life. I also

would like to thank my previous supervisor Prof. Zhou Shi and colleagues in his team, for these communications and collaboration that enhanced my knowledge related to my thesis.

Last but most important, I owe a huge thank to my wife Wan Jin and my little girl Xilan Chen. Their sweet love and accompany, immense support and encouragement, great patience and comprehension, help me enjoy this PhD journey. I highly appreciate all the supports from my parents Xiuyun Liang and Qiufan Chen, without whom I cannot go this far.

Résumé étendu en français (20 pages)

Introduction générale

Les sols jouent un rôle central dans de nombreux grands enjeux planétaires. Que ce soit pour la sécurité alimentaire, la sécurité de l'approvisionnement en eau, le changement climatique, la production d'énergie renouvelable, la santé humaine ou la protection de la biodiversité, les écosystèmes mondiaux dépendent de nombreux processus biogéochimiques intervenants dans les sols (Koch et al., 2012; Weil and Brady, 2016). Ces enjeux globaux ont fait émerger le concept de 'sécurité des sols' (Koch et al, 2013; McBratney et al., 2014) qui souligne la nécessité de préserver et d'améliorer les ressources en sols du monde pour faire face à ces défis majeurs. Les sols sont également des maillons essentiels pour atteindre les objectifs de développement durables des nations Unies comme les objectifs 2 (faim « zéro »), 3 (bonne santé et bien-être), 6 (eau propre et assainissement), 7 (énergie propre et d'un coût abordable), 12 (consommation et production responsables), 13 (mesures relatives à la lutte contre les changements climatiques), 14 (vie aquatique) et 15 (vie terrestre), (Bouma et al., 2014; Keesstra et al., 2016).

Bien que les sols soient au centre de ces enjeux globaux, leur gestion nécessite des actions et des connaissances locales. C'est la raison pour laquelle il existe une forte demande pour des informations sur les sols, à la fois aux échelles globales et locales. Ce constat est à l'origine du développement de prédictions globales à haute résolution spatiale de propriétés des sols (Arrouays et al., 2014). Au niveau mondial, environ deux pays sur trois disposent de cartes de sols conventionnelles à des échelles supérieures ou égales au 1/1 000 000, mais plus des deux tiers de la surface de terres émergées ne disposent pas de telles cartes, même à l'échelle du 1/1 000 000 (Hartemink et al., 2013). Les cartes conventionnelles sont le plus souvent obsolètes, et leurs réalisations ou actualisations sont coûteuses et nécessitent un temps de travail considérable (Grunwald et al., 2011). De plus, elles ne fournissent généralement pas d'indications sur leurs incertitudes et reposent essentiellement sur l'expertise des pédologues, ce qui rend difficile leur reproductibilité et leur mise à jour. Ces limites des cartes de sols conventionnelles sont à l'origine de l'émergence et du développement de la cartographie numérique des sols (CNS, en anglais : Digital Soil Mapping : DSM) facilité par le formidable développement des technologies de l'information géographique et de son traitement ainsi que par l'augmentation de la puissance de calcul (Minasny and McBratney, 2016; Zhang et al., 2017).

La CNS a été définie comme étant « La création et l'enrichissement des

systèmes d'information spatiale sur les sols par des modèles numériques permettant d'inférer les variations spatiales et temporelles des types de sols ou de leurs propriétés à partir d'observations et de connaissances sur les sols en utilisant des variables environnementales qui y sont liées » (traduction d'une définition en anglais de Lagacherie and McBratney, 2006). Ce concept est fondé sur la théorie des facteurs de la formation des sols de Dokuchaev (1883), puis de Jenny (1941). Il a été complété et formalisé ensuite par McBratney et al. (2003) sous le nom de « *scorpan* model », $S=f(s, c, o, r, p, a, n)$. Bien que de nombreux travaux de CNS aient été effectués bien avant 2003, le modèle *scorpan* model fut le premier à formaliser et conceptualiser la CNS comme une prédiction spatiale quantitative. Ce modèle pose que des propriétés ou des classes de sols peuvent être prédites par leur relations avec 7 facteurs : d'autres informations sur les sols (s), le climat (c), les organismes (o), le relief (r), le matériau parental (p), l'âge (a) et la position géographique (n).

Jusqu'au début des années 2010, la CNS est demeurée une activité académique de recherche, puis elle est devenue plus opérationnelle en délivrant des informations à la communauté scientifique, mais aussi aux acteurs locaux, décideurs et politiques (Minasny and McBratney, 2016, Arrouays et al., 2017a). L'un des exemples de l'émergence de l'opérationnalité de la CNS est l'initiative *GlobalSoilMap* (Sanchez et al., 2009; Arrouays et al., 2014). Son objectif est de produire des prédictions de 12 propriétés majeures des sols, selon des spécifications précises, sur le monde entier, en utilisant une approche « ascendante » (depuis les pays vers le globe). En même temps, des approches « descendantes » (depuis le monde vers les pays) ont été développées. SoilGrids est l'un des meilleurs exemples utilisant ce type d'approche. Le produit SoilGrids à 1km de résolution (10 propriétés, 2 classes de sol) a été réalisé en 2014 en utilisant 110 000 profils et 75 co-variables environnementales (Hengl et al., 2014). En 2017, le produit SoilGrids250m a été réalisé en utilisant 150 000 profils and 280 co-variables. Ces produits sont librement téléchargeables en ligne (https://soilgrids.org, v0.5.3). Leurs versions actuelles ne délivrent pas d'estimations des incertitudes. En 2017, le Partenariat Mondial sur les Sols (porté par la FAO), a produit une carte globale du carbone organique des sols (GSOCmap, http://54.229.242.119/GSOCmap/), ce qui constitue un signal que la CNS est maintenant reconnue au plus haut niveau de la sphère politique. Pour la production de cette carte, environ deux-tiers des pays ont fourni des prédictions nationales, ce qui montre que la CNS est à présent opérationnelle aux échelles

nationales grâce aux efforts de transfert de technologie. Le reste du monde a été couvert par des approches « descendantes" (principalement à partir des données LUCAS pour certains pays d'Europe et à partir de SoilGrids pour le reste du monde). Cette carte est à la résolution de 1km, ne produit des estimations que pour la couche 0-30 cm, et ne donne pas d'estimations des incertitudes.

En pratique, les pays ayant délivré des prédictions quasi-complètes conformes aux spécifications *GlobalSoilMap* restent relativement rares (Australie et USA), mais de nombreux essais de cartographie de certaines propriétés ont été réalisés dans plusieurs pays (Brésil, Chili, Chine, Danemark, France, Hongrie, Nigéria, Ecosse, Corée du Sud) et de très nombreux autres pays (e.g., Croatie, Estonie, Kenya, Madagascar, Sri-Lanka) ont produit des cartes nationales de quelques propriétés, bien que la plupart n'aient pas suivi strictement les spécifications *GlobalSoilMap*.

Cette thèse débute par une revue de l'état de l'art dans le **Chapitre 2** qui traite de la cartographie numérique sur de vastes surfaces. Cette revue est fondée sur 160 articles publiés de 2003 à 2019. Nous y identifions les principales questions et défis pour la communauté travaillant sur ce sujet. La plupart des travaux se sont concentrés sur le carbone organique des sols (COS) et la texture (argiles, limons, sables). Parmi les travaux sur le carbone, la plupart se sont focalisés sur les 30 premiers centimètres et très peu ont abordé le carbone profond ainsi que le potentiel de stockage additionnel de carbone.

La plus grosse partie de la thèse se concentre sur le COS pour deux raisons principales : (i) le stock de COS est le plus grand réservoir terrestre de carbone organique, ce qui en fait un élément crucial pour le cycle global du C ; ii) le COS est un élément essentiel de la fourniture de services écosystémiques liés au sols, comme la production de nourriture, la régulation des eaux, le contrôle de l'érosion, la biodiversité et la régulation du climat (Sanchez et al., 2009; Adhikari and Hartemink, 2016; Rumpel et al., 2018). La convention des Nations Unies pour combattre la désertification a identifié la cartographie des stocks de COS comme un indicateur pour détecter et surveiller la dégradation des terres (IUCN, 2015). En parallèle, lors de la COP21, l'initiative "4 per 1000 carbon sequestration in soils for food security and the climate" (4 per 1000, https://www.4p1000.org/understand) a été lancée avec pour objectif une augmentation relative des stocks de COS de 0.4% par an, afin d'atténuer les émissions globales de gaz à effet de serre, de combattre la dégradation des sols, d'augmenter la sécurité alimentaire ainsi que de favoriser l'adaptation de

l'agriculture au changement climatique (Minasny et al., 2017; Soussana et al., 2015, 2019). A cause de la grande importance de ces stocks de COS, leur estimation spatiale ainsi que celle de leur potentiel de stockage additionnel font l'objet d'une attention soutenue. C'est en conséquence l'un des objectifs principaux de ma thèse.

Le développement de la CNS à des échelles multiples peut conduire à de nombreuses prédictions différentes sur les mêmes surfaces, ce qui peut conduire à des interrogations sur quelle carte utiliser ou sur comment combiner ces différentes cartes. En outre, certains pays ne disposent pas encore d'assez de points d'apprentissage pour produire une carte nationale. Je traite de ces questions et je propose des solutions dans le **Chapitre 3** de cette thèse.

La densité apparente (ou masse volumique) est un des paramètres nécessaires pour la conversion des teneurs pondérales en stocks volumiques pour estimer les stocks de COS. Cependant, elle est fréquemment très peu renseignée dans les bases de données en raison de son coût d'acquisition. C'est pourquoi des fonctions de pédo-tranfert (FPT) sont souvent utilisées pour prédire des valeurs manquantes. Toutefois, ces fonctions utilisent le plus souvent des équations très générales et ne vérifient pas leur domaine de validité. C'est pourquoi, dans le **Chapitre 4**, je développe des fonctions basées sur de l'apprentissage automatique (en anglais: machine learning) et je teste une méthode pour déterminer leur domaine de validité afin d'éviter des extrapolations abusives.

Une grande proportion des cartes de COS se concentre sur les horizons de surface. Pour autant, les stocks de COS profonds (>30 cm) représenteraient environ 53% de ceux contenus dans le premier mètre (environ 1 500 Pg) et 71% de ceux contenus dans les deux premiers mètres (environ 2 400 Pg) (Batjes, 1996). Pour mieux estimer la distribution spatiale des stocks de COS profonds, l'estimation de l'épaisseur du sol (en anglais : soil thickness) est un paramètre essentiel à connaître. Cependant, les observations sur l'épaisseur du sol sont souvent « censurées à droite » (c'est-à-dire que les observations mesurées atteignent une profondeur inférieure à la profondeur réelle) ce qui rend les prédictions plus difficiles que pour des propriétés qui disposent de mesures continues sur l'ensemble de leurs valeurs réelles. La façon de traiter cette difficulté fait l'objet du **Chapitre 5** de cette thèse.

Il est généralement admis que les sols ont un potentiel limité de séquestration de carbone stable, défini par le concept de saturation des sols en COS (Hassink, 1997; Angers et al., 2011; Chen et al., 2019). En conséquence, le potentiel de

séquestration additionnelle peut être estimé par la différence entre la saturation des sols en COS et leurs valeurs actuelles. Une carte de ce potentiel peut constituer un outil d'aide à la décision afin de mieux localiser les zones où plus d'efforts doivent être concentrés en raison d'un potentiel de séquestration supérieur. Dans le **Chapitre 6**, je montre comment cartographier ce potentiel de séquestration dans les couches 0-30 et 30-50 cm, en utilisant une équation empirique proposée par Hassink (1997) et des techniques de CNS.

Le stock de COS particulaire (non associé à la fraction minérale fine du sol) peut constituer un pourcentage significatif des stocks totaux. La théorie liée à la saturation ne s'applique donc pas dans ce cas. Une autre limite de la théorie liée à la saturation est que dans de nombreux contextes agro-pédo-climatiques, elle ne peut pas être atteinte. Dans le **Chapitre 7** de cette thèse, je montre comment estimer un potentiel de stockage additionnel en utilisant une approche fondée sur des données observées dans différentes zones discriminées par des facteurs de contrôle de la dynamique du carbone. J'estime statistiquement ce potentiel dans les couches 0-30 et 30-50 cm en me fondant sur des quantiles des valeurs observées afin d'évaluer le potentiel théorique de stockage dans le cadre de l'objectif 4 pour mille.

Le dernier chapitre de la thèse (**Chapitre 8**) se concentre sur les principales conclusions et perspectives de ce travail. J'y résume les principales conclusions de mon travail de thèse. Pour les dernières discussions et perspectives, j'ai en particulier bénéficié des discussions et conclusions de la dernière conférence conjointe des groupes de travail « Digital Soil Mapping » et « GlobalSoilMap » organisée à Santiago (Chili) en mars 2019 (Arrouays et al., 2019).

Principaux résultats et conclusions partielles

Dans le **Chapitre 2, «Digital mapping of soil information at a broad-scale: A review",** je réalise une revue des essais de CNS sur de vastes espaces (>10 000 km²) et basée sur 160 articles publiés de 2003 à mai 2019. L'objectif de cette revue était de synthétiser les progrès récents, les défis et les perspectives de la cartographie numérique.

Cette revue montre tout d'abord une forte augmentation de la CNS sur de vastes espaces (Figure 2.1). La distribution géographique de ces travaux place l'Australie et la Chine en tête, suivies par la France, les USA, le Royaume Uni (principalement l'Ecosse) et le Danemark. Cette revue montre également les grandes tendances des évolutions méthodologiques de ces dernières années,

comme, par exemple, l'augmentation relative du recours à des méthodes de type « apprentissage automatique ».

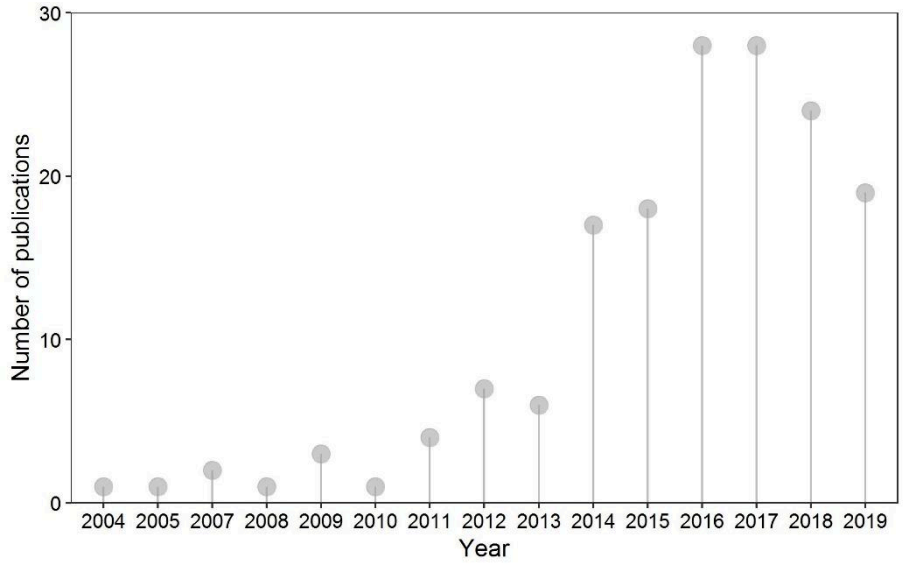


Figure 2.1 Nombre de publications par année (l'année 2019 est incomplète)

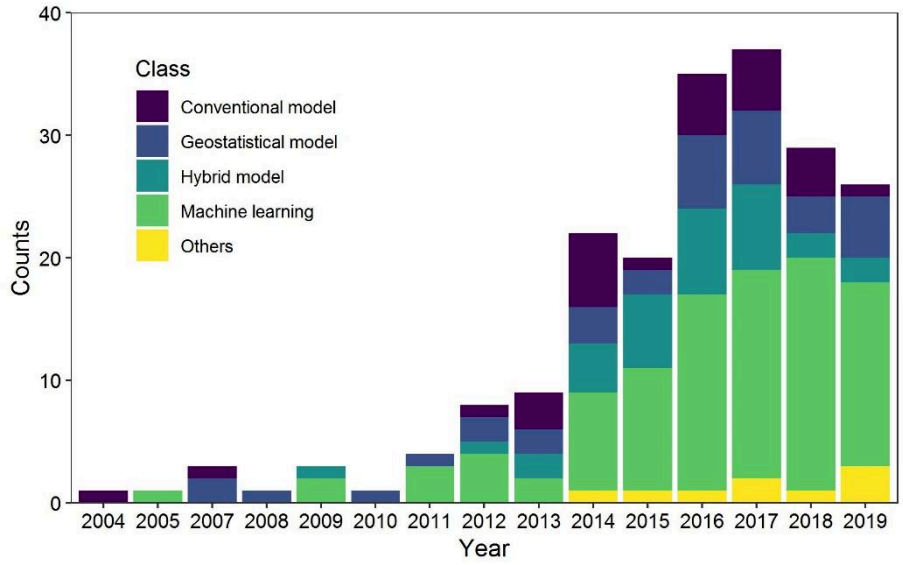


Figure 2.9 Fréquence d'utilisation de différents modèles prédictifs

Elle examine également les supports de publications les plus fréquemment utilisés. Les distributions géographique, sémantique et temporelle des points d'apprentissage et les méthodes de validation utilisées sont rarement optimales, car les études se basent souvent sur des données historiques provenant de cartographies des sols. La variable du sol la plus fréquemment cartographiée est le COS, mais se limite le plus souvent aux valeurs des couches de surface (inférieures ou égales à 30 cm) et relativement peu d'études abordent les potentiels de séquestration ou de stockage additionnel du COS. L'épaisseur du sol reste très peu cartographiée malgré son importance capitale pour l'estimation de

stocks totaux ou du réservoir en eau utile des sols. La CNS est passée progressivement de la recherche académique à la production opérationnelle, et ceci même au plus haut niveau politique. J'examine les principaux travaux portant sur des évolutions passées ou futures du COS et je montre qu'ils sont essentiellement basés sur des modèles empiriques et statiques n'incorporant pas des modèles mécanistes de la dynamique du COS. Je montre également la diversité des cartes prédictives (en particulier du COS) sur de mêmes espaces, en utilisant différents jeux de données, différents modèles, et différentes étendues spatiales. A l'issue de cette revue, je concentre mes travaux sur (i) la possibilité de tirer parti de ces différentes prédictions pour optimiser les cartes nationales de COS, (ii) la prédiction de paramètres essentiels à la prédiction des stocks de COS (masse volumique apparente et épaisseur du sol) et (iii) la prédiction des potentiels de séquestration et de stockage additionnels de COS.

Dans le **Chapitre 3** "**Model averaging for mapping topsoil organic carbon in France**" je teste la possibilité d'améliorer la précision de plusieurs cartes du COS de France métropolitaine obtenues à partir l'approches nationales, continentales, ou globales, en utilisant des modèles de mélange de type « ensemble » calibrés à partir de données sur réseau de mesures de la qualité des sols de France (RMQS). De façon importante, je montre que tous les modèles de mélange permettent une légère augmentation de la performance des prédictions et qu'un nombre relativement faible d'observations stratifiées dans l'espace (200 points, soit une densité d'un point pour 2 500 km²) permet une calibration satisfaisante des modèles de mélange. Cette augmentation de la performance des prédictions reste toutefois relativement limitée dans la mesure où les cartes nationales utilisées disposaient d'une grande densité d'informations ponctuelles. je simule des situations où les cartes nationales sont calibrées à partir de peu d'information, voire sont inexistantes. Pour ce faire, Je reproduis les prédictions nationales en diminuant progressivement la quantité de points d'apprentissage, jusqu'à simuler l'absence totale de carte nationale (Figure 3.8).

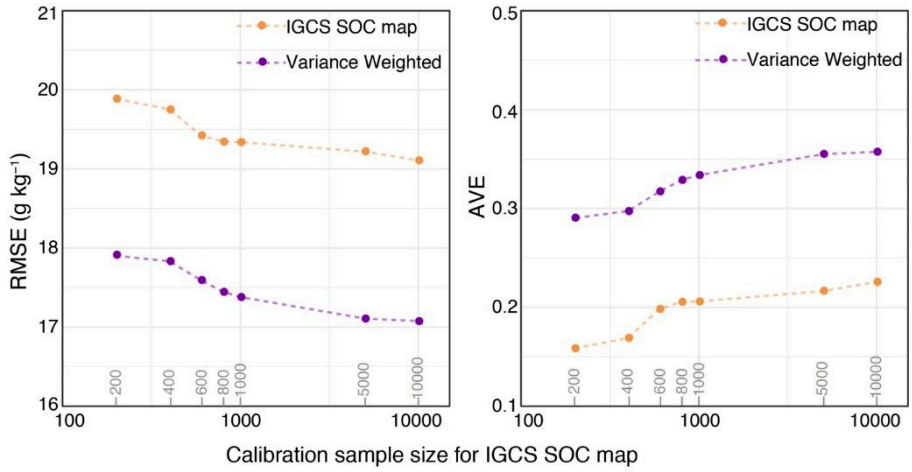


Figure 3.8 Performance du modèle utilisant la variance pondérée (WV)pour la méthode ensemble, en fonction de l'effectif des jeux de données de calibration du modèle (200, 400, 600, 800, 1000, 5000 and 10000 points)

Dans tous les cas, les modèles de mélange augmentent la performance de la prédiction, même si elle diminue très fortement en l'absence de carte nationale. Je produis ainsi une référence pour les pays disposant de très peu d'information et des recommandations sur les données minimales à acquérir afin de tirer au mieux parti des prédictions plus globales.

Dans le **Chapitre 4 « Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area",** j'établis une FPT pour prédire la densité apparente des sols en utilisant le modèle GBM calibré sur la région Centre-Val de Loire. En effet, bien que cette information soit essentielle pour calculer des stocks, elle est fréquemment absente des bases de données. Pour pallier cette absence, le recours à des FPT, calibrées sur des données plus facilement disponibles, est le moyen le plus couramment utilisé. En réduisant l'étendue sémantique et spatiale de mon domaine de calibration, puis en l'appliquant à la France entière, je simule des situations fréquentes où cette variable d'intérêt n'est disponible que sur une partie d'une étendue plus grande. J'utilise une distance euclidienne standardisée pour identifier les situations qui sont en dehors du domaine de validité de la FPT établie sur la région. Afin de concilier la précision des prédictions et le nombre de situations pouvant être prédites avec la FPT je teste différents seuils de cette distance euclidienne pour exclure des situations du domaine de validité de la FPT.

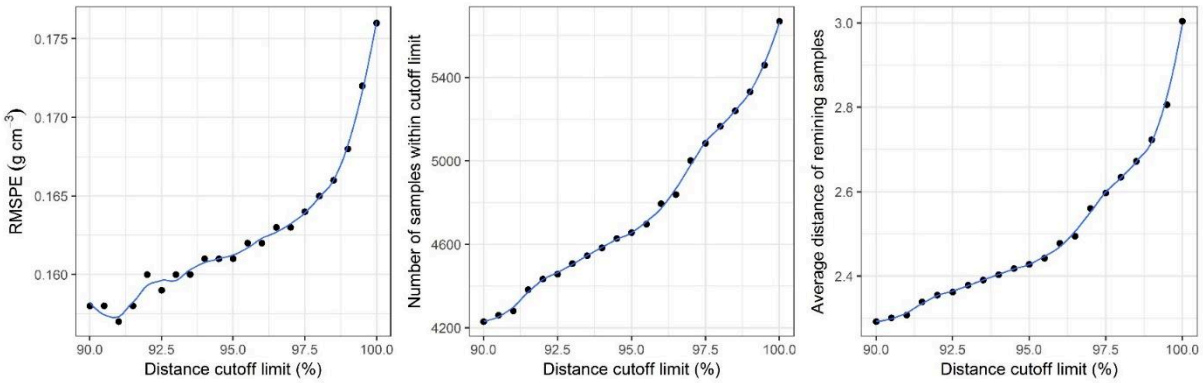


Figure 4.7 Evolution de l'indicateur RMSPE, du nombre et de la distance moyenne des échantillons restants en dessous du seuil de distance lorsque ce seuil distance augmente de 90 à 100%

Je montre que des stratégies d'échantillonnages additionnels peuvent permettre d'accroître sensiblement le domaine de validité et la précision de la FPT. Des stratégies orientées ou des densifications systématiques permettent une large amélioration de la robustesse des FPTs. Ce travail constitue ainsi une approche méthodologique pour orienter un échantillonnage en vue d'étendre le domaine de validité des FPT.

Dans le **Chapitre 5**, "**Probability mapping of soil thickness by random survival forest at a national scale",** je cartographie la probabilité qu'à le sol d'excéder une épaisseur donnée. Les données d'épaisseur (ou de profondeur) du sol ont la particularité d'être fréquemment « censurées à droite » en termes statistiques ; il est très fréquent que l'épaisseur observée soit inférieure à l'épaisseur réelle du fait de considérations très pratiques comme la longueur de la tarière utilisée (le plus souvent de 120 cm) ou de contraintes de temps lors du creusement de profils. En science du sol, très peu d'études considèrent l'effet de telles données censurées. En revanche, des techniques de traitement de données censurées à droite sont fréquemment appliquées en médecine, lorsque, par exemple, on cherche à estimer la durée probable de survie de patients. C'est ce type de méthode que j'ai appliqué à l'épaisseur du sol. Je démontre ainsi comment ces données peuvent être prises en compte pour modéliser la probabilité que l'épaisseur du sol excède une valeur donnée En utilisant le modèle « Random Survival Forest (RSF)", je modélise la probabilité qu'ont les sols d'excéder une valeur donnée en utilisant des co-variables représentant les principaux facteurs de formation des sols et j'extrapole cette modélisation à l'ensemble de la France métropolitaine. Comme exemple, je produis des cartes de la probabilité d'atteindre ou d'excéder les profondeurs standards recommandées par les

spécifications de *GlobalSoilMap* : 5, 15, 30, 60, 100, and 200 cm, dont je montre ci-après un exemple pour la profondeur de 100 cm.

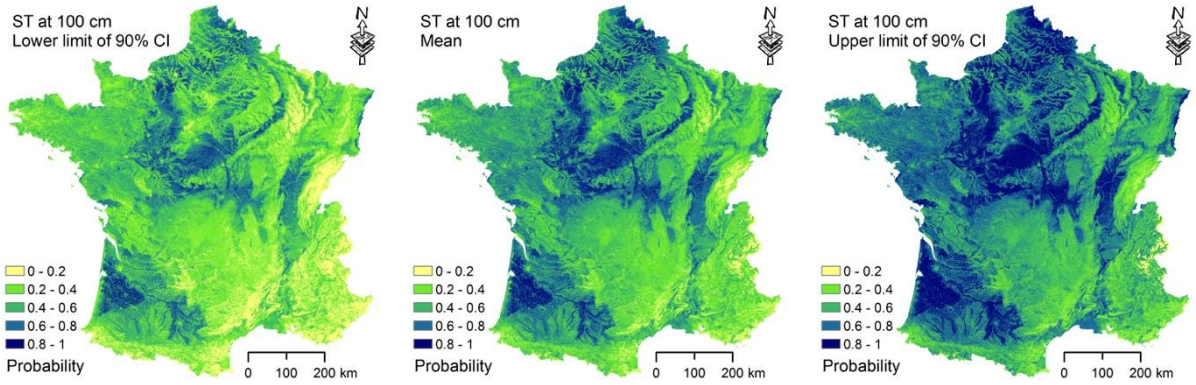
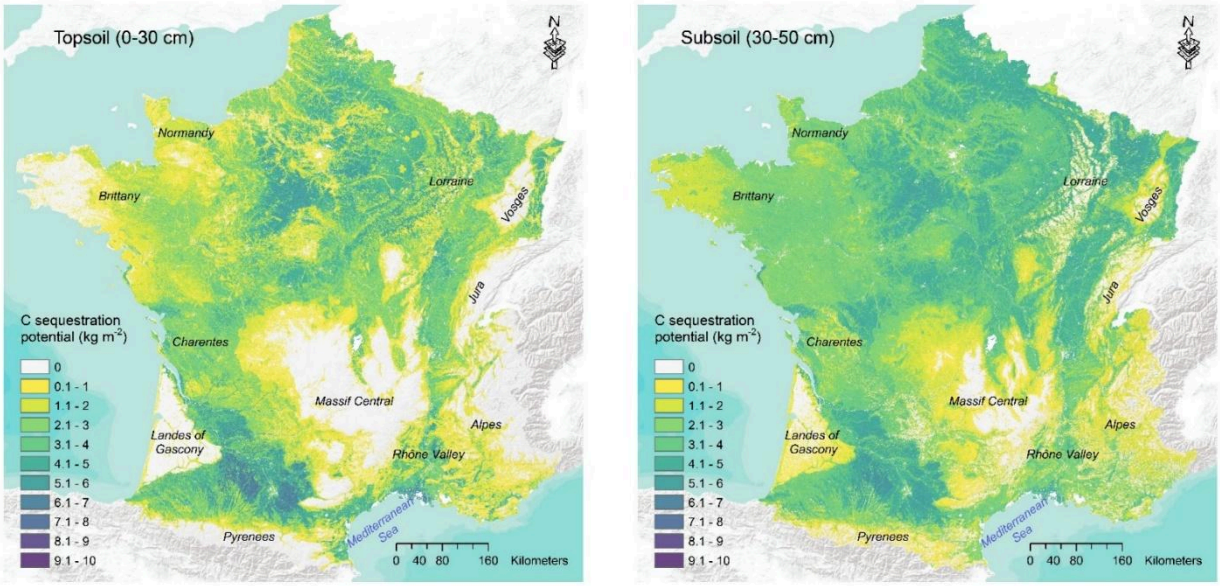


Figure 5.7 Carte de probabilité d'excéder une épaisseur de 100 cm (au milieu) et ses intervalles de confiance à 90% (à gauche et à droite)

J'utilise une approche de type « bootstrapping » pour estimer les intervalles de confiance à 90%. Je montre que la méthode RSF permet de corriger les données censurées à droite et que cette correction est plus efficace pour les sols les plus minces et les plus épais. Je propose ainsi une approche nouvelle pour modéliser ce type de données sur les sols. Elle permet de produire des cartes de probabilité pour toutes les épaisseurs inférieures aux observations les plus profondes présentes dans le jeu de calibration. Les résultats sont applicables pour définir des stratégies pour des campagnes additionnelles d'acquisition de cette donnée. Ils peuvent aussi être utilisés de façon pratique pour des problématiques d'ingénierie géotechnique.

Dans le **Chapitre 6** « **Fine resolution map of top- and subsoil carbon sequestration potential in France**" je traite du potentiel théorique de séquestration additionnelle de COS dans les sols de France. Il est communément admis qu'il existe une limite supérieure au potentiel de stockage de carbone stable, définie comme étant la saturation du sol en carbone. Dans cette partie, j'estime cette saturation théorique dans les couches 0-30 et 30-50 cm des sols de France à partir d'une équation empirique définie par Hassink (1997). Puis J'estime un potentiel théorique de séquestration additionnelle (SOCsp pour « SOC sequestration potential) de COS par différence entre les stocks observés et les stocks à saturation théorique sur l'ensemble de la grille des points du RMQS. Je cartographie ensuite ce SOCsp à haute résolution spatiale sur la France métropolitaine en utilisant une approche de régression-krigeage utilisant des co-variables environnementales externes.



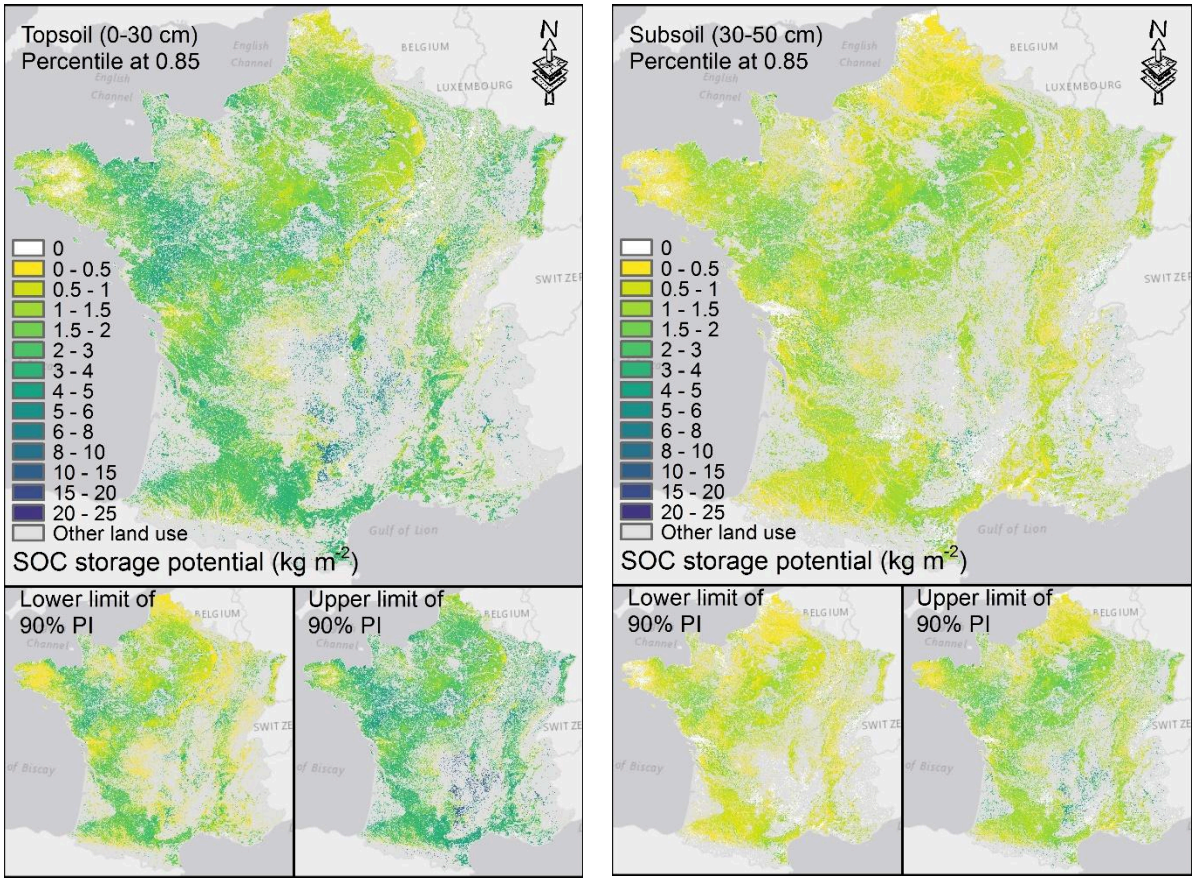
Figures 6.6 et 6.7 Carte du potentiel de séquestration additionnel de carbone organique pour les couches de surface (0–30 cm) et de sub-surface (30–50 cm) en France métropolitaine

Les résultats mettent en évidence les facteurs de contrôle de la séquestration en COS, qui diffèrent entre les stocks superficiels (0-30 cm) et les stocks plus profonds (30-50 cm). Le facteur principal de séquestration dans les couches de surface est l'utilisation du sol. Le déterminant principal de la saturation des couches les plus profondes semble être le matériau parental. Au total, le potentiel théorique de séquestration en COS des sols de France et très important (1008 Mt C pour 0-30 cm and 1360 Mt C pour 30-50 cm) si on le compare aux estimations des stocks totaux d'environ 3,5 Gt C pour 0-30 cm. Ces résultats ne signifient aucunement que ce potentiel peut être atteint, mais ils indiquent les zones les plus déficitaires, et suggèrent qu'un effort particulier de recherche doit être entrepris pour tester des méthodes permettant un enrichissement et une stabilisation du COS en profondeur. Ils mettent aussi en évidence les zones théoriquement saturées et montrent qu'environ 176 Mt C sont au-dessus de la saturation théorique et pourraient donc être très vulnérables en cas de changement d'occupation du sol.

Dans le **Chapitre 7 "National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones"** je m'intéresse plus spécifiquement au potentiel de stockage additionnel de COS total des sols cultivés de France. Pour ce faire, j'utilise une approche dite « data-driven » c'est-à-dire fondée sur les données actuelles

observées. L'hypothèse implicite de cette approche est de considérer que sous les pratiques en vigueur en France, et sous des conditions pédo-climatiques données, les stocks observés les plus importants représentent ceux qui devraient être théoriquement atteignables. L'étude est restreinte aux sols cultivés (vignes et vergers inclus) car je fais l'hypothèse que ce sont ces sols qui présentent le potentiel de stockage additionnel le plus important – hypothèse qui semble par ailleurs corroborée par les résultats obtenus dans le **Chapitre 6** – et que ce sont également les sols où des changements de pratiques sont le plus facilement susceptibles d'influer sur le stockage de COS.

Je délimite différentes zones (carbon landscape zones : CLZs) en définissant des « clusters » caractérisés par des facteurs semblables de contrôle sur stockage en COS (principalement des indices climatiques et la teneur en argile des sols). Le nombre de clusters est fixé par un compromis entre la possibilité de les différencier et le nombre de points qu'ils contiennent pour en effectuer un traitement statistique. Je calcule ensuite des estimations des valeurs les plus élevées atteignables en utilisant ces centiles correspondants aux valeurs limites inférieures observées pour 80, 85, 90 et 95% des effectifs de chaque CLZ. Puis je calcule un potentiel de stockage additionnel par différence entre les données observées et les valeurs obtenues en utilisant ces différents centiles. Quand les valeurs observées sont supérieures aux limites des centiles, le stockage additionnel potentiel est, par construction, considéré comme nul. Les centiles utilisés ont évidemment une influence très importante sur les résultats.



Figures 7.9 et 7.10 Potentiel de stockage de carbone organique additionnel pour les sols cultivés et les couches de surface (0-30 cm) et de sub-surface (30-50 cm) pour le centile de 85% (en haut). Les cartes figurées en bas correspondent aux valeurs correspondent aux intervalles de prédiction de 90%

Quand ces centiles passent de 80% à 95% le potentiel national de stockage additionnel de COS triple, passant de 336 à 1020 Mt C pour 0-30 cm et de 165 à 433 Mt C pour 30-50 cm. Ceci montre bien la grande sensibilité du choix des centiles aux résultats obtenus. Les potentiels de stockage additionnels sont en règle générale inférieurs aux potentiels théoriques de séquestration, surtout en ce qui concerne le sous-sol. Ces potentiels se réfèrent à des pratiques actuelles. Il reste donc possible que de nouvelles pratiques innovantes puissent éventuellement conduire à des stockages supérieurs. Cette approche présente des avantages certains en termes opérationnels car elle peut permettre de fixer des objectifs de stockage en prenant en compte des considérations sur leur faisabilité, tant du point de vue des politiques de soutien que de celui des agriculteurs. La robustesse de des estimations devrait être par ailleurs estimée en utilisant des approches complémentaires telles que celles offertes par des modélisations plus mécanistes.

Conclusion générale et perspectives

Principales conclusions et perspectives de mon travail

Mon travail de thèse traite de plusieurs questions qui contribuent à des avancées notables en matière de CNS et du prolongement de la CNS à la cartographie des fonctions des sols (CNFS) et de leurs services écosystémiques (CNSES). Ma conclusion synthétise en premier des résultats que j'ai obtenus et leurs perspectives.

Dans le **Chapitre 2**, je réalise une revue préliminaire des tentatives de CNS sur de vastes espaces dans des contextes variés. Je pointe, en particulier, des questions relatives aux stratégies d'échantillonnage et de validation, ainsi qu'aux tendances quant aux méthodes utilisées. J'identifie aussi dans cette revue une priorité internationale qui est la cartographie des stocks de COS, ainsi que ses faiblesses liées aux manques de données concernant la densité apparente des sols et leur épaisseur. Je mets également en évidence les difficultés inhérentes à la prolifération de nombreuses prédictions issues de la CNS, et établies avec différentes co-variables, données ponctuelles d'entrée, et différents modèles, ainsi que sur des étendues géographiques différentes. Bien que ce chapitre constitue une première vue générale de la CNS sur de vastes espaces, un de mes objectifs à court terme est d'en approfondir l'analyse pour en produire un article plus générique.

Dans le **Chapitre 3** .je teste une méthode pour agréger différentes prédictions cartographiques. Je montre en particulier que l'utilisation de modèles « ensemble » apporte toujours une amélioration à ces prédictions. Ceci constitue un premier pas vers l'agrégation de cartes d'origine différentes qui permettent de capturer différents facteurs de contrôle de la propriété étudiée (le COS). Je montre également de façon concrète comment ce cas peut être généralisé à des pays particulièrement pauvres en données nationales en optimisant les échantillonnages complémentaires à acquérir.

Dans le **Chapitre 4**, je traite du domaine de validité des FPTs. Ce travail peut être considéré comme une contribution méthodologique pour les parties du monde où les données permettant de construire de telles FPTs sont peu nombreuses, ou sont localisées sur une portion restreinte du territoire à étudier. Il apporte des éléments sur la caractérisation du domaine de validité d'une FPT et sur la façon de raisonner un échantillonnage pour étendre ce domaine de validité.

Dans le **Chapitre 5**, je teste une nouvelle méthode pour prendre en compte des données censurées à droite en ce qui concerne l'épaisseur du sol. Ce point est

important car une grande proportion des données historiques relatives à l'épaisseur du sol sont certainement censurées à droite en raison de limitations pratiques dues au creusement de sondages ou de profils. La méthode que j'utilise permet de cartographier la probabilité qu'à le sol d'excéder une épaisseur donnée. Elle est applicable dans le domaine de l'étendue des épaisseurs observées, mais est plus précise pour les épaisseurs les plus faibles et les plus grandes. Des travaux complémentaires, suggérés dans ma thèse, devraient permettre d'utiliser cette méthode pour prédire une épaisseur de sol en tout point de l'espace.

Les **Chapitres 6 et 7** constituent de premières tentatives pour passer de la CNS à une cartographie des fonctions potentielles des sols. Au lieu de prédire une variable d'état (une teneur ou un stock de COS), je tente de prédire et son potentiel additionnel d'augmentation, soit en termes de séquestration de COS « stable », soit en termes de stockage de COS total. Dans le **Chapitre 6**, je montre que les potentiels théoriques de séquestration sont très élevés, en particulier en ce qui concerne le sous-sol, ce qui soulève la question d'une meilleure compréhension des pratiques et des mécanismes permettant d'augmenter la séquestration de COS en profondeur. Dans le **Chapitre 7**, je mets en œuvre une méthode « dirigée par les données » (data-driven) fondée sur les données observées de stocks de COS. Je calcule différents centiles supérieurs de ces données et les quantités additionnelles de COS qu'il faudrait ajouter pour parvenir aux limites inférieures de ces centiles. Les calculs sont, de par leur nature même, très sensibles au choix des centiles retenus. Toutefois, un résultat majeur est que la plupart des stocks additionnels calculés en utilisant la méthode des centiles sont très largement inférieurs aux potentiels théoriques de séquestration, surtout dans le sous-sol. Ceci signifie encore que le stockage de carbone dans les horizons profonds reste une priorité à explorer. Mon analyse permet également de différencier les principaux facteurs de contrôle de la distribution du COS, qui diffèrent entre les couches de surfaces et du sous-sol. Ces cartes constituent des estimations de potentiels théoriques. Plus de recherches sont nécessaires pour évaluer dans quelle mesure des changements de pratiques ou d'usages permettraient d'atteindre des objectifs de stockage ou de séquestration. Le rapport publié récemment en France pour évaluer la possibilité d'atteindre l'objectif du « 4 pour 1000 » (Pellerin et al., 2019) constitue un bon exemple permettant de relier l'objectif d'atteindre un potentiel à sa faisabilité. Un autre défi est d'évaluer comment les effets de ces changements d'usage ou de pratiques peuvent être incorporés dans des modèles mécanistes (e.g., Century ou Roth-C).

Si cela est théoriquement possible, ces modèles demandent toutefois de nombreuses données sur les sols, le climat et les pratiques de gestion, ce qui pourrait rendre difficile leur application aux échelles nationales ou globales. Considérant les limites des différentes approches, « data-driven » et modélisation mécaniste, elles me semblent aujourd'hui complémentaires.

Pertinence vis-à-vis de certains défis posés au domaine scientifique « Pedometrics »

Dans cette partie de la conclusion, je fais référence aux "défis posés à Pedometrics" proposés par Heuvelink (2019). Je me concentre sur les défis qui sont le plus directement en lien avec mon travail de thèse et j'en tire des éléments de réflexion et de propositions dans le cadre de la CNS. Je discute ensuite d'autres perspectives concernant la CNS et la cartographie de propriétés fonctionnelles ou de services rendus par les sols.

Pouvons-nous développer des estimations des incertitudes facilement communicables?

Compte tenu de l'importance de délivrer des incertitudes, de plus en plus d'études fournissent des intervalles de confiance ou de prédiction pour les prédictions issues de la CNS. C'est ce que j'ai réalisé dans les **Chapitres 5, 6 et 7** de ma thèse. Toutefois, beaucoup d'utilisateurs finaux de ces produits ne tiennent pas compte ou ne comprennent pas ces évaluations des incertitudes (Arrouays et al., 2019). Il est de notre devoir de mieux communiquer sur ces dernières, et d'expliquer en quoi elles sont importantes et comment les utiliser. De fait, les demandes des utilisateurs concernent de plus en plus des outils opérationnels pour l'aide à la décision et des analyses de risques. Pour ces derniers, des intervalles de confiance ou de prédiction ne semblent pas les outils les mieux appropriés. Nous devrions développer des méthodes permettant la prédiction de fonctions de probabilité de distribution des propriétés, plus adaptées à l'utilisation de modèles capables de prendre en compte des probabilités de risques. Egalement, il serait sans doute plus efficace de communiquer sur l'incertitude des sorties des modèles d'aide à la décision ou de probabilités de risques que de communiquer sur l'incertitude des données d'entrée des modèles.

Pouvons-nous développer des méthodes approfondies de changement d'échelle?

Nous sommes toujours gênés par le concept d'échelle, principalement à cause d'un manque de définition précise de ce concept. Pourtant, la plupart des modèles biogéochimiques (e.g., Century, Roth-C) sont développés à l'échelle locale.

Comment appliquer ces modèles mécanistes à des échelles plus globales, en adaptant la prédiction de leurs données d'entrée à ces échelles ? Ceci reste un défi important qui reste très prégnant dans les études cherchant à prévoir les changements de SOC selon différents scénarios ou qui utilisent des méthodes « data-driven » pour prédire des potentiels de stockage (Cf. les perspectives du **Chapitre 7**). Nous suggérons deux approches possibles: (i) explorer le potentiel de données facilement accessibles pour quantifier des paramètres d'entrées des modèles mécanistes (par exemple, les analyses Rock-Eval, Cécillon et al., 2018), pour initialiser les modèles à de larges échelles ; (2) Etablir plus d'expérimentations à long-terme, couvrant au maximum les situations agro-pédo-climatiques, pour développer de nouveaux modèles mécanistes qui prennent en compte des processus et leur variation dans l'espace.

Pouvons-nous incorporer nos connaissances pédologiques déterministes dans la CNS?

Les travaux de CNS actuels reposent sur l'apprentissage automatique où la connaissance pédologique est principalement utilisée pour définir les co-variables environnementales pertinentes. Nous risquons de fuir vers la science des données et les statistiques et de perdre notre crédit scientifique si nous continuons à ignorer la connaissance pédologique en CNS. Notre communauté a identifié ce problème et quelques études récentes portant sur le concept de « Structural equation modelling » (SEM) et des modèles d'évolution mécanistes apportent des exemples sur l'incorporation de la connaissance pédologique en CNS, bien que la précision des modèles puisse être inférieure à celle des modèles d'apprentissage automatique (Angelini et al., 2017; Ma et al., 2019). Plus d'efforts futurs devraient être consentis dans cette direction.

Pouvons-nous produire des cartes globales suffisamment précises?

De grands progrès ont été réalisés dans le cadre de l'initiative *GlobalSoilMap* pour produire des cartes à la résolution de 90 m. Comme souligné par Heuvelink (2019), il est plus facile d'affiner la résolution spatiale que la précision sémantique. Nous devons avoir pour objectif de satisfaire à la fois le besoin d'une résolution spatiale fine et des standards de précision requises. Nous évoquons à ce sujet plusieurs pistes: (1) acquérir plus de données ponctuelles sur les sols en utilisant des échantillonnages optimisés, (2) explorer le potentiel de données environnementales provenant de la télédétection satellitale et aéroportée (e.g., Sentinel 2, Sentinel 3, RISAT-2B Earth-Observation Satellite and gamma-ray); (3) tester de nouveaux modèles de prédiction spatiale (e.g., deep learning, SEM).

Il existe aussi un besoin d'estimer la précision des cartes nationales ou

globales pour un usage local, ce qui peut nécessiter un échantillonnage supplémentaire qui doit rester d'un coût abordable. Des approches de modèles « ensemble », telles que celle décrites dans le **Chapitre 3**, peuvent sans doute améliorer la précision des prédictions locales et plus globales. En outre, ces approches « ensemble » peuvent être considérées comme un premier pas vers l'intégration de prédictions à des échelles variables ainsi qu'un premier progrès en ce qui concerne l'harmonisation de produits différents.

Pouvons-nous quantifier l'incertitude des observations sur les sols et analyser son effet sur la cartographie des sols ?

Les observations et les analyses de laboratoires effectuées sur les sols comportent toutes une part d'incertitude qui varie selon les observations, les analyses, et les propriétés d'intérêt. Dans les études actuelles, les observations et les mesures sont considérées comme représentant la réalité, et leurs incertitudes sont le plus souvent ignorées. Pour délivrer une information de grande qualité, la caractérisation et la prise en compte de ces incertitudes sont nécessaires. Le développement rapide de prédictions issues de mesures de télédétection rapprochée, souvent caractérisées par de fortes incertitudes, est un très bon exemple de la nécessité de prendre en compte ces incertitudes. Les erreurs de mesures ou de prédictions doivent être prises en compte en CNS car elles se propagent inévitablement sur les cartes finales. Ce défi, est clairement est pris en compte dans certaines parties de ma thèse. Dans le **Chapitre 4**, je propose une estimation du domaine de validité des FPTs afin d'exclure des incertitudes qui se propageraient ensuite sur les estimations des stocks de SOC, ainsi que des méthodes d'extension de ce domaine. Dans le **Chapitre 5** la méthode que je propose montre qu'il est possible de modéliser et de cartographier une donnée imprécise, car censurée, et que la carte résultante peut permettre de raisonner un échantillonnage complémentaire afin de réduire l'incertitude des prédictions. Dans le **Chapitre 6**, la proportion moyenne de SOC stable dans la fraction fine utilisée dans l'équation de Hassink (1997) introduit une incertitude dont j'ai par ailleurs simulé certaines conséquences dans un article (Chen et al., 2019) non inclus dans ce manuscrit. Cette incertitude pourrait être réduite en intégrant dans la prédiction les changements historiques d'usages *via* la télédétection, à la condition d'être en mesure d'estimer la conséquence de ces changements d'usage sur la dynamique du SOC pour la fraction grossière et la fraction supposée stabilisée par la fraction fine. En outre, une approche par modélisation (e.g., Century, Roth-C) simulant ces changements en fonction de changements d'usage

ou de pratiques aiderait également au choix de centiles plus réalistes pour une approche « data-driven » (**Chapitre 7**) et réduirait ainsi l'incertitude liée au choix de ces centiles.

Comment cartographier les fonctions des sols ?

L'enjeu est ci de passer d'une approche CNS à des approches CHFS ou CNSES telles que définies en début de conclusion. De très nombreux utilisateurs ont besoin de cartes des fonctions des –ou des services rendus par les - sols pour la modélisation et l'aide à la décision et nous devons porter plus d'attention ce type de cartographie. Des collaborations doivent être nécessairement établies avec d'autres disciplines pour relever ces défis. Les **Chapitres 6 et 7** constituent de bons exemples sur comment passer de la prédiction d'une propriété à celle d'une fonction potentielle. Toutefois, ils caractérisent des potentiels théoriques et manquent fortement d'une connexion avec des références indiquant comment, où, et jusqu'à quel niveau ce potentiel pourrait être atteint.

Que pouvons-nous apprendre sur les processus à partir d'outils d'apprentissage automatique?

Les outils d'apprentissage automatique se focalisent actuellement presque exclusivement sur la prédiction de types de sols ou de certaines de leurs propriétés. Bien que certains outils permettent une première analyse des facteurs de contrôle des distributions géographiques, ils ne sont pas souvent utilisés pour améliorer notre compréhension, ni pour confirmer nos connaissances, et encore moins pour découvrir de nouveaux processus. Plus d'efforts sont à conduire pour que les méthodes de CNS puissent contribuer à nos connaissances pédologiques.

Autres pistes d'amélioration

Dans la conclusion générale de ma thèse je développe quelques autres pistes d'amélioration que je n'évoque que très brièvement ici. Il s'agit en premier lieu de la question de la propriété et du libre accès aux données qui sont des questions cruciales. Mis à part la signature de conventions spécifiques, qui relève de décisions politiques, les modèles ensembles (**Chapitre 3**) constituent une solution pour partager des résultats sans partager les données brutes sur les sols. Une autre piste prometteuse récemment proposée (Padarian and McBratney, 2019) pourrait être de partager et joindre des modèles sans partager les données d'entrée privées. Un deuxième point concerne notre capacité à former et à entraîner de nouvelles équipes aux techniques de CNS. Ce point est crucial à bien des égards et constitue une des clés de la réussite de programmes de cartographie

« ascendante » tels que *GlobalSoilMap*. De gros efforts ont été fournis par de nombreuses institutions (e.g., FAO-GSP, ISRIC-World Soil Information, Université de Sydney, et NRCS-USDA). Une stratégie pertinente pourrait être d'associer dans de mêmes équipes des pédologues expérimentés en cartographie traditionnelle et des spécialistes du traitement des données, formés aux méthodes de CNS. Un troisième point concerne la fusion de données issues de capteurs multiples dont le développement est exponentiel. Ceux-ci sont souvent utilisés séparément, et leur fusion raisonnée avec d'autres informations spatialisée pourrait constituer une voie prometteuse pour l'avenir. Un dernier point concerne le choix – et la parcimonie dans le choix – des co-variables. Avec la multiplication de ces dernières, la tentation est grande de vouloir les incorporer toutes. Je considère que du point de vue de l'interprétation pédologique, comme de l'incertitude introduite par chaque co-variable, une sélection des co-variables et nécessaire, non seulement du point de vue de leur signification purement statistique, mais aussi de celui de leur signification pédologique.

References

Adhikari, K., Hartemink, A.E., 2016. Geoderma, 262, 101–111.

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., 2017. European Journal of Soil Science, 68(5), 575-591.

Angers, D., Arrouays, D., Saby, N., Walter, C., 2011. Soil Use and Management, 27, 448–452.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M.,..., Zhang, G.-L., 2014. Advances in Agronomy 125, 93–134.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., ..., Rodriguez, D., 2017. GeoResJ, 14, 1–19.

Arrouays, D., Piggio, L., Salazar, O., Mulder, V.L., 2019. Geoderma Regional, submitted.

Batjes, N.H., 1996. European Journal of Soil Science, 47(2), 151-163.

Bouma, J.,2014. Journal of Plant Nutrition and Soil Science, 177(2), 111-120.

Cécillon, L., Baudin, F., Chenu, C., Houot, S., Jolivet, R.,, ..., Savignac, F., 2018. Biogeosciences, 15(9), 2835-2849.

Chen, S., Arrouays, D., Angers, D.A., Martin, M.P., Walter, C., 2019. Soil & Tillage Research, 188, 53–58.

Dokuchaev, V.V. 1883. Jerusalem, Israel.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Soil Science Society of America Journal, 75(4), 1201-1213.

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Geoderma, 207, 256-267.

Hassink, J., 1997. Plant and Soil, 191(1), 77–87.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., ..., Gonzalez, M.R. ,2014. PLoS One, 9(8), e105992.

Heuvelink, G.B.M., 2019. Pedometron, 43, 9-13.

IUCN, 2015. Land degradation neutrality: implications and opportunities for conservation.

Jenny, H. 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York, NY.

Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., ..., Bardgett, R.D., 2016. SOIL, 2(2), 111–128.

Koch, A., McBratney, A., Lal, R., 2012. Nature, 492(7428), 186-186.

Koch, A., McBratney, A.B., Adams, M., Field, D., Hill, R., ..., Angers, D., 2013. Global Policy, 4, 434–441.

Lagacherie, P., McBratney, A.B., 2006. Development in Soil Science, 3–22.

Ma, Y., Minasny, B., Welivitiya, W.D.P., Malone, B.P., Willgoose, G.R., McBratney, A.B., 2019. Geoderma, 341, 195-205.

McBratney, A., Field, D.J., Koch, A., 2014. Geoderma, 213, 203-213.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. Geoderma, 117(1–2), 3–52.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., ..., Winowiecki, L., 2017. Geoderma, 292, 59–86.

Minasny, B., McBratney, A.B., 2016. Geoderma, 264, 301-311.

Padarian, J., McBratney, A.B., 2019. SOIL, under review.

Pellerin, S., Bamière, L., Launay, C., Martin, R., Schiavo, M., ..., Cardinael, R., 2019. Stocker du carbone dans les sols français, quel potentiel au regard de l'objectif 4 pour 1000 et à quel coût?.

Rumpel, C., Amiraslani, F., Koutika, L., Smith, P., Whitehead, D., Wollenberg, D., 2018. Nature, 564, 32–34.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., ..., Minasny, B., 2009. Science, 325, 680–681.

Soussana, J.F., Lutfalla, S., Ehrhardt, F., Rosenstock, T., Lamanna, C., Havlik, P., ..., Lal, R., 2019. Soil & Tillage Research, 188, 3–15.

Soussana, J.F., Saint-Macary, H., Chotte, J.L., 2015. Paris, Sept. 16, 2015.

Weil, R.R., Brady, N.C., 2016. The nature and properties of soils. Pearson.

Zhang, G.L., Liu, F., Song, X.D., 2017. Journal of Integrative Agriculture, 16(12), 2871-2885.

UNIVERSITE ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

AGRO CAMPUS OUEST

INRA SCIENCE & IMPACT

Titre : Cartographie à haute résolution de propriétés du sol à l'échelle de la France métropolitaine : application au carbone organique du sol et à ses potentiels additionnels de séquestration et de stockage

Mots clés : Cartographie Numérique des Sols ; *GlobalSoiMap* ; Fonctions des Sols.

Résumé : Cette thèse est une contribution à la Cartographie Numérique des Sols (CNS) sur de vastes espaces. Dans le Chapitre 1, je discute des principaux facteurs à l'origine de l'émergence et du développement de la CNS et j'en retrace brièvement l'historique. Dans le Chapitre 2, je réalise une revue générale de la CNS sur de vastes espaces en me fondant sur 160 articles publiés de 2003 à mi-2019. J'y synthétise et discute les principales avancées et défis pour la communauté de la CNS. Je me focalise ensuite sur le carbone organique des sols (COS) en raison de son importance fondamentale pour les services écosystémiques et le cycle global du carbone. Dans le Chapitre 3, je montre comment améliorer une carte nationale du COS en agrégeant différentes cartes de COS et je donne des des pistes sur la façon de tirer parti de prédictions globales pour les pays disposant de peu de données, en utilisant une stratégie d'échantillonnage peu coûteuse et efficace. Ensuite, dans les Chapitres 4 et 5, je me concentre sur le domaine de validité de fonctions de pédotranfert utilisées pour la prédiction de la densité apparente et je développe une approche nouvelle pour traiter l'épaisseur des sols en France. J'y propose également des stratégies efficientes pour améliorer la précision de leurs prédictions. Je passe ensuite de la CNS à la cartographie de fonctions des sols en prenant comme exemple la cartographie du potentiel de séquestation (Chapitre 6) et de stockage (Chapitre 7) en COS. Ces chapitres contribuent à améliorer de nombreux aspects concernant la CNS et son application au programme *GlobalSoiMap*. Je conclus cette thèse dans le Chapitre 8, où je discute les principaux résutats de ma thèse et où je les relie aux principaux défis de la discipline Pedometrics. Je discute les principaux apports de mon travail au regard de ces défis et je souligne les questions restant à résoudre dans le futur proche.

Title : High resolution soil property maps over mainland France: application to soil organic carbon and its additional sequestration and storage potentials

Keywords : Digital Soil Mapping; *GlobalSoilMap*; Soil functions

Abstract : This thesis is a contribution to Digital Soil Mapping (DSM) at broad scale, with applications on the French mainland territory. In Chapter 1, I discussed the main drivers for the rise and development of DSM and gave a brief history about DSM. In Chapter 2, I made a general review about broad-scale DSM by reviewing 160 selected articles from 2003 to mid-2019. I synthetized and discussed the main achievements and challenges for the DSM community. Then I decided to focus on soil organic carbon (SOC) because of its main importance for ecosystem services and global carbon cycle. In Chapter 3, I showed how to improve a national SOC map by merging various SOC maps and provided inputs on how to take advantage of global predictions in 'data-poor' countries using a low cost and efficient sampling strategy. Then in Chapters 4 and 5, I focused on the validity domain of pedotransfer functions used for bulk density predictions and on developing a novel approach to deal with soil thickness prediction over France. I also proposed efficient sampling strategies to improve the accuracy of their predictions. I moved from DSM to Digital Soil Assessment (DSA), exemplified by SOC sequestration potential in Chapter 6 and SOC storage potentials in Chapter 7. They contribute to improving some aspects related to DSM and GlobalSoilMap. In Chapter 8, I finished this thesis by discussing the most important findings of my work and relating them to main challenges of Pedometrics. I outlined the inputs that my work provided to reaching these challenges and highlighted the remaining issues to be solved in the near future.