Université catholique de Louvain
Earth and Life Institute

Towards a better understanding and prediction of the spatial distribution of organic carbon in cropland soils

François Stevens

Thèse présentée en vue de l'obtention
du grade de docteur en sciences agronomiques
et ingénierie biologique

Louvain-la-Neuve
Novembre 2018

Thesis committee

Professor Gerard Heuvelink ISRIC – World Soil Information, Wageningen, The Netherlands

Professor Marnik Vanclooster Université catholique de Louvain, Earth and Life Institute, Environmental Sciences

Professor Kristof Van Oost Université catholique de Louvain, Earth and Life Institute, George Lemaître Centre for Earth and Climate Research

Professor Patrick Bogaert, Supervisor Université catholique de Louvain, Earth and Life Institute, Environmental Sciences

Professor Bas van Wesemael, Supervisor Université catholique de Louvain, Earth and Life Institute, George Lemaître Centre for Earth and Climate Research

Professor Bruno Delvaux, President of the Jury Université catholique de Louvain, Earth and Life Institute, Environmental Sciences

Acknowledgment

La thèse est une aventure faite de rencontres et de partage. C'est la raison pour laquelle j'ai une pensée pour toutes les personnes qui m'ont accompagné. Tout d'abord je voudrais remercier mes promoteurs Patrick et Bas qui ont été disponibles et présents du début à la fin pour m'aider, me conseiller et m'encourager. Vos manières différentes d'aborder les questions scientifiques ont donné lieu à des réunions animées et productives. J'ai souvent eu la chance de bénéficier de l'interaction positive entre la créativité et de l'esprit de synthèse de Bas et l'analyse critique et la vision anticipatrice de Patrick, ce qui m'a permis d'en ressortir avec des idées neuves et concrètes. Il est clair que rien n'aurait été possible sans vous.

J'aimerais aussi remercier les autres membres de mon comité d'accompagnement. Au départ en contact avec l'unité ENGE pour réaliser une thèse dans le domaine de la pollution atmosphérique, les hasards du milieu universitaire ont fait en sorte que c'est finalement une thèse dans le domaine du carbone organique des sols qui s'offrait à moi. Alors que j'avais fait le pari de me jeter dans l'inconnu, j'ai pu compter sur mes promoteurs mais aussi sur Marnik, Kristof, et Sébastien pour m'immerger dans ces mondes nouveaux des sciences du sol, de la géomorphologie, des géostatistiques et de la modélisation environnementale. Les discussions que nous avons eues m'ont permis de faire mûrir ma compréhension dans ces domaines et j'ai beaucoup apprécié votre disponibilité et votre patience. Je voudrais également remercier Antoine (encore un Stevens) et Sebastian (Doetterl) pour m'avoir expliqué leur recherche, partagé des astuces et fourni l'accès à des données utiles, et bien sûr Bruno Delvaux pour avoir accepté de présider mon jury. Finally, I would like to thank Gerard Heuvelink, from the International Soil Reference and Information Centre in Wageningen, for having accepted to bring me his opinion and experience as a member of the jury.

Je remercie également ma source de financement pendant quatre ans. Le doctorat s'est déroulé dans le cadre du projet *ARC 09/14-022* (Action de Recherche Concertée), financé par la *Fédération Wallonie-Bruxelles* par l'intermédiaire de *l'Académie Universitaire Louvain*. L'image hyperspectrale a été acquise durant le projet *SOC 3D (SR/10/43-SOC3D)* au sein du *programme STEREO II* de *Belspo* (Politique scientifique fédérale).

Je voudrais souligner ici à quel point j'ai été heureux de travailler au sein de l'unité ENGE. J'y ai apprécié l'atmosphère chaleureuse et animée (vive la réunion cafet) qui y règne et que l'on doit à nos trois professeurs Pierre, Emmanuel et Patrick, bien aidés par le reste de l'équipe. Il me serait impossible de lister tous ceux que j'ai côtoyé, mais malgré les quelques années qui nous éloignent déjà, je garde des souvenirs assez précis de la vie locale. Il serait irrationnel de ne pas commencer par Julien, notre vénéré gardien des traditions et maître du temps, qui n'a pas hésité à partager ses connaissances encyclopédiques pour mon plus grand intérêt. Mais aussi Corentin qui m'a fait découvrir le Bois de Lauzelle "à 5 minutes à pied", les "Balkans" avec François, Damien et Raphäel, Marie-Julie pour son coaching, Thomas et Sarah et les autres du bureau d'à côté d'où les rires communicatifs perçaient souvent nos modestes cloisons. Et bien sûr, comment oublier notre fine équipe du covoiturage avec Céline, (Dr. Ig.) Gaëtan et (encore)

Raphaël, notre virée mémorable dans la neige et ces dizaines d'autres à l'Outremer. Et enfin tous les anciens, et tous les autres "petits nouveaux" qui ne le sont sans doute plus depuis longtemps.

Mon éprouvante gestation doctorale a été entrecoupée d'événements marquants dont notre mariage avec Maëlle ainsi que la naissance de Noé et de Maxence. Votre présence, votre soutien et vos quelques sacrifices m'ont apporté une dose essentielle d'entrain et de courage. Je voudrais également remercier ici pour leur soutien et leur bienveillance les autres personnes qui m'entourent, Catherine et la famille, ma belle famille, ma maman et bien sûr mon papa qui nous a quitté. C'est à toi que je dédie cette thèse.

Table of Contents

1 Introduction **1**

1.1 What are soil organic matter and soil organic carbon? 1

1.1.1 Origin of soil organic matter 1

1.1.2 Variables influencing soil organic matter 2

1.1.3 Importance of soil organic matter 5

1.1.4 Measurement of soil organic matter and soil organic carbon 6

1.2 How much soil organic carbon is present in the soil, where and why? 8

1.2.1 Definition of terms related to the general notion of scale 8

1.2.2 Overview of soil organic carbon stock patterns across spatial extents 10

1.2.3 Elements influencing soil organic carbon in agricultural landscapes 10

1.3 Why are spatial analysis and prediction necessary for the study of soil organic carbon? 12

1.3.1 Assess and predict response of soils to change in climate, land use and land management 12

1.3.2 Build and monitor sequestration policies 13

1.3.3 Improve soil quality and cultivation practices 13

1.3.4 Explain and model soil processes 14

1.4 Which methodologies are used for the spatial prediction of soil organic carbon? 14

1.4.1 Soil monitoring 15

1.4.2 Spectroscopy 16

1.4.3 Mechanistic modelling 16

1.4.4 Digital soil mapping 17

1.4.5 Digital terrain modelling 18

1.4.6 Statistical modelling 18
1.4.7 Machine learning 19
1.4.8 Geostatistics 19
1.5 What are the current limitations for the knowledge and the spatial prediction of soil organic carbon? 21
1.5.1 Quality and quantity of data on soil organic carbon and auxiliary variables 21
1.5.2 Importance of soil organic carbon located below the surface layer 21
1.5.3 Multiple scale variability of soil organic carbon 22
1.6 Objectives 23
1.7 Overview of the following chapters 24
1.7.1 Assessment of the factors controlling vertical and horizontal distribution of soil organic carbon in the Belgian Loess Belt 25
1.7.2 Separating scales of variability to explain spatial distribution of soil organic carbon 25
1.7.3 Detecting and quantifying field-related spatial variation of soil organic carbon 25
1.7.4 Accounting for the spatial scale of variation in total stock quantification 26

2 Assessment of the factors controlling vertical and horizontal distribution of soil organic carbon in the Belgian Loess Belt 27
2.1 Introduction 28
2.2 Material and methods 29
2.2.1 Study area 29
2.2.2 Carbon datasets 30
2.2.3 Terrain attributes 33
2.2.4 Horizons distribution reconstruction 34
2.2.5 Smoothing the profiles 36
2.2.6 Computing class-representative soil organic carbon profiles 36
2.2.7 Spatial prediction of soil organic carbon 37
2.3 Results and discussion 38
2.3.1 Horizon distribution reconstruction 38
2.3.2 Soil property influence on soil organic carbon profile . . 40
2.3.3 Terrain influence on soil organic carbon profiles 44
2.3.4 Analysis of the correlations for legacy and new data . . 46

2.3.5 Spatially explicit prediction of soil organic carbon . . . 49
2.4 Conclusions . . . 52

3 Separating scales of variability to explain spatial distribution of soil organic carbon 53
3.1 Introduction . . . 54
3.2 Material and methods . . . 55
3.2.1 Study area and datasets . . . 55
3.2.2 Overview of filtered kriging . . . 56
3.2.3 Selection of the scales . . . 57
3.2.4 Fitting of the ranges . . . 59
3.2.5 Spatial filtering . . . 60
3.2.6 Correction of a suspected bias in soil organic carbon values 61
3.3 Results and discussion . . . 62
3.3.1 Selection of the ranges . . . 62
3.3.2 Soil organic carbon . . . 62
3.3.3 Texture fractions . . . 69
3.3.4 Links between filtered soil properties . . . 69
3.4 Discussion and conclusion . . . 76

4 Detecting and quantifying field-related spatial variation of soil organic carbon 79
4.1 Introduction . . . 80
4.2 Material and methods . . . 83
4.2.1 Study area . . . 83
4.2.2 Datasets . . . 83
4.2.3 Exploratory analysis of spatial variation . . . 84
4.2.4 Mixed-effect model . . . 84
4.2.5 Estimating the spatial model with restricted maximum likelihood . . . 86
4.2.6 Optimum use of the hyperspectral image . . . 87
4.2.7 Validation of the model . . . 87
4.3 Results . . . 89
4.3.1 Exploratory analysis of spatial variation . . . 89
4.3.2 Model M1 (texture classes) . . . 90
4.3.3 Model M2 (fields) . . . 92
4.4 Discussion . . . 99

4.4.1 Rigorous assessment of texture and fields effect 99
4.4.2 Implications for soil organic carbon modeling and mapping 100
4.4.3 Limitations of hyperspectral images 101
4.5 Conclusions 101

5 Accounting for the spatial scale of variation in total stock quantification 103
5.1 Introduction 104
5.1.1 Global C budget 104
5.1.2 Carbon policies 105
5.1.3 The challenge of total soil organic carbon stock measurement 106
5.1.4 Objective and plan of this chapter 107
5.2 Mathematical formalization 108
5.2.1 Spatial integration of a random field 108
5.2.2 Influence of model error on total stock 109
5.2.3 Spatial prediction of model error using samples 112
5.2.4 Simplification of the predictor in the limit cases 114
5.3 General Discussion 116
5.3.1 Choice of the methodology of prediction 116
5.3.2 Implications for cost-effectiveness 119
5.4 Conclusion 120

6 General conclusion 123
6.1 Main findings 123
6.1.1 Identify and quantify the effect of the factors controlling the spatial variability of SOC concentration and stock . 123
6.1.2 Identify the influence of the spatial scale at which the different factors control the spatial distribution of SOC concentration and stock. 125
6.1.3 Analyze the potential, limitations and pitfalls of the different formats of data used in this study. 126
6.1.4 Determine how the prediction of spatial variation is useful for the estimation of the total SOC stock within a study area. 127
6.2 Perspectives 128

A Multi-scale decomposition and filtering of spatial variables 129

B Method REML-EBLUP 131

Acronyms

AIC Akaike's Information Criterion. 60, 66

DEM Digital Elevation Model. 18, 51

DSM Digital Soil Mapping. 17, 18, 24, 109, 126, 128

EBLUP Empirical Best Linear Unbiased Predictor. 20

ESR Effort Sharing Regulation. 105

EU-ETS European Emissions Trading Scheme. 105

GAM Generalized Additive Model. 19

LMF Levenberg-Marquardt-Fletcher Algorithm. 59, 60

LULM Land Use and Land Management. 5, 10, 13, 25, 119

LULUCF Land Use, Land Use Change and Forestry. 105

MDD Minimum Detectable Difference. 15

REML Restricted Maximum Likelihood. 20, 79, 86, 87, 92, 95, 97, 98, 131

REML-EBLUP Restricted Maximum Likelihood Empirical Best Linear Unbiased Predictor. 20, 86, 87, 95, 101

SMN Soil Monitoring Network. 15, 18

SOC Soil Organic Carbon. 1, 3–6, 8, 10–31, 33, 34, 36–38, 40–54, 56, 57, 59–63, 66–69, 75, 79–85, 87–93, 95–109, 116–121, 123–128

SOM Soil Organic Matter. 1–3, 5, 6, 8, 11, 14–16, 21, 124

TA Terrain Attributes. 18, 25, 33, 37, 38, 44–49, 51, 52, 127

UAV Unmanned Aerial Vehicle. 16

Chapter 1

Introduction

1.1 What are soil organic matter and soil organic carbon?

Soil Organic Matter (SOM) is the organic matter component of soil and plays a critical role at many levels. Soil Organic Carbon (SOC) is the C content of SOM. This section gives a general overview of SOM and SOC.

1.1.1 Origin of soil organic matter

Soil organic matter is composed of living plant, animal, and microbial biomass, dead roots and other plant residues in various stages of decay Hillel and Hatfield (2005). Organic fragments are continuously decomposed by microfauna (like earthworms and arthropods) and microorganisms (like bacteria and fungi) toward smaller molecule size (Figure 1.1). Different processes result in the formation of, on the one hand, carbon dioxide and, on the other hand, mineral components that can be again taken as nutrient by the roots of plants. As it decomposes, the potential of organic molecules to be physically protected increases. This happens through organo-mineral associations by interaction with mineral surfaces (Hassink, 1997) or through physical protection by spatial separation of C from decomposers, typically by encapsulation within aggregates (Six *et al.*, 2002).

Besides this, the concept of humification has been widely used in the scientific community. It refers to the formation of large recalcitrant dark-coloured molecules (called humic substances) from small organic molecules through the combination of microbial activity and catalytic effect at the soil matrix (Stevenson, 1994). However, the absence of evidence supporting the process of humification as well as the existence of these large molecules itself have recently been underlined (Lehmann and Kleber, 2015).

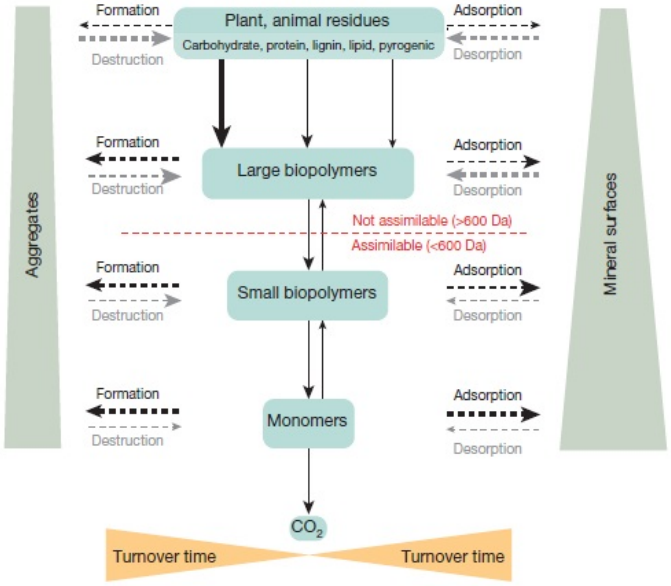


Figure 1.1: Reconciliation of current conceptual models for the fate of organic debris into a consolidated view: the Soil Continuum Model of organic matter cycles and ecosystem controls (Lehmann and Kleber, 2015)

1.1.2 Variables influencing soil organic matter

The concentration of SOM in the soil depends on the input from plant and organisms as well as on the rate at which SOM is mineralized. The input of organic material is largely influenced by land cover. It is typically lower in cultivated areas where plant material is removed by harvesting. The efficiency of SOM mineralization depends simultaneously on (i) SOM biochemical composition, (ii) soil physical texture and structure, (iii) type and abundance of organism, (iv) environmental conditions, with various interactions and feedback effects and (v) time.

SOM biochemical composition The organic pool of soil contains a wide variety of compounds. Some compound referred as labile are considered easily decomposed by microbial communities. In contrast, compounds from primary source (e.g. lignin) or secondary sources are referred as recalcitrant as they are of lower value for decomposers. However, recalcitrance is relative as under suitable conditions, appropriately adapted decomposer organisms have the ability to decompose even presumably persistent materials (Lehmann and Kleber, 2015; van Lützow *et al.*, 2006). Besides, it has been shown that decomposition of presumably recalcitrant materials is favoured by the supply of fresh labile material (Fontaine *et al.*, 2007; Klotzbücher *et al.*, 2011).

Soil texture and structure The content of SOM is also controlled by soil texture and soil structure.

On the one hand, SOM becomes stabilized due to binding to mineral surfaces. Various mechanisms are considered such as ligand exchange, polyvalent cation bridges and weak interactions such as hydrophobic interactions including van der Waals forces and H-bonding (van Lützow *et al.*, 2006)

On the other hand, a hierarchical order of aggregates exists in soil where SOM is the major binding agent. Soil particles assemble in micro-aggregates (20-250 μm) which may themselves assemble in macro-aggregates ($>$ 250 μm). The process of aggregate formation depends on the interaction of multiple factors with a crucial role played by soil mineralogy, root exudates, earthworms activity and microbial communities (Six *et al.*, 2004; Cotrufo *et al.*, 2013). Aggregate formation seem to be closely linked with C storage and stability through physical protection of SOM within micro-aggregates against microbial decomposers (Wang *et al.*, 2014a). The concept of *saturation* assumes that the content of SOC that can be stabilized in soil in steady state is not a linear function of C input but increases up to a limit called "saturation level" (Stewart *et al.*, 2007). Angers *et al.* (2011) showed a positive relationship between the amount of SOC stabilized and the size of the clay fraction together with the fine silt fractions (2-20 μm). The deviation from the level of saturation, also called "saturation deficit", was attributed to an external perturbation from the point of equilibrium, typically caused by land management. Stewart *et al.* (2008) put in evidence mechanisms of saturation related to three different fractions: SOM associated with particles, SOM protected in aggregates and biochemically recalcitrant C.

Current knowledge suggest that, while chemical composition has an influence on the faster decomposition of litter and labile SOM (mainly from amendments and root exudates), long-term (i.e. decades to centuries) decomposition is rather controlled by abiotic mechanisms of protection, i.e. organo-mineral interactions and aggregation (Stockmann *et al.*, 2013).

Type and abundance of organism The type and abundance of organisms is part of an ecology implying the soil (pH, nutrients ...), the plants and the environmental conditions. If multiple behaviors have been documented, synthesizing the links between microbial communities and SOM processing remains a major challenge (Schmidt *et al.*, 2011).

Environmental conditions Environment and more specifically climate controls the activity of micro-organisms through temperature and precipitations. The speed of chemical and biochemical reactions characterizing SOM decomposition is positively controlled by temperature with an exponential signature (Davidson *et al.*, 2006). However, the activity of organisms also depends on the availability of both water and oxygen. Due to the fact that above a certain level the presence of water may prevent soil oxygenation, decomposition is maximal inside a specific range of soil moisture and decreases as we move away (Falloon *et al.*, 2011). Specific SOC processes related to specific soils that are

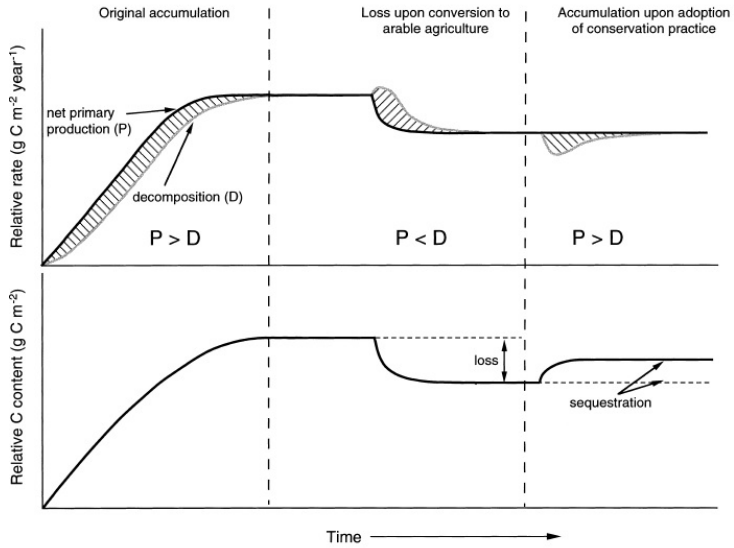


Figure 1.2: Conceptual model of SOC dynamics in a prairie agroecosystem, from Janzen *et al.* (1998). During soil formation, net primary production (P) exceeds decomposition (D), resulting in the accumulation of soil C until P and D again converge. Upon conversion of the land to arable agriculture, D initially exceeds P, resulting in the loss of soil carbon until a new steady state is approached. Adoption of C-retentive cropping practices reduces D relative to P, resulting in a gain of C until D and P again converge.

permanently dry, frozen or water-saturated (Schmidt *et al.*, 2011) will not be covered here.

Time Change in SOC is potentially a slow process.

Given enough time and stable environmental conditions, agricultural and native ecosystems evolve to a stable state in which the biota maximizes homeostasis with the physical environment (Odum, 1969). However, perturbations strong enough not to be dampened by homeostatic controls cause ecosystems to evolve towards a new steady state (Kimble *et al.*, 2000). Therefore, the content of organic carbon in a soil will evolve during the period following a perturbation such as a change of land use. This is for example represented in Figure 1.2 for a conversion from grassland to cropland. According to the type of perturbation, this transition process may be relatively long at human scale. Effect of land use history on SOC content have been reported to last for decades, centuries or even longer (Foster *et al.*, 2003; Stevens and van Wesemael, 2008). In a meta-analysis of field experiments introducing the use of cover crops in rotations, Poeplau and Don (2015) predicted that a new steady state would be reached after 155 years of cover crop cultivation. A slow rate of change implies that current distribution of SOC is not systematically representative of the current drivers.

Historic Land Use and Land Management (LULM) should also be accounted for when explaining and predicting SOC dynamics in agro-ecosystems (Luo *et al.*, 2011).

1.1.3 Importance of soil organic matter

SOM plays an important role in multiple soil functions.

Soil organic matter modifies the physical micro-structure of soils by contributing to the formation of aggregates, therefore increasing the porosity. This is related to larger water infiltration, smaller run-off of water from the surface and larger aeration, rootability and workability (Dexter, 2003). Besides, the structuring effect of SOM increases soil stability and reduces water erosion (Loveland and Webb, 2003). Some substances composing SOM can create chemical bindings with molecules of organic and inorganic pollutants and therefore limit their availability (Scheunert *et al.*, 1992). Finally, SOM has the potential to supply substantial quantities of nutrients (i.e. nitrogen (N), phosphorus (P) and sulphur (S)) for plant uptake (Sarker *et al.*, 2018).

Therefore, SOM is associated with a constellation of positive effects, making it to be considered as an overall indicator of soil quality and soil health (Dexter, 2003). The effects of SOM on water infiltration and retention, on soil stability and against erosion reduces the incidence and severity of hazards such as dryness, flooding and mudflow. Soil aeration, rootability and availability of nutrients has positive effects on soil fertility and soil biodiversity.

There is considerable concern that, if SOM concentrations in soils are allowed to decrease too much, then the productive capacity of agriculture will be compromised by deterioration in soil physical properties and by impairment of soil nutrient cycling mechanisms (Loveland and Webb, 2003). However, there is little evidence of a fixed thresholds of SOM as the effect of SOM content on soil quality depends on other elements such as soil type and composition of SOM (Loveland and Webb, 2003; Johannes *et al.*, 2017).

Besides its effects on soil quality, SOM is also seen as a means to mitigate climate change.

The soil is the largest terrestrial pool of organic carbon with an estimate of 1325 Pg SOC in the upper 1 m and just under 3000 Pg when estimates for deeper soil layers are included (Köchy *et al.*, 2015). Therefore, SOM contains more than three times as much carbon as either the atmosphere or terrestrial vegetation. Each year, due to the mineralization of fresh and older SOM, around 59 Pg C are transferred from the soil reservoir to the atmosphere, mainly in the form of methane or carbon dioxide. We may conclude that even a relatively small increase of the SOC stock could partially compensate for the human induced emissions from combustion of fossil fuels and cement industry (7.8 Pg C yr^{-1} Stocker *et al.*, 2013).

Land use change and adaptation of agricultural management practices have been identified as suitable and potentially cost-effective ways to increase SOC stock and mitigate greenhouse gases emissions (Lal, 2004). In long-term field

experiments, the conversion from arable land into grassland showed the highest SOC sequestration rates, ranging between 0.4 and 0.8 t C ha^{-1} yr^{-1}. Typically, the use of cover crop, ley in rotation or the combination of reduced tillage with 100% straw incorporation showed rates around 0.1 t C ha^{-1} yr^{-1} (Lugato *et al.*, 2014). In a review paper, Smith (2012) reports a theoretical carbon sequestration potential using a combination of agricultural practices of 200 Mt CO_2 eq. yr^{-1} but an economically realistic potential of 20 Mt CO_2 eq. yr^{-1} (1 Pg = 10^{15} g = 1000 Mt). For this economically realistic potential, the potential of mitigation practices across Europe was calculated with the MITERRA-Europe simulation model. More recently, (Paustian *et al.*, 2016) summarized cropland greenhouse gases mitigation practices and potentials within a decision tree shown in Figure 1.3.

1.1.4 Measurement of soil organic matter and soil organic carbon

SOM cannot be analyzed precisely as the components are manifold, poorly defined and form complex associations with the soil mineral matrix. However, the alkaline extraction technique is the classical method used to separate SOM in categories according to its chemical behaviour. An alkaline solution is added to the soil sample and causes the dissolution of approx. 50% - 70 % of the organic material. After an acid is added to the solubilized organic materials, a dark solid precipitates. The definition of the different fractions (not dissolved, dissolved and precipitated) and the interpretation of how they link to SOM decomposition processes varies between authors (Lehmann and Kleber, 2015).

In many circumstances, only the total content of SOM or SOC is needed and not the different chemical fractions. Various methods exist in this context but the most popular are dry combustion, Walkley and Black wet oxidation and loss-on-ignition.

The method of dry combustion is considered to be the most accurate one and is largely used as a reference. It consists in heating the soil ($\sim$ 950 °C) for burning off all C present. Released CO_2 is analyzed using thermal conductivity detection or IR spectroscopy using a CN Analyzer. The disadvantage of the method is that it requires the drying of the sample and/or a pre-treatment to remove inorganic C (carbonates). This may result in the loss of volatile organic compounds or the decomposition and loss of other organic compounds present in the sample (Schumacher, 2002). The method of Walkley and Black (Walkley and Black, 1934) is also referred to as chemical oxidation or wet oxidation. Oxidisable matter in the soil is oxidised by a potassium dichromate solution. Heat is generated when two volumes of sulfuric acid are mixed with one volume of the dichromate. The remaining dichromate is titrated with ferrous sulphate (Department of Sustainable Natural Resources, Unknown year). Walkley and Black (1934) claim a recovery rate of 77 % of total C but these rates depend on soil properties (Wang *et al.*, 2012; Meersmans *et al.*, 2009b). With the method of loss-on-ignition, the soil sample is first dried then heated to 400 °C during 3 hours and cooled. The mass concentration of SOM is then calculated as the

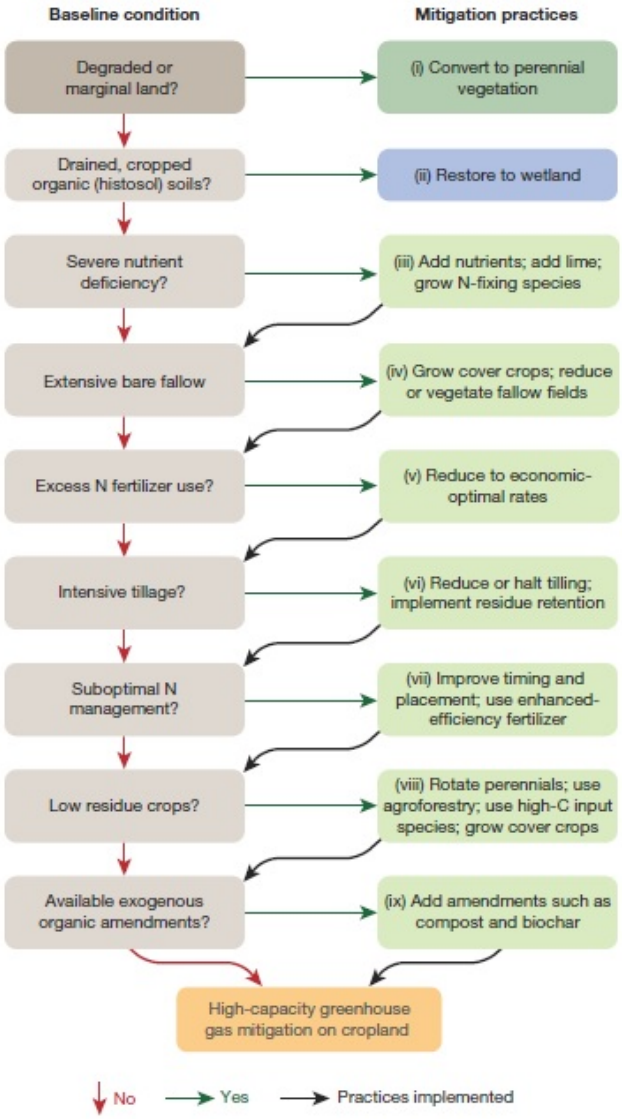


Figure 1.3: Decision tree for cropland greenhouse gases mitigating practices (rice not included) from Paustian *et al.* (2016). For degraded, marginal lands, the most productive mitigation option is conversion to perennial vegetation either left unmanaged or sustainably harvested to offset fossil energy use (cellulosic biofuels). For more arable lands, multiple options could be implemented sequentially or in combination, depending on management objectives, cost and other constraints. Practices are roughly ordered from lower-cost or higher-feasibility options towards more costly interventions (bottom of figure).

relative difference between the weight of the sample just after drying and the weight of the sample after ignition (Florence and Mengel, 2013).

The content of SOC in a soil sample is generally given by its mass concentration expressed in g C kg^{-1} or in %. A factor of 1.724 is generally used in order to convert SOC to SOM or inversely, with the relationship

$$SOM = 1.724 \times SOC \tag{1.1}$$

It is based on the assumption that SOM is made of 58% of SOC. However, the use of this single factor is challenged as, on the one hand, the exact proportion of SOC in SOM depends on soil type and soil depth and, on the other hand, its value seems overall underestimated (Pribyl, 2010).

Typical mean topsoil SOC concentrations for the different "Global Ecoregions" are comprised between 1% and 5% (Stockmann *et al.*, 2015).

The stock of SOC is routinely defined as the mass of SOC by unit of surface. It can be given either for a specified depth layer (e.g. 0-1 m), or for the whole soil profile up to the parent material. It is generally expressed in kg C m^{-2}, Mg C ha^{-1} or t C ha^{-1}, with 1 kg C m^{-2} = 10 Mg C ha^{-1} = 10 t C ha^{-1}.

For a fixed soil depth layer, the corresponding SOC stock is generally estimated from soil sampling using the expression of eq. (1.2)

$$SOC_{stock} = C \times T \times (1 - RM) \times BD \tag{1.2}$$

where C is the SOC concentration, T is the thickness of the depth layer, RM is the volumetric fraction of rock measured and BD is the soil bulk density. A specific application of this expression is for example visible in Goidts and van Wesemael (2007). Bulk density may vary widely according to factors such as soil type (Batjes, 1996) and cultivation intensity. For this reason, Ellert and Bettany (1995) introduced the concept of equivalent soil mass which proposes to quantify and compare SOC stocks based on a fixed soil mass rather than on a fixed soil depth.

1.2 How much soil organic carbon is present in the soil, where and why?

As this thesis deals with spatial prediction of SOC, we believe it is important to expose the natural scales of variation of SOC across the globe and the underlying factors. This is the scope of this section.

1.2.1 Definition of terms related to the general notion of scale

.

The term of scale has multiple, sometimes contradictory meanings and is often used is replacement of more specific terms. At this point, we consider useful to

recall the definition of some specific terms related to the general notion of scale so that these terms will be used without ambiguity afterwards. This section is largely inspired by Dungan *et al.* (2002).

In order to reduce potential confusion over myriad definitions, we distinguish three different categories to which spatial scale related terms may be applied:

Phenomenon The phenomenon being studied.

Observations The spatial units or sampling units used to acquire information about the phenomenon.

Analysis The analysis of the data, used to summarize them or make inferences.

The phenomenon, sampling and analysis categories can be thought of as three dimensions to which scale concepts pertain. A selection of specific terms related to the notion of scale is presented below.

Extent In the general context of statistical analysis, the extent is the total length area or volume that exists or is observed or analyzed. The term extent can thus be applied to the phenomenon, observations or analysis axis of scale. Within this thesis, extent will refer to the surface of the geographical area under study.

Support The support is a geostatistical concept related to the analysis axis of scale, in the sense that it is a property of a variable used to model data. It refers to an n-dimensional volume, including its geometrical shape, size and orientation, within which average values of a variable may be computed. The support can be as small as a point or as large as the full extent of the study area over which the variable is represented by a spatial field. A notable insight of geostatistics is that a variable measured on one support is a different variable than the same variable measured on a different support.

Resolution Resolution is a term that naturally applies to observations and analysis rather than to phenomenon. It is defined as the smallest object that can be reliably detected. The resolution of a data set is often defined as the size of the smallest sampling unit of its sampling design. The term resolution is commonly used in remote sensing.

Grid spacing The grid spacing is the distance between the nodes of a regular grid (the case of irregular grid is not addressed here). The general term of grid may refer to a sampling grid as well as a grid of analysis. For a sampling grid, the resolution cannot be inferior to the grid spacing. In contrast, for a grid of analysis, the grid spacing does not determine the support size. For a given grid, the support could for example be the point but could also be larger than the grid spacing.

1.2.2 Overview of soil organic carbon stock patterns across spatial extents

The most dominant trends of SOC variability over the globe are associated with climate and soil type. Variation in SOC is largely related to variation of latitude because of climatic control (Minasny, 2013). Besides, in a global inventory, Batjes (1996) reported large differences of SOC content between FAO–UNESCO soil units, whether the comparison was done for the whole profile or for a given depth interval.

The largest SOC stocks are found in wetlands and peatlands which are mainly found in regions of permafrost and in the tropics (with orders of values of 40-50 kg C km^{-2} in the 0-1 m layer) because decomposition rates in wetlands and permafrost are limited by low availability of oxygen and low temperatures, respectively (Köchy *et al.*, 2015). Deserts and mountainous areas present the smallest stocks with typical values between 1 and 10 kg C m^{-2}.

When focused on a country or a region, variables related to soil type such as soil taxon, drainage and texture are still generally the main explaining factors of SOC variability. However, influence of large patterns of land cover and land use can be consequent too (Meersmans *et al.*, 2009a; Tan *et al.*, 2004).

At field to landscape extent, additional processes of very diverse nature emerge as dominant controls (Viaud *et al.*, 2010a). Some are related to human action such as LULM. Others are related to the way topography influences the distribution of water and sediments in the landscape. See section 1.2.3 for more details on factors influencing SOC in agricultural landscapes.

Spatial variability may be relatively high even at a range of a few meters. At the extent of the micro-site (defined by 5 samples regularly spread within a circle of 4 m radius), Goidts (2009) measured coefficients of variation for SOC concentration (i.e. the ratio of the standard deviation and the mean value) of 8 % (cropland) and 15 % (grassland). This short-range variability of SOC results more generally from the short-range variability of soil properties. Lin *et al.* (2005) used hierarchical sampling to assess the variability of different soil properties (A-horizon thickness, depth to calcium carbonate and surface soil pH values) and observed that the majority of the variability ($>$ 50 %) was already present within a 12 m distance. Variability of SOC at this scale may be the result of short range heterogeneity of soil processes such as water dispersion or accumulation due to micro-topography, non-uniform distribution of fresh matter inputs like plant residues or livestock manure and the intrinsic randomness of soil and environmental processes.

1.2.3 Elements influencing soil organic carbon in agricultural landscapes

At the landscape extent, SOC spatial variability is strongly affected by natural and anthropogenic processes and linear anthropogenic elements such as hedges or ditches (Lacoste *et al.*, 2014). Conversion of native land to cropland generally causes significant decrease of SOC (VandenBygaart *et al.*, 2003b). Many

common agricultural practices like ploughing or disc-tillage accelerate the decomposition of soil organic matter and leave the soil susceptible to wind and water erosion. The stabilized SOC fractions are particularly subject to fast decline in the case of an increase in cultivation intensity (Six *et al.*, 2002). The mechanism involves breaking of aggregates through disruption and exposition of formerly encapsulated SOM (Abdollahi *et al.*, 2014; Malamoud *et al.*, 2009). Inversely, the selection of crop rotations and the application of fertilizers or organic amendments tend to raise SOC stock by enhancing crop growth and/or increasing C inputs (VandenBygaart *et al.*, 2003b).

Topography influences the distribution of water, energy and sediments within the landscape and this has an impact on SOC stock. Translocation of sediments plays a key role in biogeochemical cycle of C through processes occurring at the sites of erosion, deposition and during transport (Berhe *et al.*, 2008). At detachment site, some of the eroded C is dynamically replaced by new C inputs, mainly by production of new photosynthate (Harden *et al.*, 1999). However, severely eroded soils lose their nutrients and water retention capacity, affecting plant growth and the return of biomass (Lal and Pimentel, 2008). During transport, mineralization rate is increased because of aggregate breakdown (Wang *et al.*, 2014b). At the footslopes, deep soils are formed with partial deposition of material that eroded from upslope positions. In these landform positions, enhanced rates of plant productivity along with physical protection mechanisms of SOM through organo-mineral interactions and formation of aggregates is likely to result in higher stocks of SOM (Berhe and Kleber, 2013).

There is an interaction between natural and anthropogenic processes as sediment translocation is dramatically amplified by agricultural practices (Kirkels *et al.*, 2014). In particular, since the mechanization of agriculture, tillage operations are responsible for the major part of within field soil redistribution in intensively cropped systems (Govers *et al.*, 1996; Van Oost *et al.*, 2005; Martinez *et al.*, 2010; De Gryze *et al.*, 2008). A large part of translocated sediments is retained and buried along field boundaries, in hillslopes or in floodplains (Chartin *et al.*, 2013; Hoffmann *et al.*, 2013). Tillage erosion interacts with water erosion which is largely dependent on agricultural patchiness and non-agricultural landscape features (such as hedges, ditches, dams, grass strips, etc.), since these affect hydraulic connectivity (Fiener *et al.*, 2011).

In a catchment, the contribution of erosion to additional SOC sequestration depends on maintaining plant productivity and on the fate of translocated sediments (deposition inside the catchment or exportation to the river network Trumbore and Czimczik, 2008). Globally, Van Oost *et al.* (2007) estimated that erosion in the agricultural landscapes results in a limited carbon sink of approximately 0.12 Pg C yr-1 while Harden *et al.* (2008) state that this sink could be slightly underestimated. Besides, the ignorance of erosion may lead to bad parametrization of SOC dynamics models and limit their predictive capacity (Sanderman and Chappell, 2013). Anyway, considering the large amounts of SOC-rich sediments that may be buried at depositional sites and their generally low sensibility to mineralization, many studies have underlined the importance

of accounting explicitly for SOC buried at greater depths in SOC assessment (Meersmans *et al.*, 2009a; Doetterl *et al.*, 2013a; Harrison *et al.*, 2011).

1.3 Why are spatial analysis and prediction necessary for the study of soil organic carbon?

The previous sections proposed an overview of the functioning and natural scales of variability of SOC. Here we want to describe the many contexts and applications where a better knowledge of the spatial distribution of SOC and other soil and environmental properties is necessary.

1.3.1 Assess and predict response of soils to change in climate, land use and land management

Considering its high sensitivity to temperature and wetness (cf. Section 1.1.2), the SOC pool is expected to react to current and future climate change. However, due to the slow evolution of SOC content and the difficulty to measure precisely stocks over large areas, the future response of SOC is highly uncertain (Barraclough *et al.*, 2015; Trumbore and Czimczik, 2008).

Different regional and national inventories of SOC across Europe have resulted in contrasting conclusions. In the UK, Bellamy *et al.* (2005) reported large declines in soil carbon irrespective of land use between the periods 1978–1983 and 1995–2003 and suggested a link to climate change. More recent studies concluded however that, in agricultural soils, at most 10–20% of these observed changes could be attributed to climate change and that the rest was an effect of reduced grazing in grassland and movement to new equilibrium resulting from past land use and management change in cropland (Smith *et al.*, 2007; Kirk and Bellamy, 2010; Barraclough *et al.*, 2015).

Spatial analysis and prediction of SOC can be helpful to assess response of soils to change in climate, land use and land management. Spatial analysis can be useful to assess the relative influence of different factors on SOC evolution. Using a SOC inventory of 1960 resampled in 2006 and regional datasets of land used and management, van Wesemael *et al.* (2010b) identified a negative effect of drainage in grassland and a positive effect of conversion from cropland to grassland. Predicting the effect of change on the evolution of SOC stocks can be done by means of process-based models (see section 1.3.4). These are calibrated and run on spatially explicit data layers of soil and environmental factors. Therefore, the quality and performance of these models will depend on the availability and precision of these datasets . Improving the methods that create these datasets is thus of primary importance.

1.3.2 Build and monitor sequestration policies

Sequestration of atmospheric C in soil is regarded as a way to compensate for anthropic emissions of greenhouse gases and to mitigate climate change (cf Section 1.1.3). What is referred to as the "4 per mille" initiative was launched during the COP21 in December 2015 in Paris. It states that annual sequestration of 0.4% (4 per 1000) of the SOC stocks until 2 meters depths would – and should – exactly compensate for human-induced emissions from fossil C. Signatories (countries, regions, international agencies, private sectors and NGOs) commit in a voluntary action plan to implement farming practices that maintain or enhance soil carbon stocks in agricultural soils and to preserve carbon-rich soils (Chambers *et al.*, 2016; Lal, 2016).

As for estimating the effect of climate change, methodologies for SOC spatial estimation are necessary in the frame of SOC sequestration. Changes in SOC stocks are now taken into account in international negotiations regarding SOC sequestration and climate change (Martin *et al.*, 2011). A right and efficient economic framework where a monetary value can be attributed to SOC increase is the key to enhancing soil C sequestration (White *et al.*, 2018). However, a proper working and broader acceptance of a global SOC market would require a thorough measuring of SOC stock and change (Alexander *et al.*, 2015). Accurate and flexible methodologies are thus needed to strictly assess the initial SOC stock and dynamic condition to serve as a baseline for future SOC evolution, and monitor future change of the stock, with a particular importance of the national scale (Martin *et al.*, 2014).

Mapping of SOC and soil properties can also be used to identify areas with the largest potential for sequestration. Typically, this could be done by mapping soil saturation deficit (cf Section 1.1.2 and Angers *et al.* (2011)) as soils that are the most below their point of equilibrium in SOC content (i.e. high saturation deficit) are the ones that could the most increase their SOC content under appropriate management (Stewart *et al.*, 2009).

1.3.3 Improve soil quality and cultivation practices

Soil management may also exploit the beneficial properties of SOC in order to maintain soil stability and improve soil quality. In this respect, mapping soil carbon at landscape scale is needed as it allows isolating the effect of different LULM. For example, Meersmans *et al.* (2012a) mapped SOC in France at a resolution of 250 m, a resolution fine enough to exploit LULM data. The resulting map highlighted specific areas of particularly low or high SOC content related to LULM. This map can thus be seen as a tool for developing appropriate soil management practices.

Precision agriculture is another application that requires to map SOC and other soil properties. One of its most important aspects, referred as "site-specific crop management" is defined as "matching resource application and agronomic practices with soil and crop requirements as they vary in space and time within a field" (Whelan and McBratney, 2000). For example, variables

such as SOC and content of nutrients can be mapped at high spatial resolution within the field to optimize manure application with the aim of maximizing crop growth while limiting costs.

1.3.4 Explain and model soil processes

Mechanistic models are used to better understand SOC processes and predict SOC evolution under different conditions. Some model applications exploit and sometimes even require spatially explicit knowledge of soil and environmental variables.

Mechanistic models of SOM turnover have been widely used. Major ones, like Roth-C (Coleman and Jenkinson, 1999) and Century (Parton, 1996), partition SOM between interconnected pools of contrasting decomposition rates. They allow accounting for the effects of soil type, temperature, moisture content, plant cover or manure application. They have been often calibrated and evaluated on data from long term agricultural experiments (Falloon and Smith, 2002; Smith *et al.*, 1997), typically with the aim to better assess the effect of agricultural management practices (Powlson *et al.*, 2012; Poeplau and Don, 2015). Multiple pools models have served as basis of more complex models including for example transport of water and gases in the soil (Herbst *et al.*, 2008) or effect of soil structure (Malamoud *et al.*, 2009). However, they are now being criticized for their shortcomings and failures in the mechanistic representation of decomposition processes (Lehmann and Kleber, 2015).

Spatial application of SOM turnover model requires spatially explicit knowledge of soil and/or environmental variables. These models were linked to spatial databases of soil and terrain variables to give spatially explicit results of SOC at national or regional scale (Milne *et al.*, 2007). They were also coupled to soil redistribution models and were used to assess the impact of erosion on the carbon cycle at landscape or catchment scale (Lacoste *et al.*, 2015; Nadeu *et al.*, 2015).

Environmental models of SOC may also be based on the statistical relationship between SOC observations and an ensemble of spatially available soil and environmental variables (Gessler *et al.*, 1995). This will be discussed in Chapter 1.4.

1.4 Which methodologies are used for the spatial prediction of soil organic carbon?

Spatial prediction of SOC is the result of multiple steps from the acquisition of spatial information, the building of a model and the use of this model for predicting. At any step, there are a wide variety of methodologies that can be applied alone or in combination. Here, we propose to describe the most popular methodologies that have been used last decades.

1.4.1 Soil monitoring

Soil data may be collected during single measurement campaigns that are typically carried on for the production of soil maps and reference inventories. Soil data may also be collected on a regular basis within a Soil Monitoring Network (SMN). These are used to estimate directly SOC content and change over a given area of scale varying from the field to the region or even the world (Post *et al.*, 2001). They are composed of an ensemble of locations where properties related to soil are repeatedly measured with the same procedures in order to form a consistent database. SOC content is almost systematically measured in SMNs due to its interactions with other soil properties and its large role on the global carbon cycle. Other C components like specific fractions or above-ground CO2 fluxes may be monitored too (Smith *et al.*, 2012).

The sensibility of SMNs to changes of SOC stock with time is critical. It is generally expressed as the Minimum Detectable Difference (MDD), i.e. the minimum difference of SOC stock necessary to assess a significant change, considering both the uncertainty in SOC measurements and intrinsic spatial and temporal variability. In European countries, considering the power of national SMNs and the typical temporal trends in SOC content, the time needed to observe the MDD is generally of about ten years but sometimes a lot more (Saby *et al.*, 2008; Schrumpf *et al.*, 2011).

The sampling design is critical when establishing new SMNs (Arrouays *et al.*, 2012). There is an essential distinction between design-based and model-based sampling schemes (Brus and De Gruijter, 1997). Model-based approach is intended for the calibration and exploitation of a spatial model of variation within the frame of geostatistics. It is particularly suitable to make point-prediction and quantify uncertainty on the prediction (Arrouays *et al.*, 2012; Brus and De Gruijter, 1997). A description of geostatistics is given in Section 1.4.8. Design-based approach is based on probability sampling, which means that inferences from the data are based upon the probability that particular locations are included in the design. Stratified sampling is a probability sampling approach which is very popular with SMNs. The area is divided into different strata according to other soil and environmental variables that are known to have an influence on SOC. Climate, soil type, land use or land management datasets are generally used to define the strata (van Wesemael *et al.*, 2010a), but even a first estimate of SOC and associated uncertainty may be used to define a stratified sampling scheme for SOC assessment (de Gruijter *et al.*, 2016). Stratification, a crucial element of the design-based approach is recognized as a suited practice to maximize the precision on the total stock over the whole monitored area. As soon as environmental factors that affect SOM dynamics are identified and stratified, inappropriate spatial averaging can be minimized (Post *et al.*, 2001). Stratification is also used to highlight the effect of specific controlling factors (e.g. Goidts and van Wesemael (2007); Sleutel *et al.* (2007)).

1.4.2 Spectroscopy

Hyperspectral Vis-NIR reflectance spectroscopy is currently receiving large attention. Associated with multivariate calibration methods, models can produce direct and unbiased estimates of SOC content. Measurements may be taken in laboratory, directly in the field or by Unmanned Aerial Vehicle (UAV), plane and satellite.

Precision tends to decrease across these modes as, when the altitude increases, the resolution becomes coarser and the perturbing effect of the atmosphere becomes more pregnant (Muñoz and Kravchenko, 2011; Gomez *et al.*, 2015; Tekin *et al.*, 2012). A typical domain of application of field and UAV spectroscopy is precision agriculture (Bilgili *et al.*, 2010; Zhang and Kovacs, 2012).

If remote sensing has mainly been used to map soil properties at local scale (Mulder *et al.*, 2011; Uno *et al.*, 2005), it is increasingly seen as a tool to be used at large scale, i.e. for national soil assessment (Lagacherie and Gomez, 2014; Stevens *et al.*, 2013). Airborne and satellite spectral imaging have the potential to cover large extents with fair spatial resolution. Besides, the absence of spatial interpolation eliminates a common source of spatial dependence in the error and allows the detection of small-scale spatial variability of SOC content (Hbirkou *et al.*, 2012a; Selige *et al.*, 2006). For some soil properties, Gomez *et al.* (2012) obtained correct representations of the spatial structures with a moderate underestimation of the semi-variances that remained constant across distances.

The weakness of Vis-NIR spectroscopy is that models that are calibrated over a given domain of soil properties lead to poor prediction when they are extrapolated to different soil conditions (Stevens *et al.*, 2012). To overcome this, large and diverse soil spectral libraries were assembled and new versatile calibration methods were developed and their performance assessed (Ramirez-Lopez *et al.*, 2013; Nocita *et al.*, 2014; Shi *et al.*, 2015; Stevens *et al.*, 2013).

Recent advances using proximal sensing in laboratory (which represents much more favorable conditions than with remote sensing) were able to produce unbiased predictions with R^2 coefficients up to 0.84 for mineral cropland soils (Nocita *et al.*, 2014). Better accounting of perturbation effects from soil wetness and roughness is promising in the perspective of obtaining good precision in field condition (Nocita *et al.*, 2013).

1.4.3 Mechanistic modelling

Mechanistic modelling can provide spatial prediction of SOC. As already mentioned in Section 1.3.4, mechanistic models of SOM turnover can be applied on a spatial basis, i.e. independently on each spatial unit of a given region. These models can also be coupled to landscape evolution models, i.e. models focused on the change of topography caused by natural and anthropogenic drivers.

These coupled models reproduce the effect of water- and sometimes tillage-induced soil redistribution on the spatial redistribution of C. At the extent of a small catchment, they have been able to reproduce the spatial heterogeneity of measured SOC stock in fields with different land use types (e.g. Rosenbloom *et al.*, 2006; Van Oost *et al.*, 2005; Dlugoß *et al.*, 2012; Yoo *et al.*, 2006). With some adaptations, such models have also been applied at regional (Nadeu *et al.*, 2015) and global (Van Oost *et al.*, 2007) scales.

1.4.4 Digital soil mapping

Digital Soil Mapping (DSM) can be 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 observation and knowledge and from related environmental variables" (Lagacherie and McBratney, 2006). The key idea of DSM is to exploit, for the prediction of a hardly observable variable, related environmental covariates that are easier to measure, therefore reducing field and laboratory work and associated costs (McBratney *et al.*, 2003).

The mnemonic *scorpan* is used to describe the empirical relationships between soil and other spatially referenced factors in DSM. It is an adaptation of the *clorpt* mechanistic model of Jenny (1994) which identifies five factors (climate, organisms, relief, parent material and time) to explain soil formation. It can be defined by eq. (1.3)

$$S = f(s, c, o, r, p, a, n) \tag{1.3}$$

where

- S = soil classes or attributes (to be modeled)
- s = soil, other or previously measured properties of the soil at a point
- c = climate, climatic properties of the environment at a point
- o = organisms, including land cover and natural vegetation or fauna or human activity
- r = relief, topography, landscape attributes
- p = parent material, lithology
- a = age, the time factor
- n = spatial or geographic position

It is assumed that S and each of the seven factors are implicit function of the space coordinates x, y and an approximate time $\sim t$ (except the factor n which represents the coordinates x, y). Processes implying soil properties (texture), climate, microorganisms, human activity, relief and time were described in section 1.2.3. $f()$ is some form of empirical quantitative function linking S to the *scorpan* factors. It can be typically any function from the framework of digital terrain modelling, statistical modelling, machine learning or geostatistics. More detail is given in the following sections.

1.4.5 Digital terrain modelling

Representation of terrain elevation at regular intervals, or Digital Elevation Model (DEM), has been used for decades in hydrological, geomorphological and biological applications (Moore *et al.*, 1991). Digital terrain modelling, also referred to as geomorphometry (Pike *et al.*, 2009), stands for the exploitation of topographic information within DSM through the computation of Terrain Attributes (TA) (or, equivalently, topographic indices). Therefore, it is related to the *scorpan* factor r of the more general framework of DSM.

The value of TA on a grid node are generally based on the value of elevation of the corresponding and neighboring grid nodes. This is for example the case for slope, curvature or relative elevation indices. Sometimes TA result from more complex rules and algorithms. For instance, the topographic wetness index (Beven and Kirkby, 1979) implies the calculation of upslope contributing area and flow accumulation by an algorithm of water flow. Florinsky (2016) proposed a comprehensive review of TA.

Because of the large topographic control on SOC (cf Section 1.1.2, various TA have proved their interest as predictor at landscape scale. As an example, models using three to five TA were able to explain up to 71% of SOC variability in a forested watershed (Thompson and Kolka, 2005).

Topography may also define spatial units such as landform elements (MacMillan *et al.*, 2004) which can be used as a support to map SOC. Florinsky *et al.* (2002) delineated zones of accumulation, transit and dissipation based on the value of horizontal and vertical curvature and observed that SOC content was on average 40% larger in accumulation zones compared to dissipation zone.

1.4.6 Statistical modelling

Linear statistical modelling aims at identifying the best empirical relationships between SOC and soil and environmental covariates. The spatial distribution of SOC over study area is then predicted based on these relationships. Modelling approaches differ according to the type of model and the nature and spatial scale of covariates. These covariates may be continuous (i.e. temperature) or categorical (i.e. soil profile development class). They may be coded on a regular grid (i.e. elevation raster) or as spatial polygons (like many soil maps).

One of the simplest method consists in averaging SOC observations over spatial units. These can typically be based on soil type (Yu *et al.*, 2007) or soil type and land use (Arrouays *et al.*, 2001; Lettens *et al.*, 2007). This approach is traditionally combined with design-based sampling within SMNs (cf Section 1.4.1).

Multivariate regression modelling is another approach. Meersmans *et al.* (2008) performed multiple linear regressions of SOC on continuous variables related to soil texture and drainage and derived from a soil map. Yang *et al.* (2008) used a generalized linear model to predict SOC distribution from mean

annual temperature, soil texture and moisture in the Tibetan grasslands. Suuster *et al.* (2012) implemented a mixed-effect model to account for systematic effects associated with hierarchical sampling design.

Generalized Additive Model (GAM) incorporates the notion of proximity in regression. It is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables. The parameters of these smooth functions are inferred within the regression process (Hastie, 2017). GAMs have been used in many cases for SOC prediction. For example, de Brogniez *et al.* (2015) mapped topsoil SOC over Europe at a resolution of 500 m with a R^2 of prediction of 0.27 while Chartin *et al.* (2017) predicted the spatial distribution of SOC over Walloon Region on a 40 m grid with a R^2 of prediction of 0.64.

1.4.7 Machine learning

Machine learning is defined as a sub-field of computer science and artificial intelligence which deals with building systems that can learn from data instead of explicitly programmed instructions. It covers a large panel of methods that are typically used for data mining, feature selection or predictive modelling. Feature selection algorithms allow automatically selecting the best set of covariates. They include wrapper methods such as recursive feature elimination and embedded algorithms of penalized regression such as ridge regression, elastic net or LASSO.

Prediction methods generally used within the frame of machine learning allow identifying and exploiting complex nonlinear relationships and interactions. They are essential to deal with the rapidly growing amount of soil and environmental datasets of various format and quality. Popular methods of prediction that have shown satisfying results with SOC are typically random forest regression (Grimm *et al.*, 2008; Suuster *et al.*, 2012), boosted regression trees (Martin *et al.*, 2014), neural networks (Ayoubi and Karchegani, 2012) or the *Cubist* piece-wise regression (Lacoste *et al.*, 2016).

1.4.8 Geostatistics

Geostatistics is an ensemble of techniques based on spatial statistics and allowing data exploration and spatial interpolation. The central idea is to define variables as spatial random fields. In our case, a soil variable like C is defined at each point as a random variable explicitly depending on the location. Geostatistics is suited for localized predictions and maps and for quantifying the uncertainty associated with these maps (Arrouays *et al.*, 2012).

The empirical variogram is the dedicated tool for data exploration in geostatistics. It is a representation of the variance of the difference between the variable at all pairs of locations separated by a given distance. The empirical variogram is often approximated by a model function (Chilès and Delfiner, 1999). Interpolation methods rely on the value of the soil variable at sampling

points and on the variogram model to estimate the variable at locations where it is unknown. The most common interpolation method is kriging (Matheron, 1963) which is also recognized as the best linear unbiased predictor. Simple and ordinary kriging are used if the mean is considered stationary over the domain while universal kriging allows accounting for a spatial trend.

In addition to simple interpolation, kriging is widely used within methodologies exploiting covariates as extra source of information. On the one hand, a kriging approach may be performed following a distinct prediction method. Kriging has for example been applied on the residuals of machine learning prediction methods (Motaghian and Mohammadi, 2011). Regression-kriging combines a statistical regression models with kriging of the regression residuals (Odeh *et al.*, 1995). It is however a sub-optimal method that has been used extensively because it is easy and has been shown empirically to perform as well as other methods.

On the other hand, the covariates may be explicitly considered directly within a hybrid geostatistical method. Variants like kriging with varying local means and kriging with external drift use covariates into the kriging system to derive the local means (i.e. the trend) of the target variable (Goovaerts, 1999). Minasny and McBratney (2007) simply call this technique Empirical Best Linear Unbiased Predictor (EBLUP). Cokriging is a multivariate extension of ordinary kriging based on a linear model of coregionalization. This accounts simultaneously for spatial autocorrelation in the target variable and spatial cross-correlation between target variable and covariates (Motaghian and Mohammadi, 2011).

When covariates are considered, it is thus assumed that the spatial model is the sum of a trend related to these covariates plus a spatially autocorrelated random field component. The most robust way to estimate the variogram (i.e. the function of spatial autocorrelation of the random field) is using the method of restricted maximum likelihood (REML, first described by Patterson and Thompson (1971)). The use of Restricted Maximum Likelihood (REML) within EBLUP is logically referred as Restricted Maximum Likelihood Empirical Best Linear Unbiased Predictor (REML-EBLUP) (Minasny and McBratney, 2007).

Geostatistics are considered in many studies focusing on small extent (< 50 km2), and in particular in precision agriculture (e.g. Nourzadeh *et al.*, 2012; Casa *et al.*, 2013). Its use becomes less common as the extent increases and only a limited number of geostatistical approaches for soil mapping have been proposed at regional, national or global extent. For example, Rawlins *et al.* (2009) used airborne radiometric data and a digital elevation model as covariates within a mixed effect model to predict SOC in Northern Ireland, Marchant *et al.* (2015) used REML and kriging along with geologic data to predict topsoil inorganic C over France while Hengl *et al.* (2014) applied regression-kriging along with machine learning regression to predict multiple soil properties (including SOC) at a resolution of 1 km at global extent.

Sampling schemes that are made specifically for use with geostatistics are referred as model-based. There are commonly regular grids with some supplementary points. On the one hand, the systematic spacing between grid nodes

allows multiple observation at fixed distance lags in the empirical variogram. On the other hand, regular spacing will ensure that the whole study area is homogeneously covered. Including supplementary points will allow estimating the variogram at distances shorter than the minimal distance between grid points.

1.5 What are the current limitations for the knowledge and the spatial prediction of soil organic carbon?

Study and spatial prediction of SOC is associated with multiple difficulties. We list here the most important points that, we believe, should be considered.

1.5.1 Quality and quantity of data on soil organic carbon and auxiliary variables

Some studies comparing SOC prediction methods showed little difference in their performances (Piccini *et al.*, 2014; Ping and Dobermann, 2006). Therefore, it has been suggested that rather than adding sophistication to existing methods, efforts should be directed to collect and make available more useful and high quality data (Minasny, 2007). On the one hand, precision and accuracy is important as relying on secondary information containing errors can cause significant errors in the prediction (Simbahan *et al.*, 2006). On the other hand, consistency in and between datasets is essential too. Lugato *et al.* (2014) showed that SOC data are often not directly comparable between European countries due to the use of different conventions and methods of analysis.

The GlobalSoilMap project proposes to elaborate a fine-resolution global grid of soil functional properties with estimates of their associated uncertainties (Arrouays *et al.*, 2014a). Specifications are given to ensure consistency between data of different countries and to produce sound estimates of variables and associated uncertainties at locations where no measurements are available. The concentration of SOC is of course included in the variables. At term, the harmonized and exhaustive database will be made freely available and is expected to help solving multiple challenges related to soils.

1.5.2 Importance of soil organic carbon located below the surface layer

Important amounts of SOC are stored in subsoil, i.e. below the upper 30 cm, to a large extent in colluvial horizons (Meersmans *et al.*, 2009a; Zádorová *et al.*, 2015). Colluvial horizons contain generally a high density of stabilized SOC, likely resulting from the physical protection of SOM associated with

microaggregates and clay particles (Meersmans *et al.*, 2009a). However, recent research suggests that these buried SOC stocks may cycle biologically on annual-to-decadal time scales if the processes contributing to their protection from decomposition are altered (Chaopricha and Marín-Spiotta, 2014). In particular, how they will respond to warming is highly uncertain (Trumbore and Czimczik, 2008). A thorough monitoring and analysis of an agricultural hillslope in Belgian Loess Belt (Wiaux *et al.*, 2014) indicated that the colluvium formed at the footslope by the transfer and burial of soil material from the eroded slope was significantly enriched in labile organic carbon (c. 70 %). Authors concluded that environmental and physical factors were likely to play a more important role in the stabilization, which also suggests that it would high sensitivity to future change in environmental conditions such as climate change. For these reasons, considering subsoil is critical for SOC quantification studies.

In practice, deeper soil samples are collected using manual or mechanized auger, drill or sampling tubes. Samples are then transported to the laboratory to be analyzed. More recent methodologies also propose to scan collected soil tubes with a spectrometer to obtain a high vertical resolution SOC profile, which allows reducing costs and time requirements (Doetterl *et al.*, 2012; Cardinael *et al.*, 2015). A spatially exhaustive prediction of subsoil SOC can then be obtained using digital soil mapping techniques by depth interval. The vertical distribution can also be modeled as a mathematical function (e.g. exponential or spline) whose parameters are regressed from SOC measurements and covariates (Malone *et al.*, 2009; Meersmans *et al.*, 2009a).

Despite advances in sampling and analytic methods, many datasets provide data restricted to the upper 30 cm of the soil or at best with a coarse vertical resolution (Doetterl *et al.*, 2012). Spatial prediction of SOC over large geographic extent is therefore often limited to the topsoil layer (de Brogniez *et al.*, 2015; Meersmans *et al.*, 2012a). The GlobalSoilMap project specifications, however, mention soil variables and their uncertainty at multiple depth intervals (including total soil depth).

1.5.3 Multiple scale variability of soil organic carbon

Patterns of SOC depend on various processes interacting across a large range of spatial extents. Therefore, as the extent increases, new controlling factors contribute to spatial variability. Typically, coefficient of variation in SOC can rise sevenfold when scaling up from micro-site to landscape extent (Goidts *et al.*, 2009b). However, predicted SOC is generally aimed at specific support and extent. Efficient spatial prediction imposes that (i) the factors controlling SOC at this specific extent are correctly identified and accounted for and (ii) quality data are available for these controlling factors at adequate support.

Short-range (range below a few meters) variability of SOC is generally not easily explained from other variables (Baveye and Laba, 2015). Because of this, a lot of noise and outliers is observed when a SOC model is validated at point locations. Consequently, a quite low determination coefficient (R^2) is expected

even for models describing the system very well (Meersmans *et al.*, 2008). As pointed out by Heuvelink (1998), direct comparison of block predictions with point measurements is not valid because the point behaviour may importantly deviate from the block behaviour. Point-scale variability can be handled in the sampling design (e.g. by using composite samples) or in the prediction method (e.g. filtering of nugget component in the variogram or block-kriging). For the mapping of SOC stocks in Wallonia, (Chartin *et al.*, 2017) argues that the use of composite sampling could be a reason explaining the R^2 larger than in similar studies.

Soil models are generally developed to explain and to predict soil properties at a specific extent and support, so their transferability to other extent and support may not be guaranteed. It can be shown that both extent and support influence the correlation between variables (Corstanje *et al.*, 2007). They also have a substantial effect on the calibrated parameters of a SOC prediction model. For example, Vasques *et al.* (2012a) compared multiple configurations of study extent and support of covariates for both the calibration and the validation of soil carbon models. The support was modified by repeatedly aggregating pixels over a 2×2 pixels windows. The resolution was thus also decreased by a factor two. Results showed that using a model calibrated at a given extent (resp. support) to predict SOC at another extent (resp. support) dramatically decreased its performances. The same applies for spectroscopy. Due to strong non-linearities in the models generally used, it is clear that a model calibrated on spectral measurements at a given pixel size cannot be used for prediction using spectral measurement collected at a different pixel size without being re-calibrated.

1.6 Objectives

The first general objective of this thesis is to identify and quantify the effect of factors controlling the spatial variability of SOC concentration and stock. To this end, we focus on cropland at the landscape extent as this represents a situation where many controlling factors are expected to be involved simultaneously (cf Section 1.2.3). Factors that are considered here are profile development, texture, drainage, topography and land management. Whenever possible we also focus on SOC concentration or stock in the subsoil, as SOC at these depths is highly related to the process of soil redistribution by tillage and water erosion (cf Section 1.2.3). The first objective is specifically addressed in Chapters 2, 3 and 4.

Objective I

Identify and quantify the effect of the factors controlling the spatial variability of SOC concentration and stock.

Different processes may act at different spatial extents. For this reason, we believe that focusing separately on a specific spatial scale can help reaching

the first objective. Therefore, the second objective consists in identifying the influence of the spatial scale at which the different factors control the spatial distribution of SOC concentration and stock. The second objective is addressed in Chapters 3 and 4.

Objective II

Identify the influence of the spatial scale at which the different factors control the spatial distribution of SOC concentration and stock.

Different formats of data are exploited in this work such as measured SOC data from point sources and continuous SOC data predicted by remote sensing. Other formats are for example continuous models of elevation, discrete soil maps and legacy datasets. For these reasons, our third objective is to analyze the potential, limitations and pitfalls of the different formats of data used in this study. The third objective is addressed throughout the document

Objective III

Analyze the potential, limitations and pitfalls of the different formats of data used in this study.

The fourth objective is related to the case where only the total stock for an area is needed. It is specifically related to Chapter 5. Indeed, Chapters 2 to 4 present applications where the spatial distribution of SOC concentration or stock is central. When only the total stock is needed, the question naturally arises of the added value of the its spatial variation. The fourth objective addresses this question.

Objective IV

Determine how the prediction of spatial variation is useful for the estimation of the total SOC stock over a study area.

1.7 Overview of the following chapters

In order to meet the objectives, DSM applications will be used, fitting the general framework of the spatial mixed effect model (eq. (1.4)):

$$z(x) = f(x) + \epsilon(x) \tag{1.4}$$

where

- $z(x)$ is the predicted variable, i.e. the SOC stock at a given spatial support over a given spatial extent.
- $f(x)$ is equivalently referred to as the deterministic component, fixed effect or spatial trend and is considered to be the effect of soil and environmental factors and
- $\epsilon(x)$ is the stochastic or random component. It can be treated as a spatially correlated random field.

1.7.1 Assessment of the factors controlling vertical and horizontal distribution of soil organic carbon in the Belgian Loess Belt

In Chapter 2, two datasets of SOC profiles (the large legacy dataset *Aardewerk* and a smaller recently resampled dataset) are used to assess the contribution of different factors on the SOC variability in loam belt (defined as the area composed of both the loam and sandy loam agricultural regions). The effect of soil texture, drainage and profile development is quantified at three depth layers (0-30 cm, 30-60 cm and 60-90 cm) along with thorough significance testing. Besides, TA are computed from a digital elevation model over a 10 meters spacing grid. A multiple linear regression with quadratic terms and interactions is performed on these TA. This allows spatially predicting SOC and better understanding and quantifying the effect of topography at regional scale.

1.7.2 Separating scales of variability to explain spatial distribution of soil organic carbon

In chapter 3, spatial components of variation of SOC and texture fractions are separated in order to better understand the influence of spatial scale in the environmental control of SOC distribution. A geostatistical filtering approach similar to the one used in factorial kriging analysis is applied to separate the scales of variation of the soil properties at three depth intervals. The same legacy dataset of soil profiles restricted to the loam belt, as in Chapter 2, is exploited for SOC information. Scale components are predicted separately and the relationships between soil properties are characterized with regard to the spatial scale.

1.7.3 Detecting and quantifying field-related spatial variation of soil organic carbon

In chapter 4, a SOC image obtained by airborne spectroscopy in Grand Duchy of Luxembourg is used to estimate the part of the spatial variation of SOC related to LULM. This was defined as a component of variation which is homogeneous within the boundaries of a field but varies from one field to another. In

contrast, the variation at similar scale due to other processes such as geomorphological processes was postulated to be completely independent from field boundaries and was modelled by a spatial random field. Besides, the effect of soil type was first estimated and filtered using a soil map. The spatial model was fitted in the most rigorous (unbiased) way and uncertainty was assessed using a methodology based on repeated random sampling.

1.7.4 Accounting for the spatial scale of variation in total stock quantification

In chapter 5, we focus on the generic problem which consists in estimating at best the total stock of SOC (or equivalently the stock of SOC per surface unit) over a continuous spatial domain of interest, where SOC stock per surface unit is viewed as a two-dimensional geographically distributed variable. Total SOC stock is estimated by integrating over this domain the prediction provided in any spatial location by a model built on a vector of auxiliary variables. First, the link between the prediction variance of the total stock over the domain as obtained by the model and the range of spatial dependence of model errors is explored. Then, we assess the reduction of variance of prediction that can be reached by accounting for the spatially dependent model errors. The prediction variance obtained using the kriging interpolator (i.e. the best linear predictor) is compared to the prediction variance obtained with two simpler predictors. The influence of the ratio between the range of dependence of model error and the size of the domain on the performance of these three predictors is assessed.

Chapter 2

Assessment of the factors controlling vertical and horizontal distribution of soil organic carbon in the Belgian Loess Belt

---

This Chapter is based on the following paper: Stevens, F., Bogaert, P., Van Oost, K., Doetterl, S., and Van Wesemael, B. 2014. Regional-scale characterization of the geomorphic control of the spatial distribution of soil organic carbon in cropland. European Journal of Soil Science, 65(4):539–552

Abstract.

The heterogeneity of the spatial distribution of soil organic carbon (SOC) at the landscape scale is generally not considered in regional or national SOC dynamics models. In cropland this heterogeneity is controlled largely by topography, which influences the distribution of water, energy and sediments, and thus the SOC dynamics. Sediment re-distribution rates have increased strongly since the mechanization of agriculture. The over-simplification of landscape processes in regional models of C dynamics may add to the uncertainty in C balances. Therefore, a better characterization of the importance of landscape effects on the SOC distribution throughout a region is needed. In this chapter we characterized the relative importance of geomorphology on the SOC horizontal and vertical variability across the croplands in the loam belt. A large legacy dataset of soil horizons was exploited together with 147 recently sampled profiles. Mean SOC depth profiles for different soil types were compared. Various topographic attributes were computed from a digital elevation model, and their influence on SOC was quantified through simple linear models. Finally, SOC content was

mapped at three depth layers through multiple linear models, and results were cross-validated. The legacy dataset allowed identifying significant differences in the mean SOC profile according to texture, drainage or profile development classes. A clear relationship between SOC content and topographic attributes was demonstrated, but only for the recently sampled profiles. This may be explained by a substantial error on the location of the profiles of the legacy dataset. This study thus shows evidence that the major control on the vertical distribution of SOC is related to topography in a region where observed heterogeneities for other commonly involved factors are limited. However, the large amount of unexplained variability still limits the usefulness of the spatial prediction of SOC content, and suggests the importance of additional influencing factors.

2.1 Introduction

The SOC pool is of great importance for the global carbon cycle. Soil represents the largest terrestrial C pool, containing more organic C than biosphere and atmosphere together (Grace, 2004). However, it is the largest source of uncertainty in regional and continental C balances of terrestrial ecosystems (van Wesemael *et al.*, 2010b). For example, predictions about the response of the soil carbon store to global warming are diverging (Trumbore and Czimczik, 2008). As a result, soils are increasingly receiving attention for the potential role they can play in CO_2 mitigation (Milne *et al.*, 2007). One of the sources of uncertainty in C flux modeling between soils and atmosphere arises from the lack of consideration of the landscape processes influencing SOC (Ciais *et al.*, 2008). Models predicting the temporal change of SOC for a region typically represent the soil system as a collection of large spatially homogeneous units (Easter *et al.*, 2007). However, horizontal and vertical variability of SOC within landscapes is large (Stevens *et al.*, 2006; VandenBygaart *et al.*, 2007; Hbirkou *et al.*, 2012a). For example, (Goidts and van Wesemael, 2007) showed that SOC variability at the field extent is of the same order of magnitude as the variability inside a large map unit of cropland. The implications of this large variability for model parametrisation, predictions and soil monitoring schemes are still poorly understood.

Topography is typically related to the spatial patterns of SOC in the landscape because geomorphic landscape features control hydrological and erosional processes, and soil temperature (Moore *et al.*, 1991; Florinsky *et al.*, 2002). Transport of sediments influences C fluxes between soil and atmosphere through the dynamic replacement of eroded C at eroding sites and reduced decomposition of buried C at depositional sites (Van Oost *et al.*, 2005; Quinton *et al.*, 2010; Vandenbygaart *et al.*, 2012). Conceptual models have been developed that combine geomorphic models with models of carbon dynamics. They have been applied to eroding micro-catchments and were able to closely reproduce the observed SOC depth profiles for a wide range of erosional and depositional settings (Liu *et al.*, 2003; Rosenbloom *et al.*, 2006; Dlugoß *et al.*, 2010). Ap-

plication of these models suggests that lateral fluxes of SOC, sediments and water will further enhance the spatial heterogeneity in SOC storage within agricultural landscapes. However, these studies typically focus on small areas with large erosion rates and pronounced topography. So far, the importance of erosion-induced and topography related variability in C stocks at the regional extent remains unclear.

However, spatial prediction of SOC content accounting for variability within the landscape is still a challenging task, given the large number of additional processes occurring at the landscape scale (Viaud *et al.*, 2010a). Even if the main controls are identified, the true landscape condition differs from the ideal conditions of studies at field and hillslope extent (Van Hemelryck *et al.*, 2011). A complex lateral and vertical distribution of SOC may result from the historic and current interactions between many processes including sediment transfers or agricultural management (Sleutel *et al.*, 2007; Goidts *et al.*, 2009b). (Doetterl *et al.*, 2012) showed that even if the sub-soil SOC content is generally larger at depositional than at eroding positions along a slope, differences between similar geomorphological positions of different slopes may be important. The difficulty to predict soil horizon thickness at a large spatial resolution (Vanwalleghem *et al.*, 2010) also suggests that the landscape processes result in important point or plot scale variability in soil properties. A spatially explicit description of the soil processes and properties at landscape scale is, however, of prime importance to simulate the evolution of SOC stocks in the coming decades (Viaud *et al.*, 2010a).

To address this issue, the main objective of this paper is to assess the importance of the topographic control on SOC spatial distribution, over a large area of cropland, accounting for landscape scale variability. The effect of topography will be characterized on both the vertical SOC variability directly, and the lateral variability at different depths. The second objective is to compare the potential of a legacy dataset and a more recent dataset to fulfill the first objective.

2.2 Material and methods

2.2.1 Study area

The study area was chosen to maximize variability in topographic features while limiting variability in other environmental factors influencing SOC. To this end, the cropland of the loam belt in central Belgium was selected. In this thesis, the loam belt was defined as the loam and the sandy loam Belgian agricultural regions like in Lettens *et al.* (2004) (Figure 2.1). The loam belt is thus an area of 9921 km^2 of which 43 % is occupied by cropland. It is characterized by a rolling topography with plateaus, slopes and some incised rivers with generally well drained, dry valley bottoms. The climate of the region is a temperate oceanic climate with mild winters and cool summers. The region is part of a mosaic of loess regions over Europe (Haase *et al.*, 2007). The geological

substrate is a several-metres thick Pleistocene aeolian deposit of calcareous loess in which luvisols have developed. The northern and northwestern parts of the region, corresponding to the sandy loam agricultural region, are rather composed of sandy loess (Haase *et al.*, 2007). Loess deposits are typically thicker on south facing slopes than on north facing ones and overlay tertiary sands (Vanwalleghem *et al.*, 2010). In some locations, these sandy layers are already apparent at the surface, as soil erosion has removed several metres of the loess. At present, the main crops in the region are cereals (46 %), sugar beet (20 %), silage maize (10 %) and potato (7 %) (van Wesemael *et al.*, 2010b) with typical C inputs to the soil by plants estimated at 1.93 Mg C ha^{-1} $year^{-1}$ and by manure at 1.1 Mg C ha^{-1} $year^{-1}$. Forests, grassland and urban areas are more typically located in the floodplains or on the slopes.

2.2.2 Carbon datasets

Characterizing the SOC spatial distribution in both top-soil and sub-soil required the use of spatial datasets describing the SOC profile. In our analysis, both a legacy and a recently-sampled dataset of soil profiles were used. Only samples located in cropland at the time of sampling were selected. The legacy dataset is composed of soil profiles described during the Belgian National Soil Survey (1947–1962) (De Leenheer *et al.*, 1968). Soils were sampled by horizon from observation pits, and physical and chemical analyses were performed in the laboratory. The SOC was estimated by dichromate wet combustion. Soil profile location, horizon limits, physical and chemical properties and classification were recorded on paper.

In 1993, the legacy dataset was digitized creating the digital soil database Aardewerk (Van Orshoven *et al.*, 1988). However, because of storage constraints, data description was simplified and standardized and, as a result, part of the information was lost. A critical simplification was the coding of the horizon boundaries as intervals instead of exact values. Some horizons were also omitted during the transcription. The methodology that we developed to reconstruct detailed SOC profiles from incomplete and uncertain information of Aardewerk is described later. From all the profiles under cropland, 543 were removed because they had missing horizons or inconsistent values for some variables, leaving 2449 profiles available. In the following, the dataset containing the reconstructed SOC profiles located in the study area will be called AW93. The position of the profiles can be seen on Figure 2.2.

Recently, original descriptions of the soil profiles located in Flanders were again digitized, but this time with the exact limits of the horizons (Van De Vreken *et al.*, 2011). This dataset was used to validate the SOC profile reconstruction method (see later). We selected 944 profiles for which SOC values were coherent with the corresponding AW93 profiles. This dataset is referred to as AW10. A common factor of 1.33 was used to correct SOC concentration for the incomplete oxidation resulting from the Walkley & Black method, in AW93 and AW10.

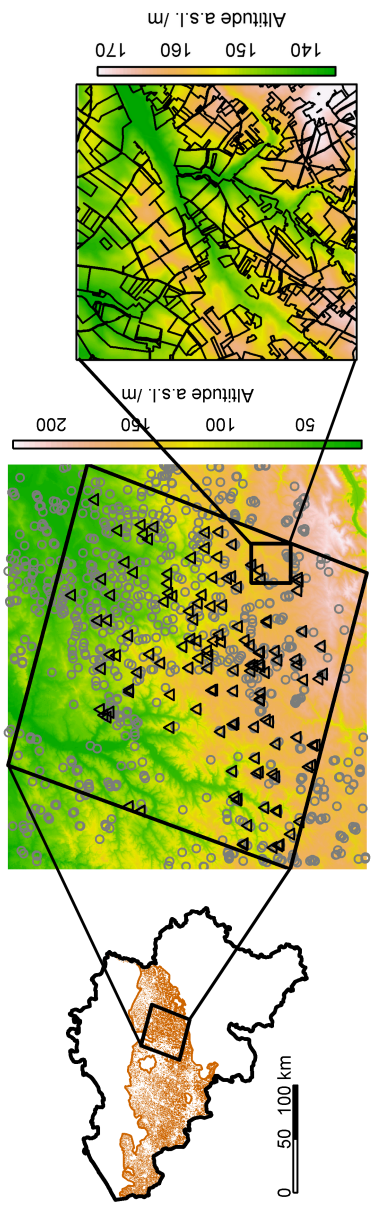


Figure 2.1: Left-hand picture: the loam belt (orange polygon) is largely occupied by cropland (orange fill). Centre picture: AW93 (blue points) and R_AW (red triangles) data-sets profile sites are presented across the re-sampled sub-area of the study region. Right-hand picture: area used for plotting the spatial prediction of SOC, with field limits drawn in black.)

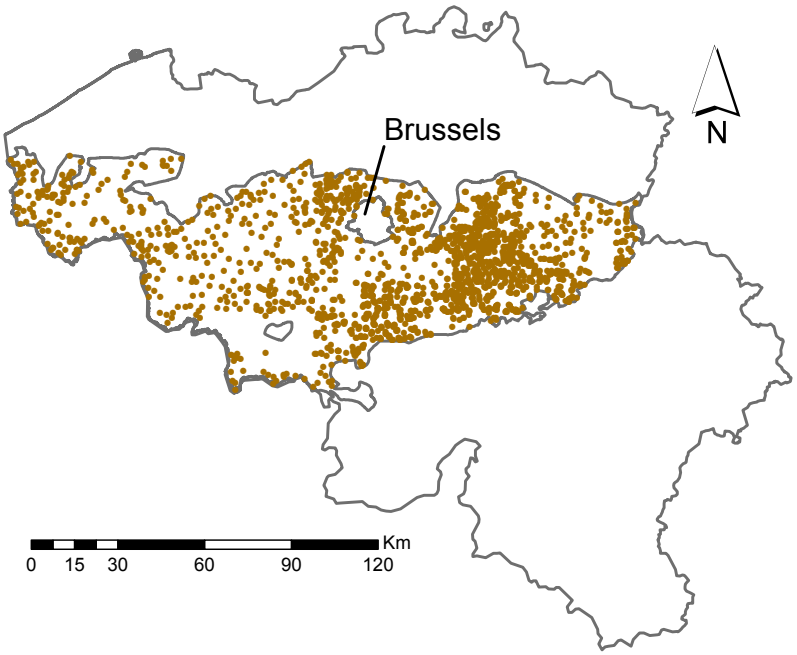


Figure 2.2: Location of the 2449 profiles of the dataset AW93 spread in the cropland of loam belt.

Finally, we also used 139 profiles from a recently sampled dataset (Doetterl *et al.*, 2013b). These profiles were randomly selected from existing AW93 profile locations within a 40 km x 40 km sub-area ((2.1)). It is not possible to affirm that their position exactly matches the position of the corresponding AW93 profile, since no marker was left in the soil during the original sampling. We also checked that the current land-use had not changed since the Belgian National Soil Survey. Between 2010 and 2012, soil cores were extracted and analysed for carbon with a spectrometer at 3 cm depth intervals up to one metre. The root mean square error associated with these estimates was small (1.22 g C kg^{-1}), and similar to the analytical error of the reference technique of dry combustion. In the following, this dataset will be referred as R_AW (for re-sampled Aardewerk).

The two datasets which were used separately for describing the spatial distribution of SOC, AW93 and R_AW, were different but complementary. The legacy dataset (AW93) contains a large number of data, with associated observations on soil properties, but a poor precision in both the coordinates and the retrieved SOC content at a given depth. The recent dataset (R_AW) contains a one order of magnitude-smaller number of profiles and no associated soil observations, but a very good precision in both coordinates and depth-explicit SOC content. In this study, the large number of profiles in AW93 is an advantage to demonstrate statistical differences in SOC depth distribution between soil classes, and the good precision in the coordinates of R_AW profiles is an advantage for three-dimensional spatially-explicit predictions.

2.2.3 Terrain attributes

TAs derived from a digital elevation model (DEM) were used in our analyses. The DEM was produced by merging two DEMs of 10 and 5 metres resolutions that respectively cover the two administrative Belgian regions of Flanders (along with Brussels) and Wallonia ((Demarcin *et al.*, 2009; AGIV, 2006). A new DEM was then created by linearly interpolating these two dataset on a grid of 10 meters spacing. The newly created dataset was then smoothed by using a 3 × 3 mean filter on the same grid, leading to a 30 m × 30 m support size.

TAs were computed using Matlab and the TopoToolbox package (Schwanghart and Kuhn, 2010). The attributes were elevation (ELEV), slope gradient (GRAD), total curvature (CURV), south orientation (SOUTH), topographic wetness index (TWI, (Beven and Kirkby, 1979)), stream-power index (STR) and the topographic position index with different neighbourhood sizes (TPI32, TPI128 and TPI512 Weiss, 2001). The flow accumulation, implied in the calculation of hydrological attributes, was computed using a multiple flow direction algorithm. TPI is the relative difference between the elevation of a cell and the average elevation of the cells within a given radius. Three radii were used (32 m, 128 m and 512 m), in order to represent different scales. The south orientation was taken as the cosine between the slope orientation vector and a vector pointing south. It was selected to reflect the anisotropy of the geomor-

Table 2.1: Predefined intervals for the horizon depth and thickness as coded in the legacy dataset AW93.

Horizon depth /cm	Horizon thickness /cm
0 - 5	0 - 2
6 - 10	3 - 5
11 - 20	6 - 10
21 - 40	11 - 20
41 - 60	21 - 30
61 - 80	31 - 50
81 - 100	51 - 100
101 - 150	>100
>150	

phic variables observed in the region, with loess deposits thickness and texture depending on slope orientation (Goossens, 1997).

2.2.4 Horizons distribution reconstruction

Because information was lost while digitizing the legacy profile dataset, the vertical position of each horizon in AW93 was not given as an exact value but as two intervals: one indicates the depth of the upper boundary of the horizon, the other its thickness class. These intervals belong to a set of predefined intervals (Table 2.1). In order to address this issue, a Monte Carlo simulation based method was developed (Figure 2.3). For a given profile, the thickness of each horizon was simulated from a uniform distribution within the given thickness interval. Then the corresponding horizon depths were derived assuming that there was no gap or overlap between horizons. The simulated set of horizon thickness values was retained only if the derived horizon depths were also enclosed in the depth intervals given in AW93, for the corresponding horizons. The simulated sequences of horizons that were not coherent with the original information given by AW93 were discarded. A smooth SOC concentration profile was then obtained from the simulated distribution of the horizons and the SOC horizon values, by equal area quadratic spline (see section 2.2.5). To ensure precision of the method, the whole process was repeated until a sufficient number of valid horizon sequences (10 000) were simulated. The estimated profile was then taken as the mean of these. This method allowed us to use all the information on horizon distribution contained in AW93 simultaneously without making an arbitrary choice potentially leading to bias (such as taking, for example, the middle of the thickness intervals as exact thicknesses).l

The AW10 dataset was used to validate the horizon distribution reconstruction method, because it contained the same profiles than AW93, but provided the exact horizon limits. The profiles of AW10 were also smoothed with a spline method (see later). Then, all the AW10 profiles were compared with the corresponding AW93 profiles. The mean profile of both datasets and the

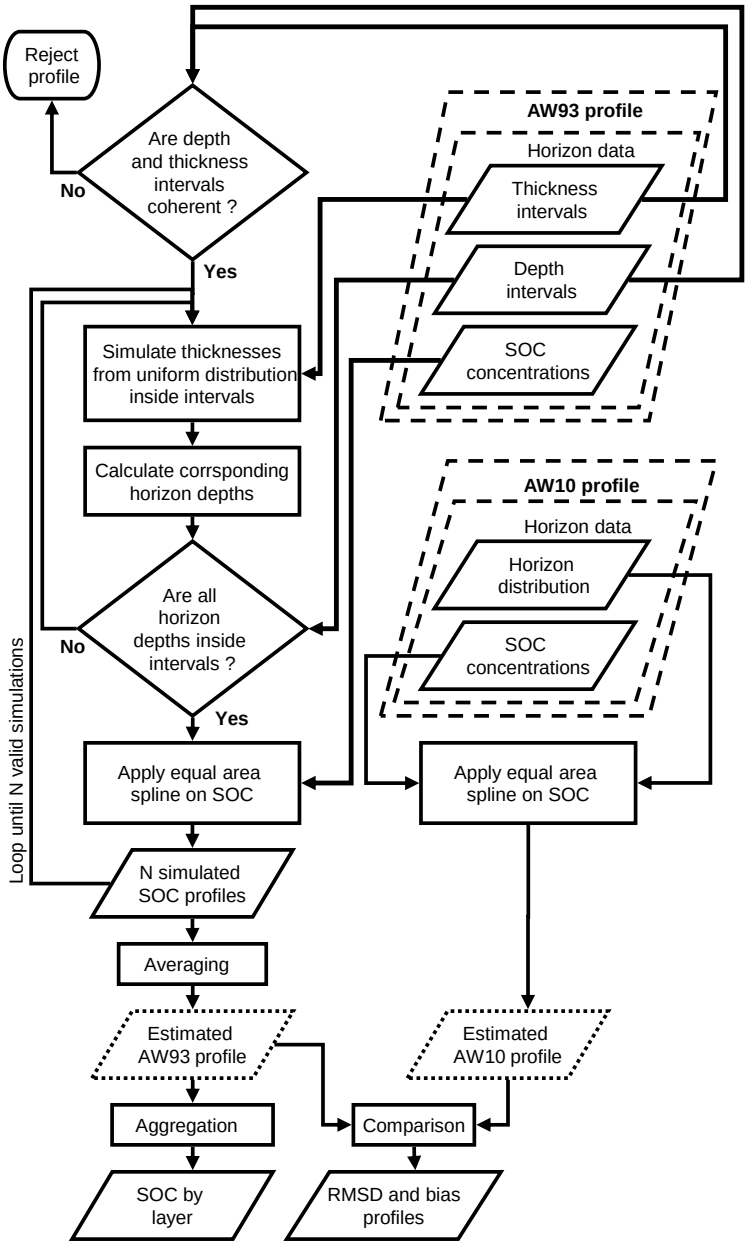


Figure 2.3: Reconstruction of the horizon distribution for AW93 data and validation of the method using AW10 data.

profile of mean difference (bias) and root mean square difference (RMSD) were computed. This resulted in an estimate of the magnitude of the error caused by the coding of the horizon limits used in the legacy dataset including its depth distribution.

2.2.5 Smoothing the profiles

The SOC concentration was provided for each horizon in AW93 and AW10. However, estimating depth-explicit profiles of SOC was needed for further analysis. Traditionally, it is assumed that the horizon value of a particular attribute represents its average value over that horizon (Malone *et al.*, 2009). In luvisols that developed in thick loess deposits, soil properties including SOC content are expected to vary continuously with depth, except at the transition zone below a mixed plough layer (Minasny *et al.*, 2006; Myers *et al.*, 2011; Kempen *et al.*, 2011). This is confirmed by the re-sampled cores that do not display SOC vertical patterns of abrupt transitions between horizons (Doetterl *et al.*, 2013b). Thus, using a step function with mean SOC value over each horizon could lead to bias and cumulative errors may produce under- or over-estimates for a given soil layer (Ponce-Hernandez *et al.*, 1986).

Although a SOC profile is often described with a negative exponential depth function, fitting this function does not ensure that the mean SOC content measured for each horizon is respected. We therefore chose to apply the equal area quadratic spline method of (Bishop *et al.*, 1999). This method produces smoothly varying profiles that are consistent with the observed mean SOC values for each horizon. However, we assume that the first horizon, the soil surface horizon, is a tillage horizon and that the SOC content is constant because of continuous mixing by tillage operations. Therefore, inside the depth interval corresponding to the first horizon o interpolation was made and the mean horizon value given in the database was used instead. The SOC concentration was then calculated for three depth layers (0–30, 30–60 and 60–90 cm) from the interpolated profiles.

2.2.6 Computing class-representative soil organic carbon profiles

Even though the legacy dataset did not permit to spatially predict SOC content (see Results and discussion), it allowed us to characterize the influence of different factors on the vertical distribution of SOC. To compare the influence of these factors on SOC profile inside the study area, profiles were grouped according to the value of a factor or a combination of two factors. Then, for each group, a representative profile was computed by taking the mean of all the profiles in this group, and the standard error on the mean was computed for all depths. Finally, mean profiles of different group were compared.

First, the grouping was made according to the soil type. The AW93 profile descriptions contain information about soil texture, drainage and profile development classes. The profile development classification was simplified to the

presence or absence of a colluvial layer above 120 cm depth (*colluvial* and *not colluvial* classes). This variable may indeed be considered as a proxy for the topography since the occurrence of a colluvium largely depends on the position on the hillslope. Mean profiles were computed for the interaction between drainage and development, and between texture and development. In order to have a number of elements in each group allowing meaningful statistical analyses, the classifications of AW93 drainage and texture were also simplified. Drainage was divided between well- or poorly-drained. Poorly drained profiles display traces of temporary water saturation (pseudo gley) above 80 cm while well-drained profiles do not. Twenty-four profiles (0.9 %) showing traces of permanent water saturation (gley) were discarded. Texture was split into three classes: *silt loam*, *heavy sandy loam* and *light sandy loam*. One hundred and three profiles (4.2 %) belonging to other texture classes (mainly clay) were not used. Only those groups containing more than 30 profiles are shown in the results.

To evaluate the direct influence of the topography on the SOC profile, a method similar to soil class representative profiles was applied, but this time using TAs and the R_AW SOC dataset. This analysis was not performed with AW93 or AW10 profiles as the extraction of TAs at the location of these profiles was problematic due to error in the spatial coordinates (see Section 2.3.4. For each TA, the profiles were sorted by increasing value of this TA, and divided into three groups of equal size. Mean profiles were then computed for the three groups, and comparisons made between them.

For both the soil property and TAs classifications, the within-class variability of SOC concentration for our three successive 30-cm thick layers was computed and displayed as a boxplot using the ggplot2 package (Wickham, 2009). For each layer, an ANOVA F-test was used to test the hypothesis that all the means are equal. When the mean of the SOC values in one or many groups showed a relevant similarity or difference with the mean of the SOC values of another group, a t-test was performed to question the hypothesis of equality of the means. The AW93 profiles are distributed all over the study area, whereas those of R_AW are restricted to a smaller sub-area. Thus, the large number of profiles increases the power of statistical tests for AW93.

2.2.7 Spatial prediction of soil organic carbon

In order to estimate the importance of topography on the spatial distribution of SOC, statistical analyses were performed with AW93, R_AW and TAs. As a first step, the Pearson coefficients of correlation between the observed SOC content in the three depth-layers and the TAs were computed. The correlations were analyzed for both AW93 and R_AW, in order to assess their ability to spatially predict SOC.

Because of the very weak correlation between TAs and SOC content in AW93 (see Table 6), a procedure was developed to check if these weak values could originate from a lack of precision in profile geo-locations. Simulated random errors of given magnitude and random direction were added to the

coordinates of the profiles in R_AW. These errors are zero-mean distributed with a standard deviation of 5 meters that corresponds to the precision of the GPS. Their mean is equal to zero due to the absence of bias. The magnitude of the simulated errors was increased gradually from 0 to 300 metres by 10-metre increments (2.7). Then, correlations between SOC layer contents and TAs extracted at modified positions were computed. Because of the issue of the precision of the AW93 samples, the next steps of the SOC spatial prediction were performed only with the R_AW dataset.

A multiple linear regression model was used as the next step to predict SOC from the TAs. TAs, TAs at second power and interaction between all pairs of TAs were all tested as predictors. All models with up to three predictor variables were tested individually. It was, however, not possible to test every model with more than three predictor variables due to the large number of possible combinations for the variables and the corresponding computational burden. Instead, a stepwise regression was used as automatic feature selection method. Overall, the final model was selected as the the one that minimized the residual sum of squares in a ten-fold cross-validation procedure. This led to the best predictive model among tested ones.

Finally, predicted SOC spatial stock from the best model was mapped in a 5 km $\times$ 5 km plotting area (see 2.1), randomly selected inside the sub-area containing the R_AW profiles. The stock was calculated by multiplying for each grid node and each depth layer the predicted SOC concentration with the estimated bulk density, expressed itself as a function of SOC concentration by the general model of Manrique and Jones (1991). This bulk density model was chosen for its simplicity and, relative to other models, its better ability to predict soil bulk density at the surface as well as down profiles in Belgium (Boucneau *et al.*, 1998).

2.3 Results and discussion

2.3.1 Horizon distribution reconstruction

The profiles calculated from incomplete horizon information (AW93) were compared with a subset of corresponding profiles that were derived from another data-base with exact horizon descriptions (AW10). The mean profile of both sets and the vertical profile of mean difference (bias) and root mean square difference (RMSD) are presented in Figure 2.4. Bias is very small across depth, and reached, between 15 and 30 cm, a minimum followed by a maximum, all having an absolute value less than 1 g C kg^{-1}.

This may indicate a trend to under-estimate the thickness of the surface horizon, since the transition between the surface horizon and the second horizon was generally also between 15 and 30 cm, and the surface horizon contained generally more SOC than the second. The RMSD profile showed peaks in the same region, reaching up to 25 % of the mean SOC content of AW10. This results because the SOC content is large, and decreased quickly below the

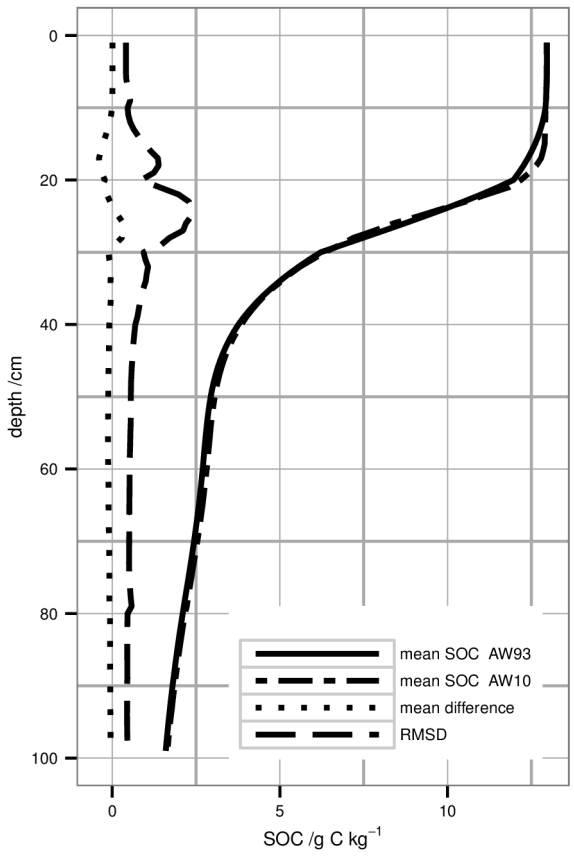


Figure 2.4: The inaccuracy and imprecision resulting from incomplete horizon distribution information is limited by using a Monte Carlo based simulation procedure. Error is at a maximum around the expected bottom limit of the plough layer (15–30 cm).

Table 2.2: Statistics for the grouping of the SOC layer contents by combined drainage and profile development classes, for the AW93 dataset, and summary of the analyses of variance.

		SOC_{0-30cm} /g C kg^{-1}	$SOC_{30-60cm}$ /g C kg^{-1}	$SOC_{60-90cm}$ /g C kg^{-1}
	Size	Means		
Well-drained colluvial	363	12.63 ± 0.18	4.68 ± 0.13	3.40 ± 0.10
Poorly-drained colluvial	179	11.92 ± 0.26	4.08 ± 0.16	2.68 ± 0.11
Well-drained non-colluvial	1072	12.16 ± 0.11	3.48 ± 0.06	2.27 ± 0.05
Poorly-drained non-colluvial	823	10.70 ± 0.13	3.02 ± 0.07	1.65 ± 0.05
ANOVA				
Between class SS		1301.7	1023.8	968.7
Within class SS		27389.7	9086.4	5554.8
Between class MS		433.89	341.26	322.90
Within class MS		11.26	3.74	2.28
F value		38.53	91.34	141.37
P value		$<10^{-15}$	$<10^{-15}$	$<10^{-15}$

plough layer, and thus the absolute error was also large. Above 10 cm and below 50cm, the RMSD was stable and varied between 0.4 and 0.5 g C kg^{-1}.

It should be kept in mind that SOC profiles in both AW93 and AW10 were estimated with the spline smoothing method and are not the true SOC profiles. The smoothing method also assumes that the SOC concentration associated with each horizon is the exact mean SOC concentration over this horizon, an assumption that could not be verified. More complex horizon reconstruction methods could be tested, for example by using the exact horizon distributions of AW10 to estimate an a priori horizon distribution for AW93. However, we were not aiming to develop complex methods for this task and these results were adequate for our present purposes.

Legacy datasets are still valuable and they also represent a previous situation. Our reconstruction method dealt with a specific problem in the data format, and cannot assumed to be transferable to other legacy datasets. We were also fortunate to have a large number of profiles (AW10) without this specific formatting problem which allowed us to validate our reconstruction methods.

2.3.2 Soil property influence on soil organic carbon profile

The soil property influence on the vertical distribution of SOC in AW93 is shown in Figure 2.5 and Tables 2.2 & 2.3. Generally, the mean profiles of each

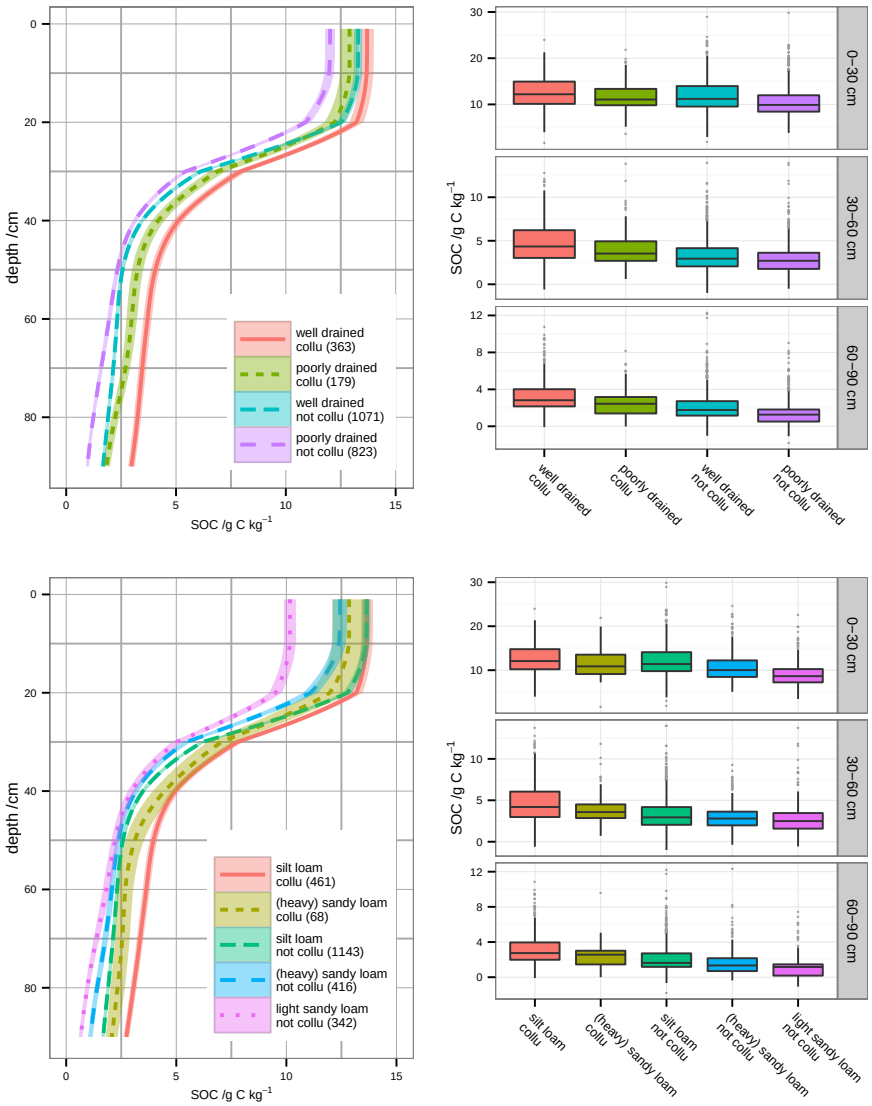


Figure 2.5: Left-hand side: mean SOC profile for soil property classes using the AW93 profiles. Standard error of the mean is represented by band along the curve. The number of profiles in each class are inside parentheses. Right-hand side: distribution of SOC for each class by 30-cm thick soil layers.

Table 2.3: Statistics for the grouping of the SOC layer contents by combined texture and profile development classes, for the AW93 data-set, and summary of the analyses of variance.

		SOC_{0-30cm} /g C kg^{-1}	$SOC_{30-60cm}$ /g C kg^{-1}	$SOC_{60-90cm}$ /g C kg^{-1}
	Size	Means		
Silt-loam colluvial	461	12.62±0.16	4.77±0.11	3.28±0.08
Heavy sandy loam colluvial	68	11.70±0.48	4.10±0.28	2.70±0.20
Silt-loam non-colluvial	1144	12.51±0.11	3.52±0.06	2.32±0.05
Heavy sandy loam non-colluvial	415	11.09±0.18	3.06±0.08	1.85±0.08
Light sandy loam non-colluvial	342	9.07±0.17	2.77±0.11	1.26±0.07
ANOVA				
Between class SS		3087.7	1060.2	967.8
Within class SS		25712.1	9134.9	5659.7
Between class MS		771.92	265.06	241.96
Within class MS		10.60	3.77	2.33
F value		72.80	70.36	103.67
P value		<10-15	<10-15	<10-15

group had similar shapes, and were differentiated well with few intersections of the curves. The intra-class variability for the SOC content inside the depth layers was large (Figure 2.5, right-hand picture). However, because of the large number of profiles in each group, significant differences between mean layer SOC content of different classes can be observed, as indicated by the very small p-values.

The upper part of Figure 2.5 and Table 2 show the combined influence of drainage and profile development on SOC distribution. In the top layer, differences between groups were small compared with their mean values. The smallest SOC content was in non-colluvial poorly drained profiles, which contained 12 % less SOC in this layer than in other profiles ($p < 10^{-15}$). In the deeper layers, however, all the classes had relatively larger differences. In the bottom layer, the poorly drained but not colluvial profiles, contained on average more than two times less SOC than the well-drained colluvial profiles ($p < 10^{-15}$). Well drained profiles are associated with higher SOC content, a relationship that is not always observed (Tan *et al.*, 2004).

The lower SOC content that is typically associated with poorly drained soils (i.e. soils experiencing temporary water saturation) can be explained by different reasons. On the one hand, water saturation can inhibit root and plant development due to the lack of oxygen in soil, therefore reducing photosynthesis and biomass input. On the other hand, the activity of microorganisms also decreases below a threshold due to non optimal wetness conditions, leading to a lower decomposition of SOC. The lower SOC content associated with poorly drained soils seems thus to indicate that the combination of reduced biomass input along with reduced decomposition from microorganisms has still on average a detrimental effect on SOC content within the context of our analysis. Nevertheless, below the top-layer, the influence of profile development was still more pronounced than the influence of drainage. Besides, we also observe a positive correlation between finer textural classes and well drained soils in the region, with thicker layers of loess deposits associated with deeper water tables. Therefore, the larger SOC concentration observed with a better drainage is partly an effect of the texture too.

Classification based on texture and development combinations (the lower part of Figure 2.5) also shows differences in SOC vertical distribution. For the top-soil, the mean SOC values were first sorted by texture, then by profile development. In particular, light sandy loam but not colluvial profiles contained on average 25 % SOC less than other not colluvial profiles ($p<10^{-15}$), showing that even when all soils are loamy, texture can have a drastic influence. In contrast, a difference for example between the mean SOC content for colluvial silt loam profiles and non-colluvial silt loam profiles cannot be observed ($p = 0.26$). Texture has thus a larger influence than profile development on top-soil SOC content. In the two deepest layers, however, substantial differences may be observed between classes with similar texture and different profile development. In the 60–90 cm layer, silt loam colluvial profiles contain on average 41% more SOC than silt loam non-colluvial profiles ($p < 10^{-15}$). The curves

Table 2.4: Statistics for the grouping of the SOC layer contents by wetness index classes, for the R_AW data-set., and summary of the analyses of variance.

		SOC_{0-30cm} /g C kg^{-1}	$SOC_{30-60cm}$ /g C kg^{-1}	$SOC_{60-90cm}$ /g C kg^{-1}
	Size	Means		
Wetness Index I	44	9.59±0.23	2.87±0.14	1.87±0.07
Wetness Index II	45	9.65±0.16	2.99±0.14	1.99±0.08
Wetness Index III	45	10.42±0.21	4.16±0.25	2.68±0.14
ANOVA				
Between class SS		19.33	45.49	17.24
Within class SS		235.75	203.31	62.10
Between class MS		9.67	22.74	8.62
Within class MS		1.80	1.55	0.47
F value		5.37	14.65	18.19
P value		0.0057	$1.8 \cdot 10^{-6}$	$1.1 \cdot 10^{-7}$

are first sorted by profile development, then by texture, showing that profile development is the most important effect in the deeper soil layers.

In summary, the influence of soil properties on SOC differed with depth and the presence of colluvium was the soil property which influences the most the deeper SOC content for our study area. Indeed, within the range of variability of texture and drainage of the cropland, the differences of deep SOC resulting from differences in profile development are larger than those resulting from differences in texture and drainage. The use of texture or drainage classes in combination with profile development classes allows the influence of profile development alone to be identified. It should be pointed out that the soil property class-representative profiles are not intended to represent actual soil profiles, but only mean profiles to be used for regional scale SOC distribution quantification.

2.3.3 Terrain influence on soil organic carbon profiles

TAs of the R_AW profiles were divided into three groups of equal size according to their order. For the sake of brevity, only the results for TWI and GRAD are shown and discussed (Figure 2.6, Tables 2.4 & 2.5). For TWI, the mean SOC profile of the third class (that with the largest TWI values) had the largest SOC content (Figure 2.6). This third class contained, on average 8 % ($p = 0.0019$), 42 % ($p = 3.4 \times 10^{-5}$) and 39 % ($p = 6.4 \times 10^{-6}$) more SOC than the two other classes, for the 0–30, 30–60 and 60–90-cm layers respectively. The positive relationship between SOC and TWI cannot be explained by the effect of the wetness itself. This would contradict the trend previously observed for the mean profiles by soil class showing that, for a given soil development class, the drainage class containing most SOC was the least wet (Figure 2.5). It can, however, be explained by the role of water erosion, because accumulation of

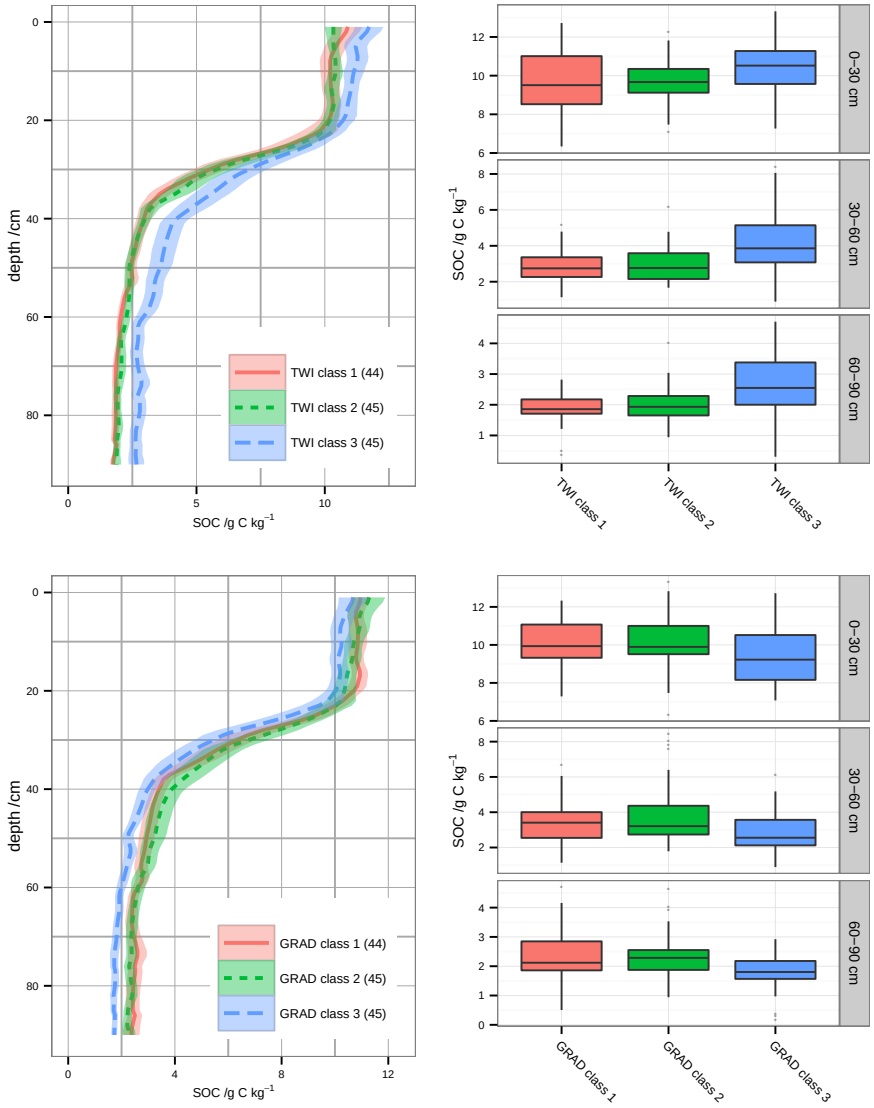


Figure 2.6: Left-hand side: mean SOC profile for TA classes using the R_AW profiles. Standard error on the mean is represented by the band along the curve. The number of profiles in each class are inside parentheses. Right-hand side: distribution of SOC for each class by 30 cm thick soil layers. (GRAD = slope gradient ; TWI = topographic wetness index).

Table 2.5: Statistics for the grouping of the SOC layer contents by slope gradient classes, for the R_AW data-set, and summary of the analyses of variance.

		SOC_{0-30cm} /g C kg^{-1}	$SOC_{30-60cm}$ /g C kg^{-1}	$SOC_{60-90cm}$ /g C kg^{-1}
	Size	Means		
Slope gradient I	44	10.13 ± 0.17	3.45 ± 0.17	2.41 ± 0.14
Slope gradient II	45	10.13 ± 0.21	3.75 ± 0.25	2.35 ± 0.11
Slope gradient III	45	9.47 ± 0.22	2.88 ± 0.16	1.76 ± 0.08
ANOVA				
Between class SS		13.10	17.95	11.53
Within class SS		247.21	230.99	71.68
Between class MS		6.55	8.97	5.76
Within class MS		1.89	1.76	0.55
F value		3.47	5.09	10.53
P value		0.034	0.0074	$5.7 \cdot 10^{-5}$

runoff in concave areas with gentle slopes induces sedimentation. This results in the burial of former C-rich soil below the plough layer where mineralization potential is reduced.

The effect of the slope classes (GRAD) was apparent mainly for the third class (Figure 2.6, Table 2.5). For the three layers successively, the mean content of the third class contained, respectively, 6 % ($p = 0.014$), 20 % ($p = 0.0014$) and 26 % ($p = 8.9 \times 10^{-7}$) less SOC than the mean of the two other classes. This third class included the profiles with a slope gradient larger than 3.4 %. This may be explained by the loss of SOC when the soil profile was truncated during erosion process. In the plough layer, the loss of carbon can be compensated by inputs from current crops which would explain why the difference between classes was small in the 0-30 cm layer. However, in the deeper layer, inputs of SOC are reduced and replacement of SOC would be only partial.

The classification of the profiles by TAs demonstrated a clear and significant influence on the SOC profile. We also grouped the AW93 profiles by TA classes on individual texture and drainage classes. However, the results did not show the relationships shown with R_AW (data not shown), possibly because of the positioning errors in AW93 (see later).

2.3.4 Analysis of the correlations for legacy and new data

The correlations between TAs and SOC layer contents, and between SOC layers themselves, were computed for the AW93 and R_AW datasets (Tables 2.6 & 2.6). Our results show a strong dependence between the SOC content of the three layers for both datasets. The correlation coefficient (r) was 0.43 between the top and the bottom layer, and varied between 0.5 and 0.65 for that between adjacent layers. There were small differences between the two datasets probably because of the characteristics of AW93 and the methods

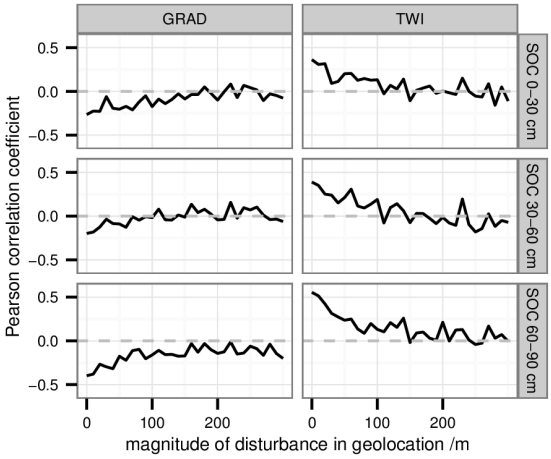


Figure 2.7: Correlations between SOC layers and TAs are sensitive to a perturbation of the originally precise geolocations of the profiles of the R_AW data-set. The weak correlations observed with the AW93 data-set can be explained by poor precision in the coordinates. (GRAD = slope gradient; TWI = topographic wetness index).

Table 2.6: Pearson correlation between TAs and SOC contents by layer for the AW93 dataset. p-values are given in parentheses.

	SOC_{0-30cm}	$SOC_{30-60cm}$	$SOC_{60-90cm}$
SOC_{0-30cm}		0.50 (<10-15)	0.43 (<10-15)
$SOC_{30-60cm}$			0.45 (<10-15)
ELEV	0.28 (<10-15)	0.11 (2.6×10^{-7}	0.17 (<10-15)
GRAD	0.05 (0.03)	0.00 (0.92)	0.04 (0.068)
CURV	0.04 (0.063)	0.05 (0.022)	0.05 (0.013)
SOUTH	-0.02 (0.46)	-0.03 (0.16)	-0.02 (0.37)
TPI32	-0.03 (0.097)	-0.06 (0.0039)	-0.07 (0.00049)
TPI128	-0.03 (0.13)	-0.14 (6.2×10^{-12})	-0.14 (2×10^{-11})
TPI512	-0.01 (0.71)	-0.14 (4.7×10^{-11})	-0.13 (5.3×10^{-10})
TWI	0.00 (0.91)	0.06 (0.0018)	0.08 (0.00011)
STR	0.09 (2.9×10^{-5})	0.08 (0.00027)	0.14 (6.8×10^{-11})

Table 2.7: Pearson correlation between TAs and SOC contents by layer for the R_AW dataset. p-values are given in parentheses.

	SOC_{0-30cm}	$SOC_{30-60cm}$	$SOC_{60-90cm}$
SOC_{0-30cm}		0.56 (2.8×10^{-12})	0.46 (4×10^{-8})
$SOC_{30-60cm}$			0.73 (<10-15)
ELEV	-0.06 (0.47)	-0.05 (0.58)	-0.05 (0.24)
GRAD	-0.34 (8.4×10^{-5})	-0.26 (0.0027)	-0.37 (1.3×10^{-5}
CURV	0.31 (0.0004)	0.45 (7.2×10^{-8})	0.37 (1.1×10^{-5})
SOUTH	-0.27 (0.0023)	-0.10 (0.25)	-0.05 (0.55)
TPI32	-0.28 (0.0011)	-0.45 (9.5×10^{-8})	-0.38 (9×10^{-6})
TPI128	-0.19 (0.033)	-0.42 (6.7×10^{-7})	-0.38 (7.2×10^{-6})
TPI512	-0.02 (0.86)	-0.28 (0.0013)	-0.28 (0.0011)
TWI	0.40 (2.9×10^{-6})	0.50 ($1.9 \times 10^{9-}$)	0.55 (1.2×10^{-11})
STR	0.17 (0.053)	0.34 (8.7×10^{-5})	0.27 (0.0016)

used to estimate its SOC profiles. The weaker precision in AW93 SOC content decreased the correlation between layers, while the Monte Carlo-based method developed to overcome the uncertainty in the position of the horizons had a smoothing effect, and this may have increased the correlation between layers.

We then considered the correlations between TAs and SOC layers, as preliminary information for the spatial prediction of SOC from topography. Results differed widely between AW93 and R_AW. In AW93, the correlations were generally weak, with the maximum r of 0.28 being achieved between elevation and top-layer SOC. The effect of elevation may be explained by the smooth variations of texture inside loam belt, because the areas in the north and the east have, on average, lower altitudes and contain less fine texture fractions. This effect was not visible with R_AW, where profiles were limited to a smaller sub-area, containing mainly silt loam. For R_AW, correlations were generally higher and all the attributes, except elevation and south aspect, had a highly significant correlation with the SOC content of at least one layer. The largest correlation was between wetness index and the deepest SOC layer ($r = 0.55$).

When the AW93 dataset was used, the weak correlations could be explained, at least partially, by an error in the profile positions recorded in the 1950's. This may result in an error in the TAs extracted at these profile positions. The possible sources of error are positioning errors on paper maps during the original survey, and transformation between different geoids. While the last error component was estimated at approximately 60 m (Van De Vreken *et al.*, 2011), that of location cannot. The correlations between SOC and TAs for R_AW profiles indicate the importance of precise geo-locations (Figure 2.7). For all TAs, the absolute value of the correlation tended to decrease when the size of the errors increased, until it reached a minimum plateau with some noise. However, the rate of decrease varied. With local attributes (slope or curvature), it decreased faster than with attributes related to position (TWI or large range TPI, results not shown). This analysis confirms that the positional uncertainty associated with the coordinates of AW93 may be the cause of the

Table 2.8: Model formulae and performance indices of the regression models for the three soil layers.

Model formula	RMSEP / g C kg^{-1}	ME / g C kg^{-1}	R2
$SOC_{0-30cm} \sim$ TWI - SOUTH + (ELEV × TPI_{512m})	1.28	0.01	0.21
$SOC_{30-60cm} \sim$ - (TPI_{32m} × TWI) - (ELEV × GRAD) + TPI_{32m}	1.28	<0.01	0.25
$SOC_{60-90cm} \sim$ - TPI_{128m} - GRAD + (GRAD × TPI_{128m})	0.67	0.01	0.33

weak correlation between SOC layers and TAs with this dataset. Besides the positional uncertainty, other factors that could contribute to explain the weak correlations for AW93 are likely the limited precision in the SOC content in AW93 (SOC concentration are given with a precision of 0.1 g C kg^{-1}), the fact that the profiles had to be reconstructed due to uncertainty in horizon depth and thickness (2.3) and the incidence of land-use conversions prior to the National Soil Survey.

2.3.5 Spatially explicit prediction of soil organic carbon

A summary of the regression model used to predict SOC over the study area is given in Table 7. TAs are sorted by decreasing importance (increasing p-values). All predictors were significant at $p < 0.05$. For the three depth layers, the number of selected attributes was equal to three. No model with more than three predictors (and thus selected by the stepwise feature selection method) gave a better score. Interaction terms were selected for four predictors out of nine, and square terms were not selected.

For the top layer, TWI was the first predictor, with a positive R^2. TWI was also the most correlated TA (Table 6). The selection of SOUTH as the next most important predictor may be explained by small differences in texture related to the hill-slope orientation (Goossens, 1997). The last predictor was the interaction between the elevation and the TPI with a range of 512 m. This may be because of small variations of texture correlated to the altitude. The terms of TWI, TPI with different radii, ELEV and GRAD were used, directly or inside interaction terms, in the models predicting SOC for the two deeper layers.

The models predicted respectively 21 % (0–30 cm) , 25 % (30–60 cm) and 33 % (60–90 cm) of the SOC variability (Table 7). The larger R^2 of the mid-layer compared with the top layer, despite a similar RMSEP and a smaller mean SOC content, is explained by its larger variance. The mean error was

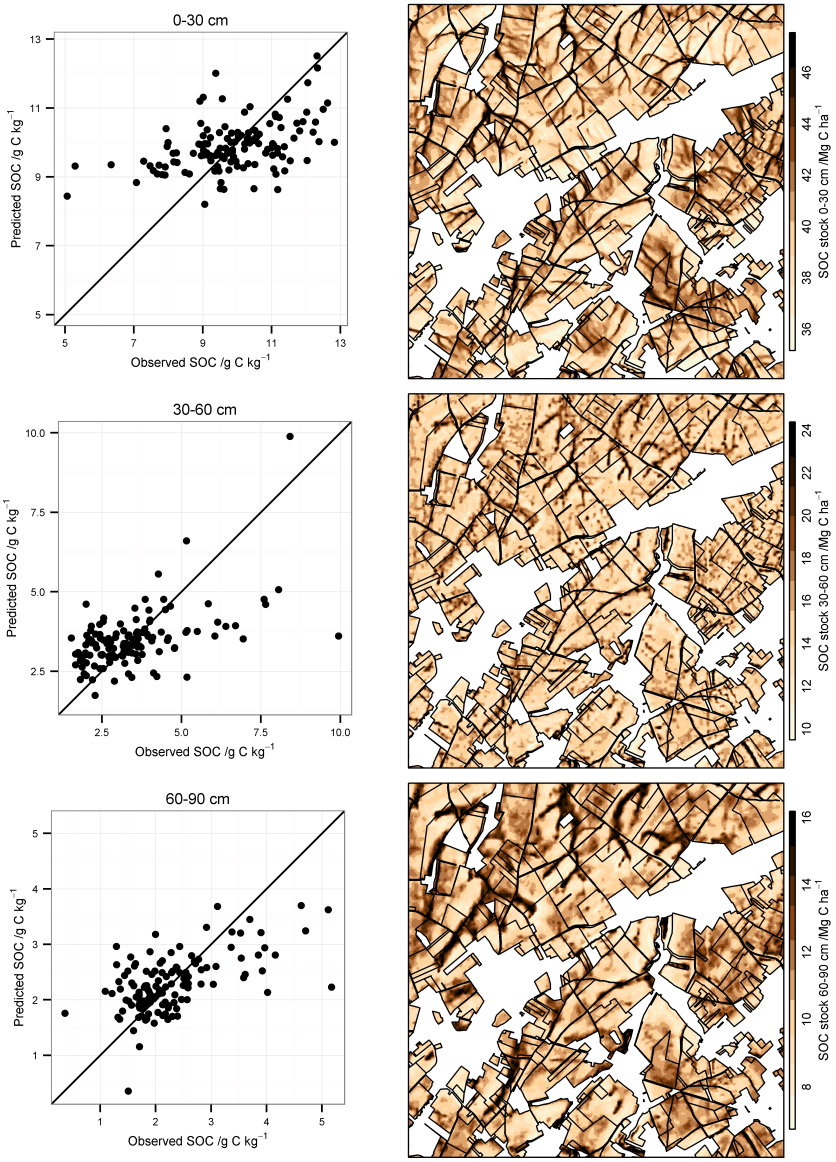


Figure 2.8: Results of the ten-fold cross validation of the R_AW (left-hand column) and mapping of the predicted SOC stocks inside the fields of the plotting area (right-hand column) for each layer. Solid line is the 1:1 line.

negligible (about 0.01 g C kg^{-1}) for all layers, confirming that the model is not biased. Points at the extremes of the range of SOC values have, in general, the larger errors (Figure 2.8, left column).

The fact that the prediction is better for deeper soil layers contradicts what has been observed in other studies (Minasny *et al.*, 2006; Kempen *et al.*, 2011). This may be because of particularities in either the study area or the prediction method. In our study area, the role of topography is particularly important since the region combines slopes and intensive tillage, and the redistribution of sediments mainly influences the sub-soil SOC dynamics through burial of SOC rich sediments. Further, as we focus on the cropland of region with homogeneous soils, we do not examine other factors which mainly influence the top-soil layers such as soil type, vegetation or land-use, as often occurs in other studies by using soil maps or remote images.

The predicted SOC layer contents were mapped into the 5 km $\times$ 5 km plotting area (Figure 2.8, right column). For the three layers, the spatial patterns of SOC were very heterogeneous, even within a field, which agrees with field or airborne observations (Stevens *et al.*, 2010). Convergent positions and valley bottoms contain more SOC than other landscape positions, and this effect is stronger for deeper layers.

The large part of unexplained variability may in variable proportions result from three factors. First, although the error in the SOC content of R_AW is limited (RMSE = 1.22 g C kg^{-1}) it could explain a small part of unexplained variability. The precision in TAs is also limited due to the support of DEM (10 m). Second, the model structure may not represent all the complexity and interaction of the processes controlling SOC. In their conceptual model, Rosenbloom *et al.* (2006) observed a sharp transition of SOC content between erosional and depositional surfaces, while Florinsky *et al.* (2002) observed a bimodal distribution for deep SOC. Finally, other factors potentially influencing SOC were not considered as exhaustive data was not easily available over the study extent. These factors could be small variations of texture within texture classes, differences in management practices or crop rotations between farms and influence of non-cropped landscape features on adjacent fields (Viaud *et al.*, 2010a) which would be difficult to determine.

Comparisons of our spatial prediction results with other studies should be made with caution. In our case, the study area was chosen to highlight the effect of topography in cropland. Usually, spatial predictions over regions are not restricted to one land use and soil type in which case the range of variability of SOC is much larger, the relative effect of hardly observable landscape scale processes is reduced, and additional sources of information (vegetation images, land use, soil type maps, can be successfully used (Vasques *et al.*, 2010).

Lastly, if in our case erosion-deposition processes are represented implicitly by TAs, other models exist which represent these processes explicitly, and the associated temporal evolution of the topography. These landscape evolution models (LEMs) use more flexible expressions for erosion, deposition and sediment transport processes, and run on a time-step basis. Their calibration

requires a large number of data, and for these reasons these models are generally applied to regions not larger than a catchment. However, some LEM have been calibrated and validated for millennia on a catchment in the loam belt (Temme *et al.*, 2011). Because theregion has relatively homogeneous soil properties and agricultural management, the model parameters derived in this catchment should be at least partially valid for all the region and could then give an estimate of the current sediment transport patterns. This could be related to SOC by empirical relationships, or integrated into a SOC dynamics model accounting for erosion.

2.4 Conclusions

We conclude that topographical information, in relation to erosion and deposition processes, allows us to envisage the spatial prediction of SOC at regional scale but that a large part of the variability still remains unexplained. The control of topography on SOC content increases with depth explaining up to 33 % of the variability in SOC content in the 60–90-cm depth layer. This is substantially less than values reported for studies conducted in single fields or micro-catchments.

Different strategies could be considered in order to improve prediction. In particular, for the top-soil layer, additional data could be used such as information from additional top-soil samples or remotely sensed images (Stevens *et al.*, 2010). The relationship between top-soil and deeper SOC could also be used to improve the prediction of deeper SOC content from the additional data. Finally, the use of landscape change models to estimate erosion-deposition patterns better could be envisaged. Our study also showed that precise geo-location of soil profiles is essential for spatial prediction in the landscape. The errors in the geo-locations of the legacy dataset AW93 probably explains why good correlations between SOC content and TAs are not observed for this dataset.

Chapter 3

Separating scales of variability to explain spatial distribution of soil organic carbon

---

This Chapter is based on the following paper: Stevens, F., Bogaert, P., and van Wesemael, B. 2015b. Spatial filtering of a legacy dataset to characterize relationships between soil organic carbon and soil texture. Geoderma, 237–238:224–236

Abstract.

The spatial distribution of soil properties often displays complex and multi-scale patterns of variation. It results from multiple soil processes acting simultaneously but at different scales. Hence, characterizing the influence of a given controlling factor on the soil property is made more difficult by the variation due to other controlling factors. In this context, separating the variation of the soil properties by spatial scales could allow disentangling the combined effect of controlling factors and would provide a qualitative and quantitative characterization of controlling factors separately. In this chapter, geostatistical tools have been used to separate the scales of variation of two soil properties (i.e. SOC and texture) coming from a legacy dataset in loam belt. Scale components were predicted separately and the relationships between soil properties were analyzed at different scales. Results illustrated that the contents of a given soil property in different depth layers were typically more correlated when only the long range components were compared. Similarly, the link between SOC and texture components was also clearer for the long range components. This means that soil processes acting at local or landscape scale influence soil properties differently according to their nature or to the depth considered. Eliminating the variation at this scale allows to better characterize the relationships between depth layers and soil properties. The study gives insights for further spatial mapping of SOC

by focusing on more appropriate variables at specific spatial scales. Furthermore, we raise the interest of spatial filtering for detecting inconsistencies inside composite datasets.

3.1 Introduction

The spatial patterns of soil properties are scale dependent as different processes dominate at different scales (Heuvelink, 1998). This is well illustrated by the spatial variation of Soil Organic Carbon (SOC). At continental or global scale, the variation is high and largely follows the variation of latitude because of climatic control (Minasny *et al.*, 2013). Geologic and pedologic properties play also a major role. In a global inventory, Batjes (1996) showed large differences of SOC content between FAO-UNESCO soil unit. In contrast, at the field or landscape scale, dominant processes are different and often of very diverse natures (Viaud *et al.*, 2010b). At a distance of a few meters, SOC variation is very large and can reach the same order of magnitude as the variation over a whole region (Goidts *et al.*, 2009a).

Soil models are generally developed to explain and to predict soil properties at a specific scale, so their transposition to other scales may not be guaranteed. The relationship between different soil/environmental properties depends on both the extent and the support of the analysis. For example, Corstanje *et al.* (2007) measured correlation between soil properties that were dependent on the extent of the sampled area. Vasques *et al.* (2012b) showed that both the spatial extent and the support of the input data had a substantial effect on the calibration parameters of a SOC predictive model. Therefore, using a model calibrated at a given extent (resp. support) to predict SOC at another extent (resp. support) generally dramatically decreases its performances and questions its validity.

In the perspective of disentangling complex and interlaced soil processes, methodologies separating scales of variability could help to highlight individual processes controlling soil properties. The variation coming from secondary controlling factors could be filtered, and the relationship between the soil property and the main controlling factor would thus appear more clearly. For example, Biswas *et al.* (2013) separated soil water storage data along a transect in various scale components and showed, among others, that short range components were positively correlated to relative elevation, while it was not the case for the original data. Wieland *et al.* (2011) applied different methods to filter a digital elevation model and found that the correlation between topography and depth of carbonates was optimal when only components with a wavelength between 60 m and 380 m were considered.

Different techniques of spatial decomposition may be applied and the choice of the method may be influenced by (1) the density, location (e.g.: scattered vs. grid data) and support (e.g.: point for field sampling vs. pixel for remote

imagery) of the variable of interest and auxiliary data, and (2) the objective of the study (e.g. to understand a particular process, or to maximize the spatial prediction accuracy). Increase of pixel size by aggregation or interpolation from scattered points to a regular grid are typical cases where small scale information is implicitly filtered out. Other methods decompose explicitly the mathematical space of the soil properties into a set of basis functions. These may be wavelets in the case of Multi Resolution Analysis (MRA, Mallat (2009)), or sinusoidal functions in the case of Fourier analysis (Butz, 2006). The empirical mode decomposition framework (Biswas and Si, 2011) also proposes to decompose the spatial variation in a set of mode function directly extracted from the data without any preliminary assumption.

In this study, we propose to spatially decompose soil properties using kriging filtering, also known as kriging analysis. This method was first described by Matheron (1982) as a component of a bigger methodology called factorial kriging analysis (FKA). Our objective is to test the hypothesis that filtering allows a better qualitative and quantitative description of relationships between soil properties. As stated by Goovaerts (1999), "accounting for the spatial scale in the study of correlations may enhance a relation between variables that is otherwise blurred in an approach where all different sources of variation are mixed, leading to a better understanding of the physical underlying mechanisms controlling spatial patterns."

3.2 Material and methods

3.2.1 Study area and datasets

To confirm our hypothesis, an analysis was carried out in the cropland of the Belgian Loess Belt in central Belgium. The reader is referred to the description of the study area in Section 2.2.1. Because of its large density of profiles for a regional dataset (2506 profiles across the whole study area, i.e. 1 profile every 4 square kilometers), we used the so-called legacy dataset "Aardewerk" as in Section 2.2.2. The dataset that was obtained from Aardewerk after calculation of the values of soil properties in the three soil layers will be referred hereafter as

Table 3.1: Variables used in the analysis. Each variables was available for three soil layers, 0-30, 30-60 and 60-90 cm. Fine represents what is usually called the fine textural fraction.

Variable	Description	Unit
SOC	Soil organic carbon	g C kg^{-1}
Sand	50 - 2000 μm texture fraction	g (100 g)$^{-1}$
Silt	2 - 50 μm texture fraction	g (100 g)$^{-1}$
Clay	0 - 2 μm texture fraction	g (100 g)$^{-1}$
Fine	0 - 20 μm texture fraction	g (100 g)$^{-1}$

Table 3.2: Main statistics of the variables used in the analysis, by soil layer. See Table 3.1 for description of the variables. Units of the statistics are the units mentioned in Table 3.1 for the corresponding variables, except for skewness which is dimensionless. Indices in the name of the soil properties represent the depth interval of the soil layer in centimeters.

	Min	Mean	Median	Max	Std deviation	Skewness
SOC_{0-30}	1.59	11.50	10.96	31.50	3.47	0.83
SOC_{30-60}	-0.97	3.44	3.04	13.95	2.05	1.38
SOC_{60-90}	-1.74	2.04	1.62	11.64	1.62	1.41
$Fine_{0-30}$	4.63	31.36	32.19	60.00	10.25	-0.20
$Fine_{30-60}$	-0.84	35.60	37.02	79.59	12.45	-0.29
$Fine_{60-90}$	0.47	36.44	38.43	90.24	14.07	-0.41
$Sand_{0-30}$	1.00	21.14	10.61	83.61	20.93	1.23
$Sand_{30-60}$	-1.21	19.63	7.92	99.22	22.97	1.39
$Sand_{60-90}$	-2.17	21.88	7.24	95.97	27.05	1.33
$Silt_{0-30}$	6.45	66.94	75.12	96.62	18.98	-1.15
$Silt_{30-60}$	1.09	64.41	72.68	94.57	20.23	-1.26
$Silt_{60-90}$	-1.64	60.86	72.07	93.40	24.33	-1.14
$Clay_{0-30}$	0.73	12.04	11.53	29.62	4.45	0.35
$Clay_{30-60}$	-0.34	15.98	15.88	49.43	6.49	0.27
$Clay_{60-90}$	-0.48	17.31	17.78	60.57	7.34	0.56

AW. These values will also be denoted as the "raw values" in the remaining part of the paper, in contrast to the filtered values. The second dataset described in Section 2.2.2, R_AW, consists in a recent resampling of 139 AW profiles inside our study area. It was used specifically here for correcting a presumed bias in AW (see Section 3.2.6).

A summary of available soil properties is shown in Table 3.1, and their main statistics are given in Table 3.2. The minimum of some variables is negative because of the spline method which has to respect both constraints on smoothness and on mean value over intervals. Although these values make no physical meaning, it was decided to keep them because it does not interfere with the spatial filtering and concerns a rather limited number of points. Histograms (not shown for the sake of brevity) suggest that the distribution of SOC, Fine and Clay at all soil layers does not depart much from a normal distribution. In contrast, the distribution of Sand and Silt looks rather bimodal.

3.2.2 Overview of filtered kriging

The main idea of filtered kriging is to decompose the spatial variation of soil properties in multiple scales, and predict scale components separately. For this purpose, soil properties were modeled as the sum of a trend (here a constant spatial mean) plus a finite number of orthogonal spatial random fields with a mean of 0. The method was first described by Matheron (1982) and

Table 3.3: Features of the scale components. Ranges in bracket represent ranges obtained with SOC after correction of a bias in the data, as explained is Section 3.2.6

Name	Range (km)	Typical spatial unit	Typical processes
Nugget	0	<10 m	Errors, intrinsic local variability
Short range	0.84 (0.72)	Field	Management, landscape features
Mid-range	6.1 (5.5)	Valley, watershed	Wetness, water erosion
Long range	28	Agricultural regions	Texture, climate

Galli *et al.* (1984) where it was combined with principal component analysis (PCA) and linear model of coregionalization (LMC) inside the methodology FKA (Goovaerts, 1992). Filtered kriging was used as is by Wen and Sinding-Larsen (1997) and Oliver *et al.* (2000) for filtering noise in images. It was also used inside the FKA methodology to explain the spatial distribution of soil water or chemical properties (Goovaerts *et al.*, 1993; Goovaerts and Webster, 1994; Bourennane *et al.*, 2012). Filtered kriging assumes stationary spatial variation, when this is not necessary with other methods like multiresolution analysis (MRA). On the other hand, filtered kriging allows more flexibility in the choice of the scales of variations.

3.2.3 Selection of the scales

In the shape of the empirical semi-variogram of SOC for the three depth layers, we could identify at first glance two breaks in the slope (cf Section 3.3.1). It seems to us that the first break appears for the three layers while the second is mainly present for the 60-90 cm layer. Besides, the shapes of the empirical variograms of SOC seem to start above zero and to slowly reach a plateau towards large distances ($>$ 30 km). Therefore, we made the working hypothesis that the model of variation could be represented by four components characterized by different ranges.

The semi-variogram model was thus decomposed into four components, namely a nugget component and three spherical models, i.e.

$$\begin{aligned}\gamma(||\mathbf{h}||)(\boldsymbol{\sigma},\boldsymbol{\theta}) = Nug(||h||;\sigma_0) &+ Sph(||h||;\sigma_{SR},\theta_{SR}) \\ &+ Sph(||h||;\sigma_{MR},\theta_{MR}) \\ &+ Sph(||h||;\sigma_{LR},\theta_{LR})\end{aligned} \tag{3.1}$$

where SR, MR and LR stand for short, mid and long range,

$$\boldsymbol{\sigma} = [\sigma_{Nug},\sigma_{SR},\sigma_{MR},\sigma_{LR}] \tag{3.2}$$

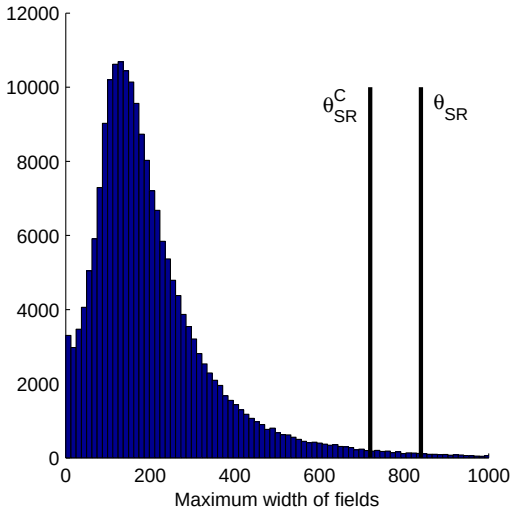


Figure 3.1: Histogram of the maximum length of the fields of the study area. θ_{SR} and θ_{SR}^{C} represent the fitted short range before and after correction of the data. Values suggest that intra-field variability is largely included in the short range component of variation.

is the part of the variance explained by each submodel and

$$\boldsymbol{\theta} = [\theta_{SR}, \theta_{MR}, \theta_{LR}] \tag{3.3}$$

is the vector of corresponding range parameters. For the semi-variogram of the fine textural fraction, despite the fact that the breaks in the slope were not apparent, we decided to use the same model structure and the same range parameters.

The spherical model was selected because, among widely used models (Gaussian, exponential, spherical), it is the only one that strictly reaches a plateau for a given distance (i.e. the range θ). In contrast, the exponential and Gaussian models are asymptotically reaching the variance and the range θ is defined for purely descriptive reasons (i.e it corresponds to the distance at which 95 % of the variance is reached). Therefore, the use of spherical models reduces numerical problems that occur when simultaneously fitting the sills of the variogram model components.

The selected ranges were expected to be partly related to environmental processes, and thus to be of the same order of magnitude as the spatial features associated to these processes (see Table 3.3 for a summary). The nugget component represents in theory the point scale variability. In practice, it represents the variation at a distance too small to be estimated from the empirical semi-variogram, since there are not enough pairs of points so closely located for allowing reliable estimations. In our case, this distance is about 50 - 100

meters and the nugget component thus represents the fine scale variability. Additionally, the nugget component also incorporates the error of analysis and the effect of the lack of precision in the coordinates of AW profiles (see Stevens *et al.* (2014)). The SR component is expected to represent the variation due to processes at the field and at landscape scales. These are, for example, erosion and deposition of sediments, the effect of non agricultural features (hedges, banks, ...) or the effect of differences of agricultural practices between adjacent fields. The short range is typically of the size of the largest fields of the study area, as represented in Fig. 3.1. The mid-range component is an intermediate level representing the variability inside catchments, which we expect to be related to difference in wetness and intensity of water erosion. The mid-range is roughly of the same order of magnitude as the width of the valleys. Finally, the long range component represents difference between subregions due to slowly varying processes, like the smooth textural differences in the original deposits of loess, or hydrologic/climatic differences due to large topographic features.

3.2.4 Fitting of the ranges

The fitting procedure for the six parameters σ_0, σ_{SR}, θ_{SR}, σ_{MR}, θ_{MR} and σ_{LR} was made at two different levels. First, for both θ_{SR} and θ_{MR}, a suite of candidate values was created. Then, for each combination of candidate values, the four $\sigma_.$ parameters were fitted on the empirical variograms of SOC by the Levenberg-Marquardt-Fletcher Algorithm (LMF). The range parameter θ_{SR} was tested by intervals of 20 m between 200 m and 2 km and θ_{MR} by intervals of 100 m between 1 km and 10 km. The fact that candidate values of θ_{SR} and θ_{MR} were tested among a wide range of values is a way to minimize the effect of arbitrary choices in the fitting procedure. θ_{LR} was fixed to 28 km because we observed that the empirical variograms of SOC became flat around this distance. Moreover, we checked how the fixed value of θ_{LR} influenced the fitted value of θ_{SR}, θ_{MR} and the $\sigma_.$'s and we observed that this influence was negligible at least in the interval 25 km $< \theta_{LR} <$ 80 km (results not shown). By contrast, adding θ_{LR} in the fitted parameters of the LMF was not feasible as it led to singularity problems. The sum-of-squares (weighted over the three depth layers) were thus produced for each combination of θ_{SR} and θ_{MR} and the values of θ_{SR} and θ_{MR} with the smallest sum-of-squares were selected, along with corresponding values for the $\sigma_.$'s.

The quantity $\mathrm{SS_T}(\boldsymbol{\theta})$ representing the departure of each model from the empirical variograms of SOC for the three soil layers was calculated as

$$\mathrm{SS_T}(\boldsymbol{\theta}) = \frac{\mathrm{SS}_{0-30}(\boldsymbol{\theta})}{\sigma^2_{0-30}} + \frac{\mathrm{SS}_{30-60}(\boldsymbol{\theta})}{\sigma^2_{30-60}} + \frac{\mathrm{SS}_{60-90}(\boldsymbol{\theta})}{\sigma^2_{60-90}}$$

where, for SOC in the ith soil layer, SS_i stands for the sum of squares of the residuals and σ_i^2 stands for the sample variance (i.e. unbiased estimator). Fitting on the three depth layers together was important because we wanted to use the same range parameters for all depth layers. The use of the variance in the denominator aims at giving the same weight to each soil layers, since the

values of the variograms of SOC (and thus the corresponding mean squares) are typically smaller for deeper soil layers.

3.2.5 Spatial filtering

Once the ranges were selected, the semi-variogram models were fitted by the least square method using the modified LMF. As well, this has been done for all soil properties. The distances used to compute the empirical variograms were not regularly spaced but were chosen so that the jth lag contain $50 + 200 \times j$ pairs of points. This choice was intended to provide a finer resolution at small distances (useful for the fit of σ_0, σ_{SR} and θ_{SR}) and, compared to regular spacing, to increase the number of observations for each point of the empirical variogram as distance increases (useful for the fit of σ_{MR}, θ_{MR} and σ_{LR}). A constraint was added in order to avoid negative sills for the submodels. Even though we assumed the same ranges for all soil properties, all the submodels associated with these ranges were not necessarily relevant in each case. Therefore, a submodel was included in the variogram model of a soil property at a given soil layer only if it resulted in a lower Akaike's Information Criterion (AIC) for the global fit. The AIC for a soil property in a given soil layer is calculated according to Burnham and Anderson (2002) with

$$\mathrm{AIC} = n \ln(\mathrm{SS}/n) + 2k$$

where n is the number of distance classes in the empirical variogram, SS is the sum of squares of the residuals of the fit and k is the number of ranges included in the model. The relative part of the variance explained by each submodel was then interpreted according to the current knowledge of soil processes and to the specificity of the study area.

In order to map the individual scale components, a regular grid of 1 km per cell was superimposed over the study area. Scale components of all soil properties were predicted both at the grid nodes and at the original sampling locations, along with their kriging standard deviation. The nugget component was only predicted at the original sampling location since outside these points, the nugget covariance is zero and the prediction would result in the (constant here) spatial mean. For each component, the kriging weights were estimated using only one submodel of the variogram at a time. Although only samples located in cropland were used for the analysis, the scale components were mapped over the whole study area and no cropland mask was applied. Therefore, for land uses other than cropland, the map represents the value of the scale component that would occur if the land use had been cropland. Once the maps were produced, an interpretation was sought for the most remarkable spatial patterns that were observed.

It is worth emphasizing that the spatial filtering aims not at giving the best prediction at each point. For this purpose, the kriging estimator, as the best linear unbiased predictor, would give better results. In this respect, the sum of the filtered components plus the constant mean is precisely equal to the measured values at the sampling points and to the ordinary kriging estimator

everywhere (see (A.1)). Instead, the aim of the filtering method here is to decompose the spatial variation in multiple components which represent the variation at specific scales. The kriging standard deviation of a scale component still represents the error that would be done if the true spatial variation was estimated with this scale component.

Finally, soil variables were plotted against other soil variables for the raw values and separately for each spatial scale component too. Weighted Pearson correlation coefficient between any variable or scale component y and z was calculated with

$$r = \frac{\sum_i w_i(y_i - \mathbf{w}^\top \mathbf{y})(z_i - \mathbf{w}^\top \mathbf{z})}{[(\sum_i w_i(y_i - \mathbf{w}^\top \mathbf{y})^2)(\sum_i w_i(z_i - \mathbf{w}^\top \mathbf{z})^2)]^{1/2}}$$

where $\mathbf{w}$ is a vector of normalized weights defined as $w_i = \sigma_i^{-2} / \sum_j \sigma_j^{-2}$. When the correlation coefficient was clearer for a specific spatial component than for the original values, the implications of this result are discussed.

In the maps and scatterplots of scales components (Figures 3.4, 3.6, 3.8, 3.10, 3.11), the constant mean was added to the scale component. Since scales components are centered on zero, this centers the displayed scale components on the mean and increases its physical meaning. For simplicity, the sum of the scale component and the mean is referred abusively to the scale component itself in the caption and discussion of these figures.

3.2.6 Correction of a suspected bias in soil organic carbon values

Although this was not the initial objective of the study, first results indicated the presence of a probable bias in the SOC observed for a group of profiles sampled by a specific surveyor. Since no obvious explanation could be found, it was decided to correct it in a pragmatic way in order to improve further analyses. The R_AW dataset was used because it contains resampled profiles that are in the region initially sampled by this specific surveyor, along with resampled profiles that are not. Although six decades separate the initial surveys and the resampling, the transformation relies on the assumption that, on average, the SOC content of the profiles in the problematic region and the SOC content of the profiles in the rest of the area have changed similarly during this time. The correction factor was calculated as

$$f = \frac{\mu^*_{R_AW} / \mu^*_{AW}}{\mu^{!*}_{R_AW} / \mu^{!*}_{AW}} = \frac{f^*}{f^{!*}}$$

where $\mu^*_{R_AW}$ is the mean of the SOC content of the 61 R_AW profiles corresponding to AW profiles sampled by the specific surveyor and μ^*_{AW} is the mean of the SOC content of these AW profiles. $\mu^{!*}_{R_AW}$ and $\mu^{!*}_{AW}$ are the equivalent for the 78 R_AW corresponding to AW profiles originally sampled by any other surveyor. Figure 3.5b represents the location of the R_AW profiles. Green triangles are the profiles used in $\mu^*_{R_AW}$ and blue circles are the profiles used in

$\mu^{!*}_{R_AW}$. In this definition, the SOC content of the profile is calculated as the sum of the SOC of the three soil layers. Therefore, the same correction factor was used for the three soil layers. f^* is thus an estimate of the evolution of SOC between $\approx$ 1955 and $\approx$ 2012 for profiles sampled by the specific surveyor and $f^{!*}$ an estimate of that evolution for the other profiles. If these values were largely different, and no explanation could be found to explain these different behaviors, this would add some credibility to the hypothesis that some profiles in AW were originally biased. Finally, the procedure of range selection, semi-variogram fitting and filtered kriging were performed again for corrected SOC and the impacts of the correction on the results were assessed.

3.3 Results and discussion

3.3.1 Selection of the ranges

The normalized residual sum of squares for the fit of variogram models of SOC for the three soil layers was computed for all the combinations of θ_{SR} and θ_{MR}. Figure 3.2a shows clearly the presence of a unique and thus absolute minimum inside the domain of variation, at the values of θ_{SR} = 840 m and θ_{MR} = 6.1 km. The uniqueness of the minimum indicates that the ranges representing the variability of SOC are coherent at the three soil layers.

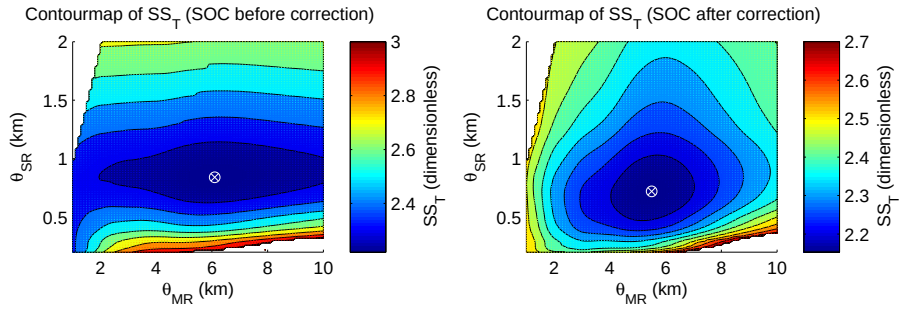
3.3.2 Soil organic carbon

In the following sections are presented the semi-variograms of SOC and the maps for the long range component after the spatial filtering. These maps allowed us to identify the presence of a probable bias in AW SOC data. The results of the correction of the bias, as explained in Section 3.2.6, are given. Then, the maps and semi-variograms for the corrected SOC values are displayed.

Results before correction

For the three soil layers, the model shows a proper fit of the empirical variogram over the entire domain (Figure 3.3). The nugget and short range components together represent the main part of the variation (Table 3.4). The long range component is mostly present in the top layer, with up to 24 % of explained variance, while the mid-range component is mostly present in the 60-90 cm layer, explaining up to 15 % of the total variance. These specific results are not discussed here because we assume they are partly caused by a bias in the data. The variograms of SOC after correction are discussed instead in Section 3.3.2.

The map of the long range component of SOC (and the associated variance of prediction) is displayed by Figure 3.4 for the three soil layers. The maps of the short range and mid-range components were not displayed because no



(a) (b)

Figure 3.2: Normalized residual sum of squares of variogram models of SOC in function of θ_{MR} and θ_{SR} representing the manual range optimization procedure for (a) SOC before correction and (b) SOC after correction. The same ranges were used and fitted for the three soil layers together, while θ_{LR} was fixed. In both cases, the contourmap shows a clear and unique minimum. The difference of optimal θ_{SR} and θ_{MR} between the two cases is small.

Table 3.4: Relative part of the variance attributed to each spatial component. Soil properties are given by depth layer when applicable. Total for each row is 100 %. The values in parentheses represent the relative part of the variances of *SOC* before the bias related to the surveyor was corrected. Blank cell occur if the submodel was removed from the model before the fit, for the corresponding scale.

	Nug	SR	MR	LR
SOC_{0-30}	.31(.40)	.40(.34)	.06(.12)	.24(.14)
SOC_{30-60}	.54(.58)	.28(.27)	.04(.09)	.15(.06)
SOC_{60-90}	.41(.40)	.31(.34)	.15(.26)	.13
$Fine_{0-30}$	.38		.09	.53
$Fine_{30-60}$	.58			.42
$Fine_{60-90}$	.54		.09	.37
$Sand_{0-30}$	.14	.18	.14	.54
$Sand_{30-60}$	.42		.16	.43
$Sand_{60-90}$	.48		.10	.43
$Silt_{0-30}$	.15	.18	.15	.52
$Silt_{30-60}$	.43		.15	.41
$Silt_{60-90}$	.49		.06	.45
$Clay_{0-30}$	.54		.11	.34
$Clay_{30-60}$	.72		.03	.25
$Clay_{60-90}$	.57	.07	.12	.23

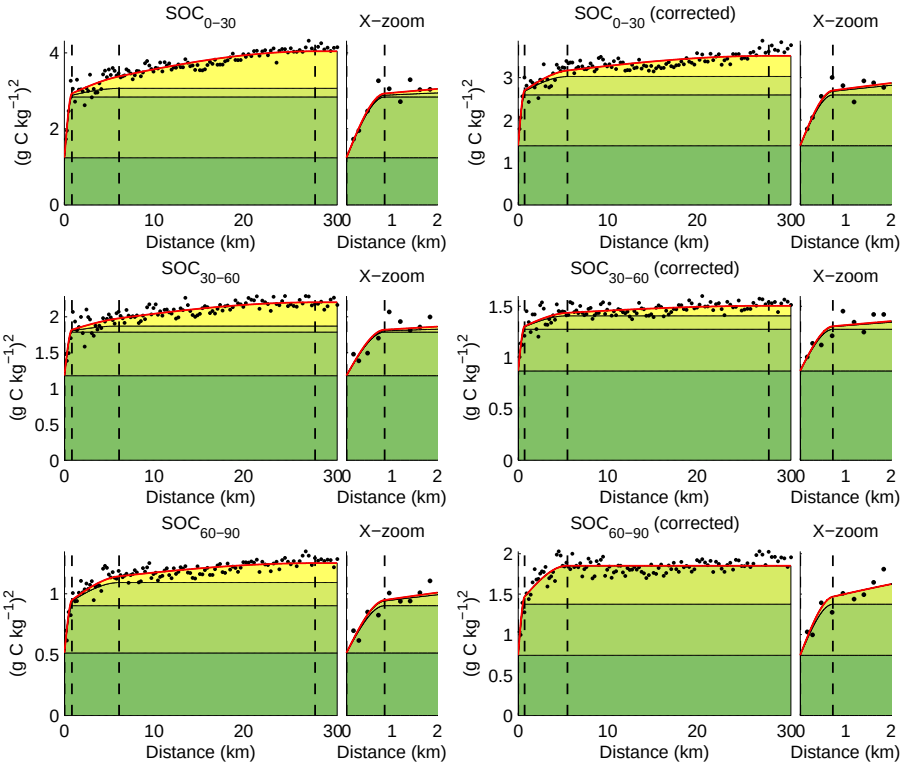


Figure 3.3: Semi-variogram and fitted model of *SOC* before correction (left column) and after correction (right column), for the three depth layers. Filled areas represent the partition of the variance between components. Vertical dashed lines indicate the ranges of the various spherical submodels used in the variogram model. Right-hand side of each graphic is a zoom over the left part of the distance axis.

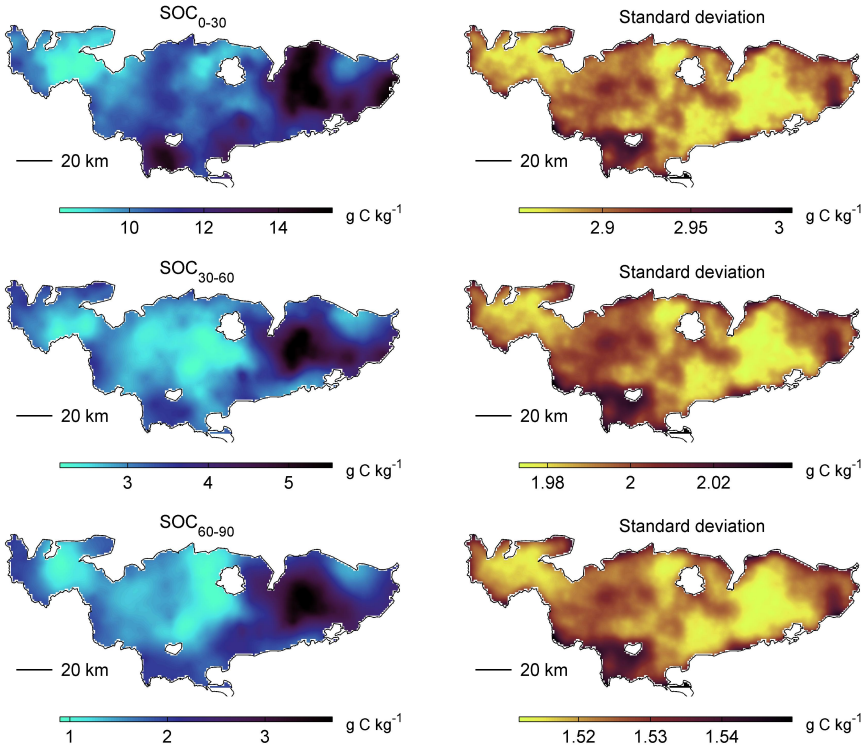
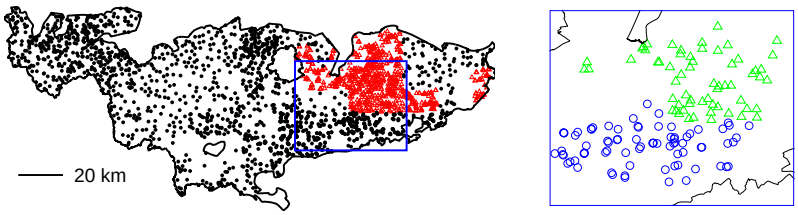


Figure 3.4: Spatial prediction (left) and associated standard deviation (right) for the long range component of *SOC* for each depth layers.



(a) (b)

Figure 3.5: (a) Positions of the AW profiles that were historically sampled by a given surveyor (red triangles) among other profiles (black dots). The area where the profiles of the first category are located coincides with an area where the long range component of SOC displays particularly high values (see Figure 3.4). This suggests the presence of a bias related to that surveyor. (b) Region corresponding to the blue rectangle in (a) with the positions of the 139 R_AW profiles which were used to calculate the correction factor. Green triangles are resamplings of AW profiles that were originally sampled by the specific surveyor while blue circles are resamplings of AW profiles that were originally sampled by any other surveyor.

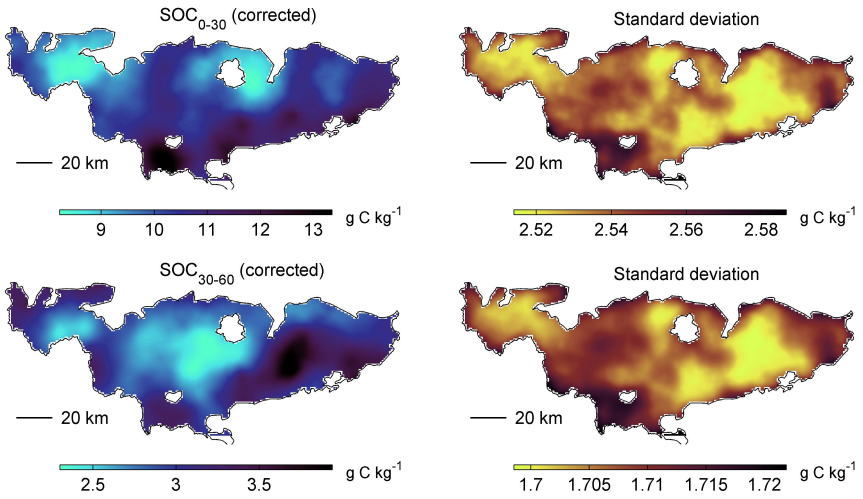


Figure 3.6: Spatial prediction (left) and associated standard deviation (right) for the long range components of SOC for each depth layer after correction for the bias related to a given surveyor. The long range component was not computed for SOC_{60-90} because the long range submodel was omitted from the variogram model, since it did not improve the modeling according to the value of AIC.

useful interpretation of the spatial patterns could be done. These maps show juxtapositions of a large number of patches of lower or higher values with a width approximately equal to size of the range. No obvious spatial correlation with geographic features was observed. This may be due to the fact that a large part of the area is not cropland, in particular in the lower part of the valleys, and the fact that the density of samples is still limited (around 1 sample for 4 km^2). For a specific prediction of the spatial variations of SOC at the field or catchment scale at high resolution, larger sampling densities are needed, so that the studies are generally limited to smaller study areas (Lacoste *et al.*, 2014).

The maps of the long range component show a large area of higher SOC content in the eastern part of the loam belt. The pattern is similar at the three depth layers, and may be described as a patch of about 20 km radius around the city of Tienen. In the deepest layer, the SOC content in the patch corresponds to twice the average SOC content outside the patch. Besides, the sampling density is particularly large in this area, resulting in lower prediction variances. For all depth layers, this area of large values shows a large similitude with the area that was sampled by a specific surveyor (Figures 3.5a). This seems to indicate the presence of a bias in the data linked to the surveyor.

Correction of the dubious profiles

The suspected bias in SOC values related to the specific surveyor was corrected as described in Section 3.2.6. The SOC content of the 706 profiles from this surveyor were modified. Values of 1.58 and 1.08 were found respectively for f^* and $f^{!*}$, leading to a correction factor $f = 1.41$. This indicates that, if there was no bias in the legacy data, the SOC content in the area originally sampled by the specific surveyor would have undergone a decrease which is 7.25 (0.58/0.08) times larger than profiles in an adjacent area sampled by other surveyors, between years around 1955 and 2012. This is not really expected since the evolution of landscape and agricultural practices is relatively homogeneous for the entire study area. Therefore, this reinforces the hypothesis that the profiles of the legacy dataset AW contain a bias associated with this surveyor. The SOC values of the 710 AW profiles related to this surveyor (red triangles in Figure 3.5a) were thus divided by the correction factor.

We notice that the value of 1.41 is also close to the value of the the correction factor for incomplete oxidation of 1.33 that was historically used in the analyzes (Van Orshoven and Vandenbroucke, 1993). This may suggest that the dubious profiles of AW were erroneously multiplied twice by this factor. However, since AW is legacy data, no material is available to check that possibility in a rigorous way.

Finally, it is worth pointing out that the SOC content of resampled profiles is very different from the content of their corresponding original profiles and that no correlation was found (results not shown). This may be caused by the combination of the large intrinsic variability of SOC at short range with the error in the location of the R _AW resamplings. Besides, SOC content may

have changed since AW measurements because of local differences in agricultural management, artificial drainage or landscape features. Nevertheless, the correction procedure was based on the assumption that the SOC content of original and resampled profiles are related at least on average over the surface of the sub areas used for the correction.

Results after correction

The ranges were fine-tuned, this time from the corrected data, using again the procedure described in Section 3.2.4. The new ranges are close to the ranges that were obtained before correction, with θ_{SR} moving from 840 m to 720 m and θ_{MR} moving from 6.1 km to 5.5 km (Table 3.3). This relative stability shows that the correction does not contradicts the hypotheses about the scales of variability. Besides, the contourmap of SS_T for *correctedSOC* also displays a unique minimum (Figure 3.2b). Compared with the contourmap of SS_T before correction (Figure 3.2a), SS_T increases more rapidly when θ_{MR} moves away from its optimal value. We can interpret this as a confirmation of the legitimacy for the presence of the mid-range component.

The relative parts of the variance of the semi-variogram models for *correctedSOC* values are indicated in Table 3.4, and variograms are displayed in Figure 3.3. Compared to *SOC*, a decrease of the relative variance for the long range component was observed, from 24 % to 14 % in the 0-30 cm layer and from 15 % to 6 % in the 30-60 cm layer. The long range component is no more included in the variogram model for the 60-90 cm layer. These changes suggest that the apparent long range component of SOC was partly due to the bias related to the surveyor.

In contrast, the relative variance of the nugget and the short range component together increased for the 0-30 cm and 30-60 cm layers. These two components together represent between 74 % (0-30 and 60-90 cm) and 85 % (30-60 cm) of the total variance. This shows the importance of processes that influence SOC at a spatial scale of around 500 m or less. These may be various processes like erosion and deposition of sediments, effect of past land use, land consolidation and drainage, influence of landscape elements like hedgerows, grass stripes, mounds and ditches, differences in the use of amendments between adjacent fields, etc.

The mid-range submodel has a small relative importance except in the 60-90 cm layer when it accounts for 26 % of the variance. Since this range corresponds roughly to the order of magnitude of the width of the valleys, a possible explanation is the fact that in some slopes or valleys a substrate of Tertiary sands was frequently incorporated in this soil layer and, since it was not sampled, we considered that the sandy substrate contained no SOC.

Maps of the long range component of *correctedSOC* for the 0-30 cm and 30-60 cm layers are given in Figure 3.6. The patch with larger and dubious values of Figure 3.4 is no more visible in the top layer and is much less clear in the mid-layer than before the correction. Instead, we observe larger values

along the southern border of the study area, where the texture is finer (see Figure 3.8).

3.3.3 Texture fractions

In order to investigate the factors controlling the SOC spatial distribution at different scales, the semi-variograms and spatial predictions of textural fractions were also computed. For the sake of brevity, only the semi-variograms of Fine are displayed (Figure 3.7). The experimental semi-variograms of all texture fractions do not seem to reach a plateau at distance larger than the largest range θ_{LR}(i.e. 28 km). We attribute this to the non-stationarity of the random field representing texture fractions, since the mechanisms of deposition of the aeolian loess originally caused a gradient in the soil texture across the loam belt. To address this issue, we assume local stationarity : since the models were only fitted on the distance classes inferior to θ_{LR}, they were only considered valid within this interval, and only points within a neighborhood of radius θ_{LR} around the point of prediction were used for kriging.

The semi-variogram models of the four texture fractions that were spatially decomposed display many similarities (Table 3.4). The nugget and the long range submodel together explain at least 67 % of the variance. The short range submodel is absent in two out of three cases and is limited to 18 % when it is present. The relative variance of the long range submodel is large and superior to 37 % in all cases, except for the clay content where the nugget submodel largely prevails. The mid-range component is always small and is absent for $Fine_{30-60}$ The large variation at long range illustrates the fact that the textural variation in our area is very gradual over space, with progressively sandier textures in the north and the west parts (see Figures 3.7). Besides, except for the fine fraction, the nugget submodel dominates as we go deeper. This tends to show that the soil texture in deeper layers depends less on the original composition of the loess deposit than on processes acting at local or landscape scales. These processes are, for example, burial of sediments or the incorporation of substrate in deeper layer after the truncation of the soil profile by erosion, with preferential erosion of some texture fractions.

The maps of the long range component of the topsoil texture fractions show a dominance of sand in the north west direction (up to 60 %) along with a clay content of about 10 %, except for a stripe of larger values (up to 20 %) in the south that runs parallel to the Mosel and Sambre rivers (not shown). Silt content is typically above 30 % with large areas reaching up to 75 % in the center and the south-east areas (not shown). For the long range component, the patterns for different depth layers are consistent (Figure 3.8).

3.3.4 Links between filtered soil properties

It was established that the link between soil properties at different layers depends on the scale at which we look at. Here we compare the link between

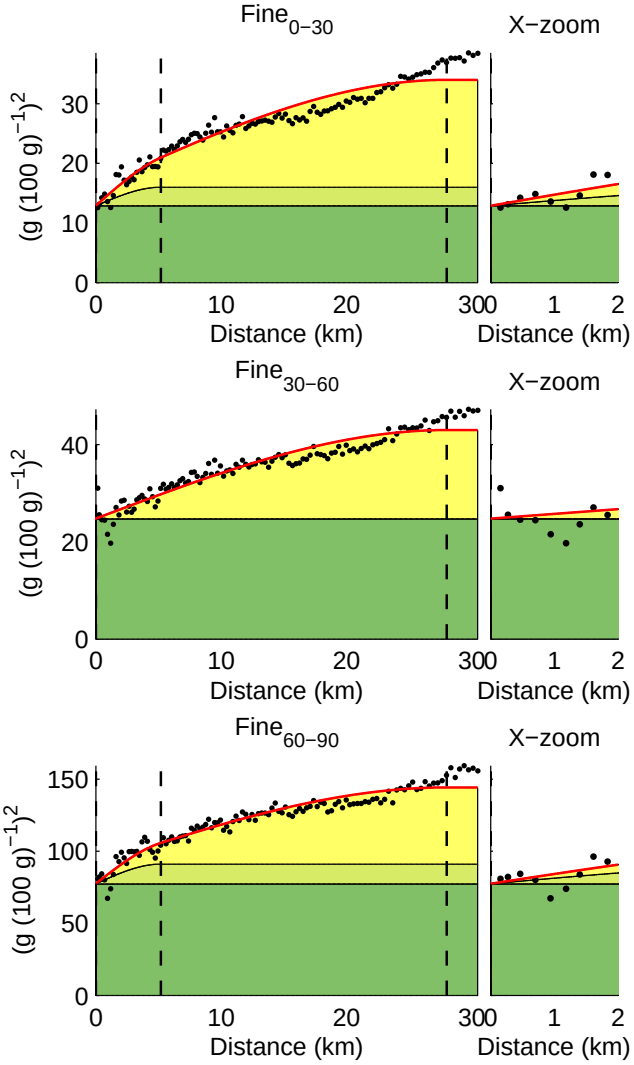


Figure 3.7: Semi-variogram and fitted model of *Fine* for the three depth layers. Filled areas represent the partition of the variance between components. Vertical dashed lines indicate the ranges of the various spherical submodels used in the variogram model. Right-hand side of each graphic is a zoom over the left part of the distance axis.

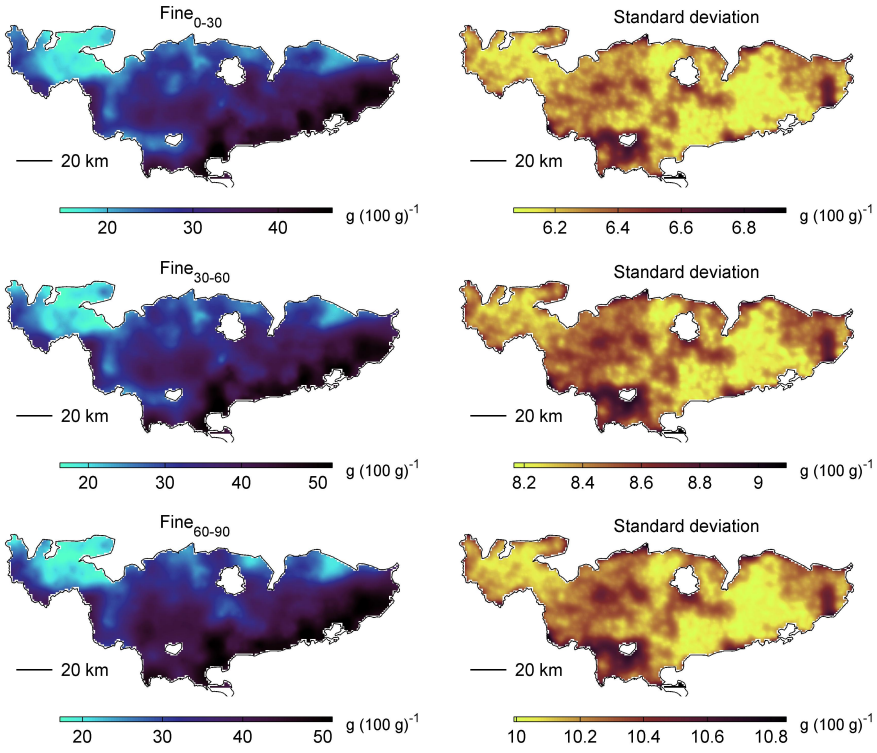
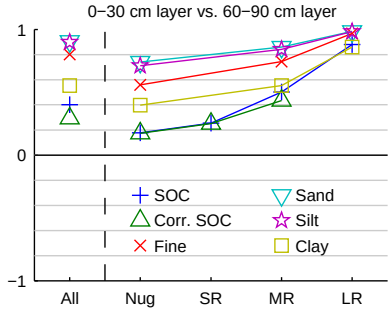
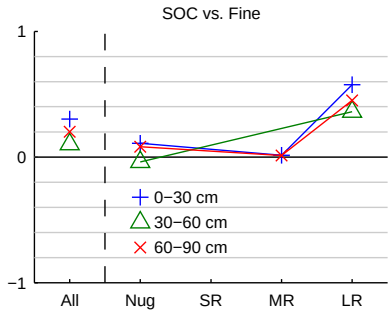


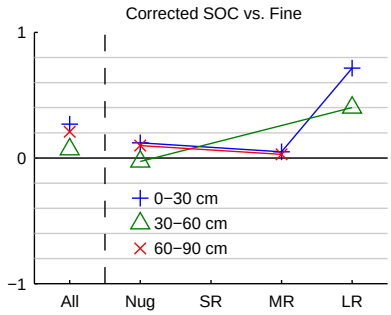
Figure 3.8: Spatial prediction (left) and associated standard deviation (right) for the long range components of *Fine* at the three depth layers.



(a)



(b)



(c)

Figure 3.9: Pearson correlation coefficients (a) between topsoil (0-30 cm) and subsoil (60-90 cm) content of soil properties (including *correctedSOC*), (b) between *SOC* and *Fine* by layer and (c) between *correctedSOC* and *Fine* by layer. Correlations are given for the raw values and component by component. SR, MD and LR stand for short range, mid-range and long range component.

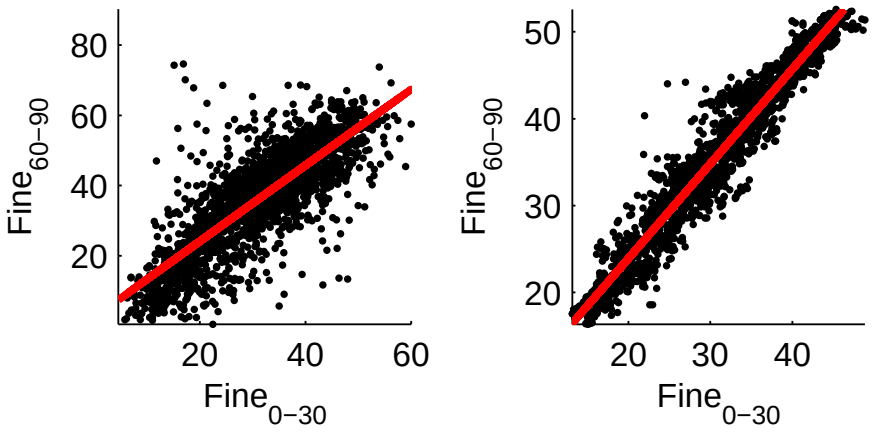


Figure 3.10: $Fine_{0-30}$ as a function of $Fine_{60-90}$ (a) for raw data and (b) for the long range components only. Red lines represent linear regressions. Correlations are stronger for the long range components.

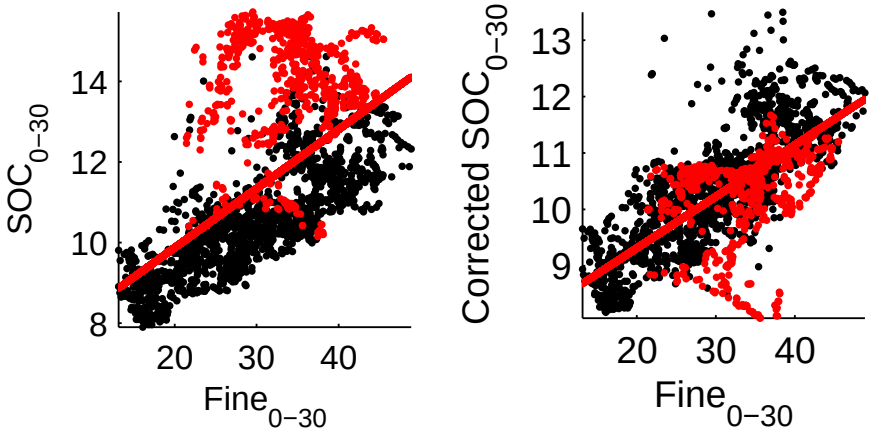


Figure 3.11: Effect of the correction of the bias on the link between (a) SOC_{0-30} and $Fine_{0-30}$ and (b)$corrected SOC_{0-30}$ and $Fine_{0-30}$. Points in red represent corrected profiles. Red lines represent linear regressions.

Table 3.5: Correlation coefficients by spatial scale between (a) the content at 0-30 cm and the content at 60-90 cm of a same soil property, between (b) the content of *SOC* and *Fine* at the same depth layer and between (c) the content of *correctedSOC* and *Fine* at the same depth layer. SR, MR and LR stand for short range, mid-range and long range component. Blank cells occur if the corresponding scale component was not filtered for at least one of the two correlated variables.

	Nug	SR	MR	LR	All
SOC	.18	.25	.50	.88	.40
Corr. SOC	.17	.25	.43		.29
Fine	.56		.74	.97	.80
Sand	.74		.86	.98	.91
Silt	.71		.84	.98	.90
Clay	.40		.55	.86	.55

(a)

ex	Nug	SR	MR	LR	All
0-30 cm	.11		.01	.58	.30
30-60 cm	-0.04			.36	.10
60-90 cm	.08		.01	.45	.20

(b)

	Nug	SR	MR	LR	All
0-30 cm	.12		.05	.71	.27
30-60 cm	-0.03			.40	.07
60-90 cm	.10		.03		.21

(c)

variables for raw values and for filtered components. In the first step, the values of a same soil property at different depth layers were compared. Then, in a second step, the values of different soil properties at the same depth layer were compared.

Link between soil properties at different depths

The correlation coefficient between the content of each soil properties at 0-30 cm and 60-90 cm is plotted for the raw values and for each available scale component separately (Figure 3.9a). Corresponding values are displayed in Table 3.5a. The correlation is always the strongest when focusing on the long range component only. The correlation coefficient between topsoil and subsoil, for the long range components, ranges from 86 % ($Clay$) to 98 % ($Sand$ and $Silt$). For the raw values, it ranges from 29 % ($correctedSOC$) up to 91 % ($Sand$). Figure 3.10 shows a scatterplot of $Fine_{60-90cm}$ in function of $Fine_{0-30cm}$, for (a) the raw values and (b) the long range component only. Results suggest that differences in the vertical distribution of soil properties between profiles are mainly due to processes acting at the slope or the landscape scale. In contrast, soil processes acting at larger scale influence soil properties in a way that is more similar between soil layers.

Link between soil organic carbon and texture at the same depth.

The strength of the link between soil organic carbon and the fine texture fractions is given for SOC values before correction (Figure 3.9b) and after correction (Figure 3.9c). Corresponding values are displayed in Table 3.5b and 3.5c. The correlation between the large scale components is always stronger than the correlation between raw values or other scale components, also if other texture fractions are used instead of Fine (not shown). The correction of SOC values increases the correlations between long range components, except in the 60-90 cm layer where corrected SOC has no long range component. Figures 3.11 illustrates the impact of the correction of the bias related to the surveyor with respect to the link between the long range component of SOC and Fine in the top layer. After the correction, the content of SOC in each point is generally lower. Besides, the points of the dubious area that were clearly grouped out off the trend line follow better the general trend after the correction (red points). Only a small part of these points show peculiarly low values after the correction, indicating that maybe not all the samples of the specific surveyor were concerned by the bias. This is possible since points from the same reviewers were sampled over different periods.

Absolute values of correlation coefficient between corrected SOC and Fine or Sand are equal or superior to 0.4 for the long range component but close or inferior to 0.1 for the other scale components. This seems to indicate that the link between SOC and texture is stronger for the top layer and on a large scale. This can be interpreted in the context of the chemical stabilization of SOC through binding to fine particles (Six *et al.*, 2002; Hassink, 1997), occurring

as a background effect. However, at the slope, landscape or catchment scale, additional processes disturb this equilibrium and moves SOC content away from the value that could be expected from texture. These processes could be for example water and human induced erosion and deposition (Harden *et al.*, 1999), differences in agricultural management or influence of non agricultural features (Lacoste *et al.*, 2014).

3.4 Discussion and conclusion

The methodology proposed in this paper allowed us to decompose the intrinsic multiscale variability of soil properties with the aim of better identifying the effect of specific controlling factors.

For all the decomposed variables, but especially for SOC, the contribution of the nugget submodel was substantial. This includes both the variation at a range inferior to the minimum sampling distance (that cannot be estimated from the empirical semi-variogram) and the effect of various possible errors, like the error due to the sampling by horizon and the error in the geolocations of the legacy dataset (with an order of magnitude of about 100 m (Stevens *et al.*, 2014)). These drawbacks are unfortunately inherent to the legacy data. Current methods of profile sampling (Doetterl *et al.*, 2013b) and geolocation would have led to smaller errors. Furthermore, the use of a more appropriate sampling design (e.g. nested sampling, Oliver and Webster (1986)) could have given a better estimation of variance at very short distances. If the exact composition of the nugget component is not exactly known, it appears anyway that for SOC the sum of the variance of the nugget and the short range (500 m) components is quite larger than the sum of the variance of the other components. This means that the variation of SOC is dominated by short range variations. On the contrary, texture is dominated by long range variations.

The large variance at short range may affect the mapping of the SOC spatial distribution at a given scale. Typically, estimations are point based. A SOC map could for example be built by estimating SOC in the center of each pixel of a grid. For pixel sizes equal to or larger than the short range, the classic kriging estimate may not reliably represent the SOC of this pixel. In contrast, block kriging would be more reliable since it considers the covariance between measurements and all the surface of the pixel. Our filtering method is an alternative to block kriging in that case, as it allows filtering the high frequency variations. The long range component estimated in the center of the pixel provides a more reliable estimate of the SOC content of this pixel. In general, the spatial component selected should account for on the resolution at which the variable is aimed to be predicted.

The choice of the resolution at which a soil variable is predicted depends on the planned applications but should also consider the type and quality of predictor data and the link between the predictors and the predicted variable. If the aim is to predict SOC at the very fine resolution of a few meters, as it is typically sought for in precision agriculture (Ladoni *et al.*, 2009), kriging

will necessitate a very dense sampling to be efficient. Given the cost of terrain campaigns, alternative methods like remote imagery (Stevens *et al.*, 2014) could be considered for fine resolution mapping of surface properties over large areas. In contrast, an application like the estimation of the subsoil SOC content from a map of surface SOC obtained by airborne spectral imagery (see Lagacherie *et al.* (2013) for corresponding application based on texture) does not require the spatial resolution of the image to be unquestioningly high. It should better match the long range component of surface SOC, which correlates better with deeper SOC.

Besides its original goal, the filtering method proved to be a simple way to identify possible anomalies in the dataset. After the filtered method was applied and the long range component was selected, the correspondence between a well-delimited area of larger values and the area historically sampled by a specific surveyor was obvious. In contrast, if kriging had been applied without filtering, the bias in the SOC content due to the surveyor would have been hidden due to the large natural short range variation. Furthermore, even though the correction of the suspected error was not optimal, it substantially moved the link between the long range components of topsoil SOC and fine textural fraction towards a relationship that is more in agreement with what can be expected from our knowledge of the process.

Chapter 4

Detecting and quantifying field-related spatial variation of soil organic carbon

---

This Chapter is based on the following paper: Stevens, F., Bogaert, P., and van Wesemael, B. 2015a. Detecting and quantifying field-related spatial variation of soil organic carbon using mixed-effect models and airborne imagery. Geoderma, 259–260:93–103

Abstract.

Evidences exist that the spatial variability of soil organic carbon (SOC) in cropland is partially controlled by environmental or human factors acting on a field basis (e.g., agricultural management, landuse history, landscape structure). However, few studies have quantified the relative importance of the fields-related variability at the regional scale. Recent airborne hyperspectral imagery methods provide SOC estimates at high resolution and over large surfaces. They may be used to quantify and explain the spatial variation of SOC. In this chapter we used a SOC hyperspectral image over Luxembourg to separate SOC variation in three components: the effect of the texture class (as defined by a texture map), the effect of fields (as defined by a cadastral map) and spatially dependent residuals. The relative variance of these components and the spatial structure of the residuals were rigorously assessed by REML. Results indicated that 65.7 ± 0.3% of the variance of SOC in the study area was explained by texture classes. The intensity of the field effect was largely dependent on the location. In some subareas of homogeneous texture class, no significant effect could be observed while in others, field explained up to 68.8 ± 12.0% of the variance. In contrast with other methods like ANOVA, the method developed here measures the variation related to spatial units (soil map units or fields) while taking explicitly into account spatial dependencies. As soon as the boundaries of these spatial units and a spatially extensive (or at least very dense) knowledge of a soil property over a

region of interest are available, it allows rigorously determining if the soil property depends on these spatial units. In the present application, findings pointed out the importance of considering fields-related variability in SOC modeling and mapping studies.

4.1 Introduction

Soil organic carbon is a property that plays a role in multiple soil functions, like nutrient cycling, water control, physical stability, biodiversity and climate change mitigation. It is itself influenced by multiple factors acting at various spatial scales, like climate, soil type, land use, land management and topography (Minasny *et al.*, 2013).

Studies that characterize the spatial variations of SOC at regional scale for cropland demonstrate an important role of the soil type. Kempen *et al.* (2010) show for example large and significant differences between the mean topsoil SOC content for different soil types in a province in the Netherlands, with the most SOC-enriched soil type (thick peat soil) containing on average 5.7 times more SOC than the least enriched type (sandy soil). Soil type (or related properties such as texture class, drainage class or parent material) is also selected as predictor in many regional SOC prediction models (Zhang *et al.*, 2012; Meersmans *et al.*, 2012b).

If we focus on cropland with relatively homogeneous soil type, the variation of SOC observed at short range (i.e. in a radius of a few hundreds of meters) reaches generally the same order of magnitude as the variations over the whole region (Goidts *et al.*, 2009a). Short range variation is complex because of the large number of possible processes that are involved. Although interactions between them may exist, there is a clear distinction between processes that act almost continuously in space and processes that are rather constant over an entire field. Continuous processes are basically the effect of water and tillage erosion (Van Oost *et al.*, 2005; VandenBygaart *et al.*, 2012) or differences in moisture and temperature conditions related to the topography and affecting SOC equilibrium (Florinsky *et al.*, 2002). Factors affecting SOC on a field basis are rather linked to agricultural management or landuse history (tillage practices, manure and amendments applications, crop rotations, residue management, former presence of grassland or forest, etc). Landscape features may also affect SOC in a way that is in some cases spatially related to the field structure like hedges or buffer zones (Viaud *et al.*, 2010b).

Processes acting at the field scale have been demonstrated in specific cases or in field experiments and they are widely implemented in SOC modeling tools. A field divided into multiple plots undergoing different treatments may display large and significant differences in the SOC content after several years (VandenBygaart *et al.*, 2003a). The long-term effect of land use/land management changes has been studied in many geographical contexts and SOC dy-

namics models have been coupled to land use evolution models to understand past changes or estimate future changes in SOC carbon stocks (van Wesemael *et al.*, 2010b; Smith *et al.*, 2005). However, evidence of spatially explicit and systematic differences in SOC content between fields for a given landscape is sparse in the literature. Large scale SOC dynamics studies are extrapolations based on field experiments and local surveys (Meersmans *et al.*, 2012b). Digital soil mapping studies, for their part, are mainly divided between precision agriculture studies and regional mapping studies. The first ones generally focus on individual fields (Casa *et al.*, 2012) while the second ones do not generally aim at mapping SOC at such a fine resolution. Moreover, observations in the landscape are not always sufficient to assess the presence of field-related variation in a landscape and sound statistical methods are often needed for consolidation.

Nevertheless, differences between the mean SOC content of contiguous fields have been mentioned in SOC images from soil or airborne spectroscopy (Stevens *et al.*, 2006). Hyperspectral imagery from remote sensing is a recent (and still evolving) method in soil science which has tremendously increased the extent of the area over which soil properties can be mapped at high resolution and at relatively low cost, and has been used for SOC estimation in different contexts (Gomez *et al.*, 2008; Uno *et al.*, 2005; Hbirkou *et al.*, 2012b). Typically, airborne imagery techniques produce maps of soil properties at a resolution of a few meters and over a region of several hundreds of square kilometers. As observed by Gomez *et al.* (2012), the map of a soil property, if the value of each pixel is predicted only from the spectral information of this pixel (i.e. no interpolation), may be used to describe in detail its short range variation. In our case, a SOC image could be exploited (in totality or after sub-sampling) to better explain the role of the various controlling factors on the spatial distribution of SOC at different scales.

The main objective of this study was to quantify the relative influence of field-related processes on the topsoil organic carbon variation in the cropland of a 420 km^2 region in Luxembourg. To achieve this, a SOC image derived from hyperspectral imagery was used along with spatial mixed linear models fitted by restricted maximum likelihood estimation method (REML). After a preliminary exploratory analysis, the effect of texture was first isolated and quantified using texture units from a soil map as fixed effect. Then, spatial clusters composed of neighboring fields of the same texture class were selected and, for each of them, the field indices derived from a cadastral map were used as fixed effect in order to separate field-related variation from spatially correlated residuals. The influence of the field effect in each cluster was estimated as the part of the variance explained by the field component. In order to test if the effect of texture and field were significant, the model was then applied on randomly permuted data and on data simulated as a random field with equivalent mean and variance.



Figure 4.1: Study area and texture zones. Fields cover 42 % of the area while prediction of the surface SOC content from hyperspectral measurement is available for 9 % of the total surface of these fields. The red square in the South indicates the position of cluster 12 for which SOC predictions are displayed in Figures 6 & 7.

Table 4.1: General statistics of the datasets. All the values are in g C kg^{-1}.

Min	$q_0.25$	Median	$q_0.75$	Max	Mean	Stdev
-4.42	14.92	21.88	29.15	50.48	22.06	8.34

4.2 Material and methods

4.2.1 Study area

The study area is a stripe of approximately 60 km × 7 km that crosses the Grand-Duchy of Luxembourg in north-south direction (see Figure 4.1). The two geological regions of the country are represented (Marx, 2013). The northern third of the study area is part of the Oesling region and the Ardennes massif. It is a relatively homogeneous region with an average altitude of 450 m and a substrate of Devonian slate underlying shallow loamy cambisols and regosols. The most common crop rotation is a 6 to 8 years rotation where cereals alternate with forage crops by 3 or 4 year periods. In the south, the Gutland region lays on Secondary sandstone, marl and limestone substrates and presents a more varied topography with an average altitude of 244 m. It is covered by (sometimes gleyic) cambisols and luvisols mainly displaying clay or sandy textures. The most common crop rotation is a three years rotation of winter wheat, winter barley and silage maize. Overall, the region has a temperate oceanic climate with mean monthly average temperatures ranging from 0 °C to 17 °C and 863 mm of average annual rainfall. Fields cover 42 % of the area.

4.2.2 Datasets

The SOC values that were used to meet our objective were taken from the SOC map produced by Stevens *et al.* (2012) from hyperspectral imagery and covering our study area. The reader is referred to the original paper for a complete description of the method. Briefly, approx. 400 reference samples were taken in the field at a depth of 5 cm and SOC was measured in the lab by dry combustion. For the sample points the corresponding spectra from the airborne campaign were extracted and a calibration was established after stratification of the samples by texture class. Finally, SOC was predicted over 5 m × 5 m pixels. Since each calibration sample was actually made of 10 subsamples randomly taken inside a 7.5 m-wide square centered on the sample coordinates, SOC values predicted at pixels represent SOC values in a similar support, and should not be considered as point values. The root mean square error of prediction was estimated to 4.7 g C kg^{-1}, associated with a R^2 coefficient of 0.73. The bias was limited to -0.23 g C kg^{-1}. In Stevens *et al.* (2012) again, SOC content was predicted only for pixels that had bare soil at the time of the overflight (according to a vegetation index). Therefore, only fields for which at least half of the surface was covered by SOC pixels were retained, representing 766 fields and 18.6 % of the total cropland surface. Table 4.1 displays general

statistics of the SOC content over the fields. Although they are meaningless, pixels with negative SOC values were highly exceptional (< 0.005 %) and were not modified.

To assess the effect of the soil type, the soil association map of the Grand Duchy of Luxembourg was used (Marx, 2013). Soil classes were aggregated (see Figure 4.1) and four classes were used for the analysis, indicating the texture of the upper 20 cm of the profile, namely clay, clay-loam, loam and sandy loam. This seems an adequate classification to account for soil type, as Stevens *et al.* (2012) have shown large differences between the mean SOC content of these classes. The cadastral map of the country was also used to obtain the boundaries of each field.

4.2.3 Exploratory analysis of spatial variation

The empirical variogram was first estimated to give an overview of the variation of SOC in the study area. For a convenient computation time, 5000 pixels were selected by simple random sampling and the empirical variogram was computed by the method-of-moments. A nested variogram model was fitted to characterize the multiple scales of variation similarly to what was done in Section 3.2.4. The method-of-moments was chosen for its convenience to fit a nested model. We underline that at this stage the fit of a model was only performed for exploration and visual control.

The model was composed of a nugget and three spherical models of distinct range parameters, r_1, r_2 and r_3. Spherical models were selected for the same methodological reasons as described in Section 3.2.3. The choice of using three spherical models comes from our perception that, as in the loam belt in Chapter 3, the slope of the shape of the empirical variogram shows three breaks at different distances. Therefore, each range parameter was fine tuned by testing it over a grid of values covering the specific distance at which we perceived the corresponding break. The combination of range parameters that minimized the sum of squares of the residuals of the fit was selected. Based on first visual evaluation, r_1 was tested by 20 m increments between 20 m and 2 km, r_2 was tested by 100 m increments between 5 km and 15 km and r_3 was tested by 1 km increments between 20 km and 500 km. The presence of a unique minimum for the sum of squares was visually checked from the square loss objective function.

4.2.4 Mixed-effect model

In order to measure the effect of texture and the effect of fields, these two factors were successively used as fixed effects in a spatial linear mixed effect model according to the general structure given by Equation (B.1) in B. No intercept was included in the model. First, the effect of texture was assessed using model M1, with

$$\text{M1} : \text{SOC} = \mathbf{M}_{\text{text}}\boldsymbol{\beta}_{\text{text}} + \boldsymbol{\epsilon}_{\text{text}} \tag{4.1}$$

where $\mathbf{M}_{\text{text}}$ is the design matrix for texture classes, $\boldsymbol{\beta}_{\text{text}}$ are the associated parameters and $\boldsymbol{\epsilon}_{\text{text}}$ are the spatially correlated residuals which, in this case, would include field-related variation, if any. The part of the total variance explained by the texture was estimated as

$$u_{\text{text}} = \frac{\text{Var(SOC)} - \text{Var}(\boldsymbol{\epsilon}_{\text{text}})}{\text{Var(SOC)}} \tag{4.2}$$

To estimate the effect of fields, the field indices from a cadastral map were used as fixed effects in model M2, with

$$\text{M2} : \text{SOC} = \mathbf{M}_{\text{fields}}\boldsymbol{\beta}_{\text{fields}} + \boldsymbol{\epsilon}_{\text{fields}} \tag{4.3}$$

with $\text{Var}(\boldsymbol{\epsilon}_{\text{text}})$ the variance of the residuals and Var(SOC) the variance of the SOC content.

Model M2 was separately fitted over multiple clusters of neighboring fields and of homogeneous texture class, so that the fixed effect did not implicitly account for the effect of texture. For a more thorough description of the way model M2 was applied to the SOC image, the reader is referred to Section 4.2.6. For a given cluster i, the part of the total variance explained by the fields was given by

$$u_{\text{fields},i} = \frac{\text{Var}(\text{SOC}_i) - \text{Var}(\boldsymbol{\epsilon}_{\text{fields},i})}{\text{Var}(\text{SOC}_i)} \tag{4.4}$$

with $\text{Var}(\boldsymbol{\epsilon}_{\text{fields, i}})$ the variance of the residuals and $\text{Var}(\text{SOC}_i)$ the variance of the SOC content over the cluster.

It should be noted that repeated samplings were done to increase the significance of the results (see Section Section 4.2.6). Eqs. 4.1 to 4.4 were applied on single repetitions and the results were summarized over multiple repetitions. In each model, for each repetition the β's were estimated by eq. B.5. Their variances and therefore their standard deviations were given by the diagonal of Σ_β in Equation B.7. Each $\bar{\beta}$ was estimated as the weighted mean of the value of β over all repetitions, with the weight equal to the inverse of the variance. A symmetric confidence interval was estimated on this $\bar{\beta}$. Its half-width was estimated as the weighted standard deviation, with again the weights equal to inverse of the variance. The value of $\bar{u}_{\text{text}}$ and $\bar{u}_{\text{fields},i}$ were estimated as the mean of u_{text} and $u_{\text{fields},i}$ over all repetition. The half-width of the confidence interval on $\bar{u}_{\text{text}}$ and $\bar{u}_{\text{fields},i}$ was estimated as the standard deviation of u_{text} and $u_{\text{fields},i}$ divided by the square root of the number of repetitions.

Sometimes, Equations 4.2 and 4.4 can lead to negative u_{text} and $u_{\text{fields},i}$. This can happen when there is spatial dependency between $\mathbf{M}_{\text{text}}\boldsymbol{\beta}_{\text{text}}$ and $\boldsymbol{\epsilon}_{\text{fields},i}$ or between $\mathbf{M}_{\text{fields}}\boldsymbol{\beta}_{\text{fields}}$ and $\boldsymbol{\epsilon}_{\text{fields},i}$ so that we may have $\text{Var(SOC)} < \text{Var}(\boldsymbol{\epsilon}_{\text{text}})$ or $\text{Var}(\text{SOC}_i) < \text{Var}(\boldsymbol{\epsilon}_{\text{fields},i})$. In this case the value of u_{text} and $u_{\text{fields},i}$ should be considered as zero since a negative part of variation is meaningless. The original values are however shown in the results for the sake of clarity.

4.2.5 Estimating the spatial model with restricted maximum likelihood

The spatial covariance model was fitted by the method of Restricted Maximum Likelihood (REML) (Patterson and Thompson, 1971). This spatial covariance was then used to calculate the coefficients $\boldsymbol{\beta}$ by generalized least squares. The prediction from the model using these parameters is called the empirical best linear unbiased predictor (EBLUP). The whole procedure is thus called REML-EBLUP and has been widely used to estimate covariance parameters in the framework of mixed model with spatial covariance (Chai *et al.*, 2008; Lark *et al.*, 2006; Santra *et al.*, 2012). The method proved to be very stable (Kitanidis and Vomvoris, 1983) and the quality of prediction is generally superior to using the method-of-moments, in particular where fixed effects are large compared to spatial covariance (Minasny and McBratney, 2007). The theory of REML-EBLUP is detailed in B.

A flexible model with a single range of variation was used to describe the spatial covariance. The structure and the parameters of the model found in the exploratory analysis (Section 4.2.3) were thus not imposed, in order to avoid too strong constraints. Instead, only one range was fitted because (i) the fixed effect included already a part of the spatial structure and (ii) the limited number of samples used in the REML (see Section 4.2.6) impaired the possibility to fit satisfactorily more than one range. A model based on the Matérn covariance function was selected, with

$$C(h) = \begin{cases} c_0 + c_1 & \text{if } h = 0 \\ c_1 \left[\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{r} \right)^{\nu} K_\nu \left(\frac{h}{r} \right) \right] & \text{if } h > 0 \end{cases}$$

where $C(h)$ is the covariance between two points separated by a distance h, Γ is the Gamma function, K_ν is the modified Bessel function of the second kind of order ν and the parameters $\boldsymbol{\theta} = [c_0, c_1, \nu, r]$ represent respectively the nugget, the sill, the smoothness and the range of the model.The Matérn function is able to take different shapes and to reproduce many different usual covariance or variogram functions with a limited number of parameters (Minasny and McBratney, 2005).

The fit of the four parameters of the variogram model was done by the profile likelihood method which is able to give a good estimation of the maximum for a combination of all the parameters (Minasny and McBratney, 2007). First, different values of r and ν where selected at regular spacing. For each combination of r and ν, the values of c_0 and c_1 which maximized the conditional log-likelihood $\ell(c_0, c_1, \mathbf{y}|r, \nu)$ were found by a method of constrained minimization. The *fmincon* function of the ®Matlab Optimization Toolbox was used with the interior-point algorithm. Convergence was visually checked with a countourplot of $\ell(c_0, c_1, \mathbf{y}|r, \nu)$ as a function of r and ν.

All individual variograms were controlled visually. To this aim, the 95% confidence interval on the variogram was estimated. First, the covariance matrix for c_0 and c_1 was computed as the inverse of the Hessian matrix, which was itself calculated by finite differences. The confidence interval was then computed by a Monte Carlo method, assuming a normal joint probability density function for c_0 and c_1. One variogram model along with its confidence interval was chosen randomly for the illustration (Figure 4.3).

4.2.6 Optimum use of the hyperspectral image

Methodological choices had to be made to comply with practical constraints and to set favorable conditions for fitting the models to the hyperspectral image. On the one hand, it was not possible to use simultaneously all the pixels of the image for the REML procedure because of computational burden. On the other hand, when fields were used in the model, the geographic distribution of the calibration pixels was expected to strongly influence how well the model would explain the spatial variation of SOC. For instance, we assumed (i) that isolated fields were less helpful than adjacent fields for a method separating the field effect from a spatially dependent continuous component, (ii) that fields separated by a large distance did not have to be included simultaneously as they shared no dependencies (the spatial covariance was only present at shorter range) and (iii) that the model needed at least a few points in a given field to reliably estimate its field effect.

In order to combine these limitations and assumptions, we used different strategies for model M1 and M2. Model M1 was repeatedly fitted over groups of 300 SOC pixels randomly selected over the whole study area, and the results were summarized over hundred repetitions (cf 4.2.4). For model M2, the minimal distance between each field was computed and fields were grouped in geographical clusters. In practice, a field belonged to a cluster if it had the same texture class and was at maximum hundred meters from at least one other member of the cluster. Only clusters containing at least five fields were retained. For each field cluster, 300 pixels were selected on a regular grid (whose mesh size thus depended on the surface of the cluster) while hundred additional pixels were distributed between the fields (proportionally to their surfaces) and randomly selected inside. This operation was repeated twenty times and, as for model M1, results were summarized over these repetitions.

4.2.7 Validation of the model

It is worth emphasizing here that the goal of the model was not to spatially predict SOC but to characterize quantitatively the controlling factors. To this end, a point-by-point validation of REML-EBLUP predictions would have been poorly informative. Because of the spatial dependence, it seems clear that the performance indicated in this case would have been largely determined by the spatial density of the calibration samples (as it would also be the case, for example, with kriging interpolation). On the contrary, a poor validation score

would not mean that the model was not explaining well the influence of the controlling factors.

According to the preceding remarks, the following analyzes were performed as validation. For models M1 and M2, every time pixels were selected, the values of the categorical variable representing the fixed effect (i.e. texture class or fields) were randomly permuted. Results of the model obtained with permuted indices were compared with the ones obtained with true indices. This allowed us to check if the fixed effect explained significantly more variation than would have been explained using some random grouping (thus preserving the same number of degrees of freedom in the model).

Furthermore, another validation procedure was performed to check that models were not confounding variation attributed to spatial units (texture classes or fields) with spatially continuous variation acting at similar spatial scale. First, the variogram of SOC was modeled by a spherical model along with a nugget component. Then, a spatial random field was simulated by the method of Monte Carlo using the covariance matrix constructed with the estimated variogram parameters. This created a fictional SOC distribution having roughly the same spatial properties as the distribution given by the hyperspectral image, but completely independent from the spatial units. Finally, the model was fitted on this simulated SOC distribution.

The two validation procedures were applied repetition by repetition (Section 4.2.6). Besides, for model M2, validation was performed cluster by cluster. In practice, for each repetition, pixels were first selected, modifications related validation procedures were applied and the model was fitted. For the original values and for any of the two validation procedures, the results obtained for each repetitions were averaged. Results obtained with the original SOC image were compared to mean results obtained within each validation procedure. In particular, the mean parts of the variance explained by the fixed effects in the three cases, namely $\bar{u}_i^{\text{orig}}$, $\bar{u}_i^{\text{permut}}$ and $\bar{u}_i^{\text{simul}}$, were compared. If the fixed effects explained more variance when models were based on the original values than when they were based on modified values, this proved that the model was able to identify variation which was actually related to these fixed effects. The level of confidence on $\bar{u}_i^{\text{orig}}$ was estimated by

$$p_{\text{conf}} = \frac{1}{N} \sum_{i=1}^{N} [\text{I}(u_i^{\text{orig}} > u_i^{\text{permut}}) \cdot \text{I}(u_i^{\text{orig}} > u_i^{\text{simul}})] \tag{4.5}$$

with u^{orig}, u^{permut}, u^{simul} the ratio of variance explained by the fixed effects for one repetition when respectively original values, permuted indices and simulated SOC values were used, I(.) the indicator function and N the number of repetitions. We assumed that the fixed effects were significant when $p_{\text{conf}} \geq 0.95$.

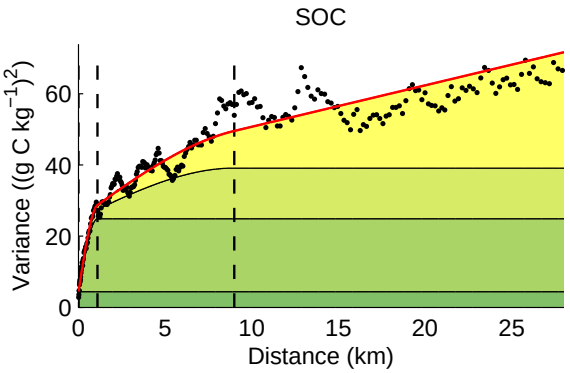


Figure 4.2: Empirical variogram and nested variogram model for the SOC content in the area. Variograms were estimated from 5000 randomly selected pixels of the SOC image. It displays a relatively tiny nugget effect plus three distinct regimes of variation with the characteristic ranges of 1080 m and 9500 m.

4.3 Results

4.3.1 Exploratory analysis of spatial variation

The empirical variogram was computed for the 5000 selected pixels and the multiscale nested model was fitted. The empirical variograms and the model (with a representation of contribution of each scale components) are shown in Figure 4.2. The values of 1080 m and 9500 m were found for r_1 and r_2. No well-defined optimum was found for r_3, which tended to be infinite. For r_3 set to the maximal tested value (500 km), the values found for r_1 and r_2 were clearly located in a unique and well defined minimum of the square loss objective function (results not shown).

The empirical model was roughly divided in three segments which were consistently represented by the model (Figure 4.2). At short distance (i.e. 0-2 km), the variogram model was completely superimposed to the observations, while fluctuations in the empirical variogram increased with the distance. The submodels associated with r_1 and r_2 accounted respectively for 34 % and 25 % of the total variance, while the part of the nugget effect was very small (5 %).

The presence of typical ranges of variation can be tentatively related to processes that influence SOC content. First, r_1 has an order of magnitude close to the maximum field length (the 99 % quantile of the maximum field lengths equals 669 m). Therefore, the fast increase of variance between 0 and r_1 may be related to the fact that, at these short distances, there is a higher probability that two pixels belong to the same field. Secondly, r_2 is of the order of magnitude of the size of the clay and sandy-loam patches in the Gutland region. Finally, the absence of plateau at long distance is coherent with Minasny

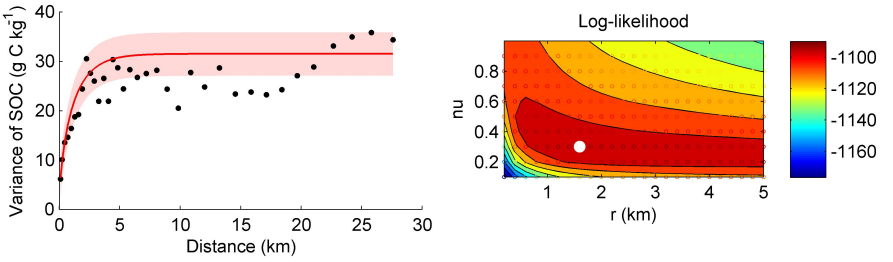


Figure 4.3: (a) Variogram of the residuals of model M1 fitted on 300 randomly selected pixels shown for one repetition. Black dots represent the empirical variogram, red line represents the modeled variogram and red patch represents the 95% confidence interval. In the study, pixels were repeatedly selected and the results of all fits of model M1 were averaged. (b) Corresponding contourplot of the log-likelihood of the parameters c_0 and c_1 of the Matérn covariance function conditional to given ν and r combinations. The estimated parameters (white dot) correspond to an actual maximum of the log-likelihood.

et al. (2013) who have shown that the variance may reach very large values for increasing distances in the north-south direction due a to climatic gradient related to the altitude.

4.3.2 Model M1 (texture classes)

Hundred groups of 300 randomly selected pixels were repeatedly built. For each group, model M1 was fitted in a first case on the original values, in a second case with randomly permuted texture class indices and in a last case after replacement of the SOC values by simulated values. In any of these three cases, the mean of the covariance function parameters r, ν, c_0, and c_1, the part of the variance explained by the texture $\mathbf{u}_{\text{text}}$ and the coefficient of the the texture classes β_C, β_{CL}, β_L, β_{SL} were computed over all repetitions, along with the standard error (Table 4.2). In all repetitions, no samples were located in the soil classes *urban* and *colluvial-alluvial* so the corresponding coefficients were not estimated. The extreme values and quartiles of $\mathbf{u}_{\text{text}}$ over the repetitions are also presented in Figure 4.4.

When original data were used, texture classes had stronger effects on average (i.e. the $\bar{\beta}$'s were more different), explaining thus a larger part of the SOC variance than the two other cases. In that case $u_{\text{text}}^{\text{orig}}$ was between 58.9 % and 75.2 % with a mean $\bar{u}_{\text{text}}^{\text{orig}}$ equal to 65.7 ± 0.3 %. When texture class indices were randomly permuted, $\bar{u}_{\text{text}}^{\text{permut}}$ was always smaller than 0.1 % while when SOC was randomly simulated (using covariance parameters estimated from original SOC values), $\bar{u}_{\text{text}}^{\text{simul}}$ was estimated to 15.1 ± 1.6 %. This indicates that the method could attribute some variation to the texture classes even if these have obviously no effect (since SOC was simulated independently

Table 4.2: Mean results of model M1 and associated standard error on the mean, calculated over hundred repetitions and for the three compared situations (Orig.: true SOC image and texture units; Perm.: true SOC image but randomly permuted texture classes; Simul.: Simulated SOC spatial distribution and true texture classes). When original data was used, texture classes had stronger effects, explaining thus a larger part of the SOC variance.

	$\bar{r}$	$\bar{\nu}$	$\bar{c}_0$	$\bar{c}_1$	$\bar{u}_{text}$	$\bar{\beta}_C$	$\bar{\beta}_{CL}$	$\bar{\beta}_L$	$\bar{\beta}_{SL}$
	km	/	g C kg^{-1}	g C kg^{-1}	%	g C kg^{-1}	g C kg^{-1}	g C kg^{-1}	g C kg^{-1}
Orig.	2.4 ± 0.2	0.3 ± 0.0	0.6 ± 0.1	26.7 ± 0.3	65.7 ± 0.3	21.5 ± 0.1	15.2 ± 0.2	29.5 ± 0.1	13.4 ± 0.1
Permut.	7.7 ± 0.3	0.3 ± 0.0	1.0 ± 0.1	57.2 ± 1.5	0.1 ± 0.0	22.0 ± 0.2	22.0 ± 0.2	22.0 ± 0.2	22.1 ± 0.2
Simul.	2.4 ± 0.2	0.5 ± 0.0	2.7 ± 0.1	54.4 ± 1.9	15.1 ± 1.6	22.3 ± 0.2	21.9 ± 0.3	21.5 ± 0.2	22.4 ± 0.2

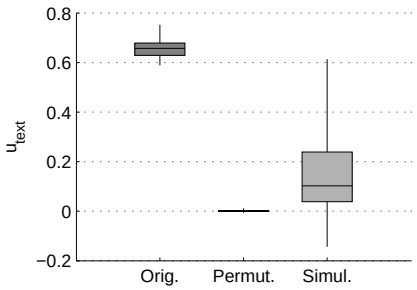


Figure 4.4: Proportion of variance attributed to the texture classes by model M1. The boxplots represents the minimum, maximum and quartiles over hundred repetitions and for the three compared situations (Orig.: true SOC image and texture units; Perm.: true SOC image but randomly permuted texture classes; Simul.: Simulated SOC spatial distribution and true texture classes). The proportion of variance which was explained by texture classes was systematically higher when true values were used.

from them). However, for any repetition $u_{\text{text}}^{\text{orig}}$ was systematically higher than both $u_{\text{text}}^{\text{permut}}$ and $u_{\text{text}}^{\text{simul}}$. This lead to a confidence level p_{conf} of 100 % on $\bar{u}_{\text{text}}^{\text{orig}}$. Note that since the distribution of soil classes is not independent from the altitude, it is likely that the effect of soil type is in fact, at least partly, an effect of climatic variation due to the altitude.

The variogram of the residuals and the contourplot of the log-likelihood are given by (Figure 4.3) for an arbitrarily chosen fit of model M1. This illustrates the use of REML along with the profile likelihood method. It shows a good coherence between the variogram model and the empirical variogram of the residuals, along with a narrow confidence interval on the variogram model (a). The clear maximum of the log-likelihood in function of r and ν (b) also indicates that the profile likelihood method converged efficiently. A similar level of quality of fit was observed in the applications of model M1 for all repetition.

Going further, we may interpret the similarity between the spatial structure of the residuals of M1 (i.e. once the effect of texture had been removed) and the structure of the fields as an indicator of the role of fields on the spatial variation of SOC. Indeed, the mean range parameter $\bar{r}$ was equal to 2400 m which is roughly of the same order of magnitude as the maximum length of the fields (i.e. 669). The fact that r was still larger than the length of fields could possibly be caused by the probability of similar management on neighboring fields as they are more likely to belong to the same farm. However, the interpretation of r has to be done cautiously since with the Matérn model, ν also influences how quickly the variograms increases with the distance.

4.3.3 Model M2 (fields)

Table 4.3: Results of model M2 averaged over twenty repetitions for each cluster and for the three compared situations (orig: original SOC image and texture units; permut: true SOC image but randomly permuted texture classes; simul: Simulated SOC spatial distribution and true texture classes). C, L and CL refer to clay, loam and clay-loam texture classes. p_{conf} represents the confidence level on $\bar{u}^{\text{orig}}_{\text{fields}}$. It was equal or superior to 95 % for twelve of the clusters, for which fields explained up to 68.8 ± 12.0 % of the variance.

Clust.	Text.	# fields	μ_{SOC}	σ_{SOC}	$\bar{u}^{\text{orig}}_{\text{fields}}$	$\bar{u}^{\text{permut}}_{\text{fields}}$	$\bar{u}^{\text{simul}}_{\text{fields}}$	p_{conf}
/	/	/	g C kg^{-1}	g C kg^{-1}	%	%	%	%
1	C	5	22.05	19.64	30.8 ± 5.4	0.3 ± 0.1	2.6 ± 0.5	90
2	C	16	23.52	17.47	52.7 ± 9.2	1.3 ± 0.2	25.0 ± 4.3	90
3	C	13	17.18	23.33	61.9 ± 10.8	0.7 ± 0.1	27.8 ± 4.8	95
4	C	8	26.72	13.80	12.1 ± 2.1	0.5 ± 0.1	5.9 ± 1.0	75
5	C	19	20.41	15.13	43.6 ± 7.6	1.0 ± 0.2	28.6 ± 5.0	60
6	C	9	26.15	25.91	39.6 ± 6.9	0.7 ± 0.1	4.1 ± 0.7	100
7	C	6	33.94	10.93	29.9 ± 5.2	0.5 ± 0.1	5.9 ± 1.0	100
8	C	14	14.81	27.59	50.0 ± 8.7	0.3 ± 0.1	19.1 ± 3.3	100
9	C	5	21.18	7.18	35.1 ± 6.1	0.2 ± 0.0	7.4 ± 1.3	75
10	C	10	21.57	17.92	54.6 ± 9.5	0.7 ± 0.1	29.6 ± 5.1	80
11	C	23	15.62	26.88	45.2 ± 7.9	0.9 ± 0.1	19.5 ± 3.4	95
12	C	17	18.39	21.37	48.1 ± 8.4	0.6 ± 0.1	14.5 ± 2.5	90
13	C	5	21.07	23.41	68.8 ± 12.0	0.2 ± 0.0	16.2 ± 2.8	100
14	C	15	17.34	19.02	12.2 ± 2.1	0.6 ± 0.1	5.8 ± 1.0	55
15	L	11	30.39	9.72	62.0 ± 10.8	-6.4 ± 1.1	22.3 ± 3.9	95
16	L	13	30.82	17.57	58.4 ± 10.2	0.6 ± 0.1	32.0 ± 5.6	90
17	L	5	27.33	9.61	15.2 ± 2.6	0.1 ± 0.0	6.0 ± 1.1	70
18	L	6	28.35	10.64	4.6 ± 0.8	0.3 ± 0.1	6.4 ± 1.1	40

19	L	5	32.17	9.91	10.3 ± 1.8	0.2 ± 0.0	18.8 ± 3.3	25
20	L	7	29.90	9.39	59.5 ± 10.4	0.3 ± 0.1	13.3 ± 2.3	100
21	L	7	29.75	15.25	41.0 ± 7.1	0.2 ± 0.0	11.4 ± 2.0	100
22	L	6	28.92	8.21	-7.5 ± 1.3	0.3 ± 0.1	13.7 ± 2.4	5
23	L	7	32.06	11.54	33.7 ± 5.9	0.1 ± 0.0	13.2 ± 2.3	85
24	L	6	35.14	5.21	-26.0 ± 4.5	0.5 ± 0.1	18.7 ± 3.2	15
25	SL	18	11.42	2.44	30.5 ± 5.3	1.8 ± 0.3	8.2 ± 1.4	90
26	SL	8	11.80	4.53	47.9 ± 8.3	0.4 ± 0.1	8.3 ± 1.4	95
27	SL	9	12.37	2.89	-4.8 ± 0.8	1.0 ± 0.2	10.0 ± 1.7	25
28	SL	7	14.03	3.03	0.5 ± 0.1	0.5 ± 0.1	1.6 ± 0.3	40
29	SL	16	9.73	3.81	23.3 ± 4.1	1.5 ± 0.3	7.5 ± 1.3	100
30	SL	5	12.37	6.46	46.5 ± 8.1	-0.0 ± 0.0	14.0 ± 2.4	95
31	SL	5	9.80	5.40	29.5 ± 5.1	0.1 ± 0.0	12.8 ± 2.2	80
32	SL	6	15.17	1.87	15.4 ± 2.7	0.2 ± 0.0	6.7 ± 1.2	70
33	SL	5	16.88	4.33	11.3 ± 2.0	0.2 ± 0.0	14.3 ± 2.5	40

Fields were grouped according to texture class and proximity as described in Section 4.2.6. For each cluster, model M2 was applied for the three different cases (original data, permutation of field indices and random simulation of SOC values) and in each case, results were averaged over twenty repetitions. Two clusters were not retained because the fit of the covariance function was not satisfactory. In one case, this seemed to be caused by a patch of pixels of abnormally high SOC values in the field. In the other case, a stripe of high SOC values along a field boundary was responsible. Ultimately, 33 clusters were selected, containing between five and 23 fields.

Table 4.3 displays a summary of the main results cluster by cluster while Figure 4.6 also displays the distribution of the part of the variance explained by the fields. Results first indicate that the mean and the standard deviation of the SOC content, μ_{SOC} and σ_{SOC}, vary a lot between clusters. This seems largely related to the texture class, with typically higher SOC contents in loamy areas but higher variation coefficients (i.e. $\sigma_{\mathrm{SOC}}/\mu_{\mathrm{SOC}}$) in clay areas.

The mean part of the variance explained by the fields when original values were used, $\bar{u}^{\mathrm{orig}}_{\mathrm{fields}}$ was superior to its equivalents $\bar{u}^{\mathrm{permut}}_{\mathrm{fields}}$ and $\bar{u}^{\mathrm{simul}}_{\mathrm{fields}}$ for 23 out of 33 clusters. For a given cluster, the spread of $u^{\mathrm{orig}}_{\mathrm{fields}}$ (which only resulted from the selection of different calibration pixels between repetitions) was relatively small. On the contrary, the spread of $u^{\mathrm{simul}}_{\mathrm{fields}}$ was generally much larger, while $u^{\mathrm{permut}}_{\mathrm{fields}}$ was always very close to zero. According to Equation 4.5, the effect of fields was significant for 12 cluster, corresponding to 36 % of the total number. The part of the variance explained by the fields in these clusters varied from 23.3 ± 4.1 % for cluster 29 up to 68.8 ± 12.0 % for cluster 13. $\bar{u}^{\mathrm{orig}}_{\mathrm{fields}}$ was negative in three cluster, indicating that not only no field effect was observed but also that on average, two fields were more similar than would have been two patches of the same footprint as these fields, but located inside a same single field.

In order to illustrate how the method separates the effect of the spatial units (fixed effect) from the spatially continuous variation, Figure 4.6 maps the components of the REML-EBLUP prediction over cluster 12 both for the original SOC image (Figure 4.6) and for a randomly simulated SOC image used in the validation procedure (Figure 4.7). The position of cluster 12 in the study area is visible on Figure 4.1. For original SOC values, the map of the fields effect (4.6c) shows large differences between fields, which confirms the importance of considering fields effect in that cluster. The map of the residuals (4.6d) illustrates how the method fitted a spatial component with a single range of variation that had to be continuous inside the fields but also over the boundaries between adjacent fields. This is a strong working hypothesis since the REML method had to do a trade-off between estimating the effect of each individual field and fitting a spatial covariance function representative for the whole cluster. The map of predicted SOC (4.6b) resembles the hyperspectral image (4.6a) with the main patterns reproduced, but not the very local ($\approx$ 10 m) variation. This is due to the fact that the sampling density of the image was not sufficient with regard to the range of the variogram model to predict perfectly the residuals in each pixel. When model M2 was fitted on a randomly

Figure 4.5: Proportion of variance attributed to the fields classes by model M2, cluster by cluster. The boxplots represent the minimum, maximum and quartiles over twenty repetitions and for the three compared situations (Orig.: original SOC image and texture units; Perm.: true SOC image but randomly permuted texture classes; Simul.: Simulated SOC spatial distribution and true texture classes). u_{fields} is generally higher when original data are used, but large disparities appear between clusters.

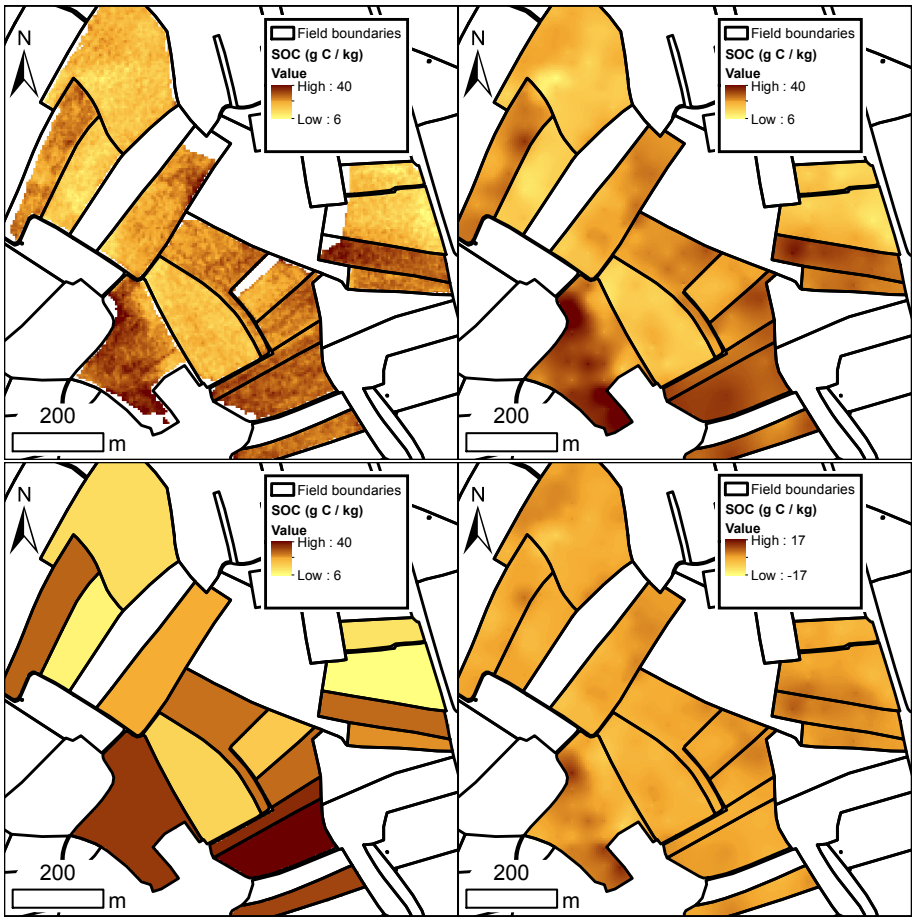


Figure 4.6: Spatial prediction of the SOC content by the model M2 on cluster 12 for the original SOC image with (a) this original image, (b) the REML-EBLUP prediction, (c) the estimated component representing the field effect and (d) the estimated component representing the spatially correlated residuals (note the difference in scale). The prediction in (b) is equal to the sum of the separated components (c) and (d). The effect of different fields could already be suspected from (a) but our methodology further allowed us to rigorously assess with a confidence of 90 % that fields counted for 48 $\pm$ 8 % of the SOC variance inside the cluster.

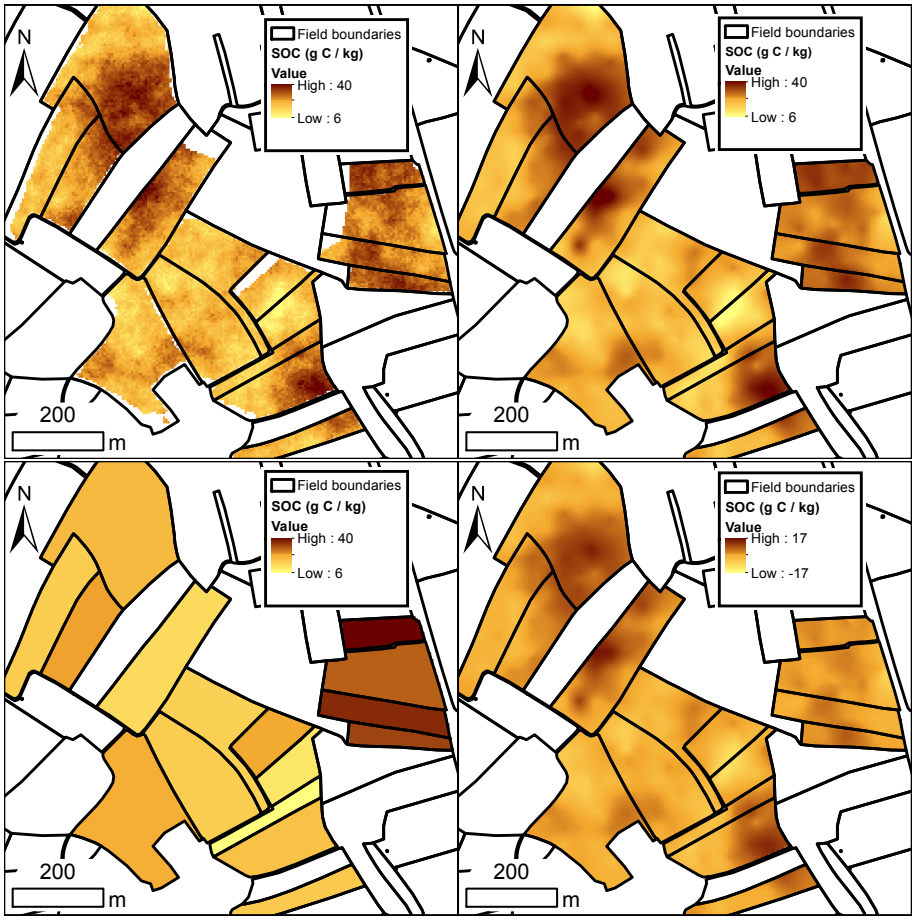


Figure 4.7: Spatial prediction of the SOC content by the model M2 on cluster 12 for a SOC image randomly simulated using a covariance function estimated from the original image. (a) displays the simulated SOC values, (b) the REML-EBLUP prediction, (c) the estimated component representing the field effect and (d) the estimated component representing the spatially correlated residuals (note the difference in scale). When the current figure is compared with the fit of model M2 on the original SOC image (Figure 4.6, same SOC scale), it appears that a larger part of the SOC variance is explained by the fields and a smaller part by the spatially correlated residuals when the original image is used, which contributes to evidence an actual field effect in this case.

simulated SOC image, the component representing the field effect (Figure 4.7c) displayed generally smaller differences between adjacent fields than with the original SOC image. However, some grouped fields on the East of the image, which are not adjacent to the others, showed higher values. This shows that even when SOC values are not linked to the field distribution, the model may confound the spatially continuous variability with a field effect.

4.4 Discussion

4.4.1 Rigorous assessment of texture and fields effect

If the simple observation of the hyperspectral image already highlights patterns of variation that coincide with texture regions and field boundaries, our mixed effect model approach allowed us to rigorously assess the presence and the magnitude of texture- and field-related variation in landscape as well as its relative importance at the regional scale. In particular, the validation method consisting in simulating SOC with the same spatial properties as in the original image was particularly useful for assessing rigorously the presence of a field effect. The fact that sometimes, simulated SOC distribution seemed at a first glance to follow the field boundaries although they were by definition independent, demonstrates the usefulness of an objective method. If field effect can be conceptually distinguished from a spatially continuous variation, in the reality the two effects were partly confounded. When model M2 was fitted on simulated SOC, it often indicated the presence of a field effect too (i.e. $u^{\text{simul}}_{\text{fields}} > 0$), although there were obviously none (Figure 4.5). Most likely, this is due to the fact that the scale at which spatially continuous variation occurs partially overlaps the scale of the fields. Despite this, in multiple clusters fields explained significantly more SOC variation when model M2 was fitted on the original SOC image than on the simulations. This allowed us to rigorously assess the presence of an actual field effect with a given level of confidence. The contrast in the part of variation related to field effect between clusters can be related to differences in local context. For example, we could have clusters where agricultural management is particularly homogeneous over the fields because they belong to the same farm, and so the effect of the field factor would be lower.

Nevertheless, the ability of the method to distinguish field effect is limited and could maybe be improved. Too often, a field effect was only predicted with small confidence, as illustrated by the frequent occurrence of p_{conf} values superior to 50 % but inferior to 95 % for the clusters. A reason for this could be the fact that the method intended to model the spatially continuous variation with a covariance model of unique range parameter, while the processes behind this variation may be very different between locations. In particular, local SOC variation related to sporadic landscape features (e.g. past presence of hedges or in-field storage of manure) may be difficult to integrate in the spatial covariance model. On the contrary, when assessing the importance of field-related variation *at regional scale*, fitting a single covariance model over the whole area may be a more elegant option. This turned out to be unfortunately

not possible because of the large number of calibration samples that would have been needed and the associated computational burden of the method. Finally, the fact that the SOC image was available for 18.6 % of the cropland surface may be also limiting since it adds information gaps between the sampled fields of a cluster (e.g. Figure 4.6). We may reasonably assume that if the area was more intensively covered by the image, the fit of M2 would be more constrained and it would be easier for the method to distinguish between the components.

If the analysis showed that at regional scale texture was the main factor controlling SOC with 65.7 % of variance explained, at landscape scale the influence of fields may also be very large. Processes acting on SOC content at the scale of the fields are for example differences in agricultural management and the residual effect of former (up to a hundred of years in the past) land uses. Field boundaries may also coincide with landscape features (e.g. hedges, streams, mounds) that have an influence on SOC by preventing lateral transport of sediments and SOC between adjacent fields (see for example (Viaud *et al.*, 2010b; Ghariani *et al.*, 2014). The presence of a field effect may explain the difficulty of Stevens *et al.* (2014) to efficiently predict topsoil SOC (in an area of homogeneous texture) using only topographic information (terrain attributes). Doetterl *et al.* (2013a) also showed that, within our study area, the topography became a good predictor of SOC only for the subsoil at depths below 20 cm, where the influence of field-related processes was expected to be weaker.

4.4.2 Implications for soil organic carbon modeling and mapping

The influence of field-related processes on the surface SOC content (and thus on the topsoil layer mixed by tillage) has direct consequences for modeling and mapping SOC at a resolution finer or equal to the average field size. In the context of modeling, the point-scale validation of a coupled erosion-SOC dynamics model (for example SPEROS-C, Van Oost *et al.* (2005)) would lead to poor coefficient of determination if the processes acting on a field basis are not properly recognized. In the topsoil, validation should therefore be based on larger spatial units. In the context of digital soil mapping, incorporating explicitly the field effect is not always possible. For example, the exact practices of farmers over large regions are rarely recorded in a centralized database, and the long-term land use history is rarely known at a sufficient spatial and temporal resolution. To predict SOC considering the specific effect of each field using traditional survey methods, each field should therefore contain enough soil samples to estimate this field effect consistently. This is hardly feasible over large regions considering the cost of sampling, in particular if samples are scattered over the whole region. Therefore, remote SOC imaging presents an interesting alternative since it can cover large regions at a resolution finer than the field size and with limited costs compared to sampling.

4.4.3 Limitations of hyperspectral images

Attention should also be paid to the limited precision for SOC prediction using remote spectral measurements. For a panel of soil remotely sensed soil properties, Gomez *et al.* (2012) observed that the variance of predicted properties was inferior to the variance of observed properties, for all distance classes. In a stringent validation procedure, Stevens *et al.* (2012) observed with the best predictive model at hand a limited prediction ability. In their analysis, the leave-one-field-out cross-validation gave the lowest estimation of the precision, showing that fields containing no calibration samples are the ones containing the largest errors on average. This means that specific differences of SOC content between fields observed on SOC images produced by remote spectroscopy may be partly artificial due to the difficulty of the prediction model to generalize over unsampled fields. Typically, the surface reflectance of individual bare fields may be influenced by features which are not related to the SOC content but are still related to agricultural management, such as surface roughness, crop residues or the use of SOC-free amendments that alter the soil color. For these reasons, the spatial structure of the error in SOC remote sensing imagery applications should be analyzed and the specific differences between fields be corroborated by irrefutable soil observations. However, with regard to the image that was used in the present study, Stevens *et al.* (2012) assures that variograms of the squared residuals were generated and did not show strong evidence of spatial structures.

4.5 Conclusions

In summary, using spatial linear mixed effect models within the rigorous framework of REML-EBLUP allowed us to better describe and understand the spatial structure of SOC variation as it provided some insights about the nature of the various acting processes. We sum up the four main findings.

First, the predominant influence of the soil texture on SOC at regional scale was confirmed. It counted for 65.7 $\pm$ 0.3 % of the variance over our study area. This may be underestimated since texture may also influence SOC at a scale finer than the soil map that was used in model M1. On the contrary, because the distribution of some texture classes is correlated with latitude and altitude, it is likely that the effect of soil type estimated here is in fact, at least partly, an effect of climatic variation due to the altitude.

Secondly, the study proved the presence of an actual field effect on SOC, with large local disparities. Fields explained up to 68.8 $\pm$ 12.0 % of the SOC variance in clusters of neighboring fields of homogeneous texture class. Model M2 detected a field effect with a confidence level at least equal to 95 % for more than one third of the clusters. The processes which are driving these field difference are most certainly related to land-use and land management history, but this could be further investigated.

Thirdly, considering the large magnitude of field-related variation in some location, we recommend to take it into account in SOC inventory, modeling and digital mapping studies. This may for example consist in an adaptation of sampling designs and validation methodologies.

Finally, the study demonstrated that, despite its precision generally inferior to on-site samples, the large amount of information (i.e. high resolution over large areas) provided by remote sensing techniques may be exploited in order to analyze and model the multi-scale variation of soil properties. Following this idea, we suggest to further develop the methodology presented here by transposing it to different soil properties and in different pedologic, topographic, agricultural or climatic contexts.

Chapter 5

Accounting for the spatial scale of variation in total stock quantification

Abstract.

There are different cases of SOC prediction applications where only the total stock over the study area is required instead of the spatially explicit distribution. For example, total stock is of interest for conducting global C budgets and for verification purposes within C sequestration policies. Currently, efficient quantification of total stock is hampered by the large intrinsic variability of SOC, the complexity of the mechanisms of sequestration, the lack of standard and broadly accepted methodologies and the cost of data acquisition. In this chapter we propose to start from a mathematical formalization of the problem of total stock estimation. The general framework consists in the spatial integration over an area of interest of the prediction provided in any spatial location by a model built on a vector of auxiliary variables. This allows us to highlight the link between the prediction variance of the total stock as obtained by the model and the range of the spatial dependence of model error. In a next step, we generalize previous results in order to improve model prediction by accounting for the spatial dependence of model error. We compare the prediction variance of the kriging predictor with the prediction variance corresponding of two specific cases. Theoretical results indicate that, if kriging shows the lowest variance of prediction for the total stock (as it is the optimal linear predictor), there are situations in which spatial interpolation of model error is not necessary and simplification can be made without any serious loss of precision. The relevance of using simplified approach depends on the contrast between the range of dependence of model error and the size of the study area. Here we propose a decision support tool for total SOC stock quantification when a spatially exhaustive prediction model is available along with covariates to feed the model and, potentially, SOC point measurements. We also provide recommendations to improve a priori, i.e. before data acquisition and model development, the cost-effectiveness of total stock quantification. In conclusion, the findings proposed here can contribute

to develop a set of standard, cost-efficient and broadly accepted methodologies for total SOC stock estimation that are also flexible towards different purposes, different spatial scales and different soil and environmental contexts.

5.1 Introduction

There are many different contexts in which the spatially explicit prediction of SOC over an area of interest is necessary. For example, spatial distribution may be required to set up a C sequestration policy or to implement precision agriculture, as in both case the best suited agricultural management may differ from one place to another. Another example is soil-process models – a key element in the understanding of SOC-related processes and in the prediction of SOC stocks evolution – which generally require spatially explicit data layers for calibration and run.

In contrast, there are contexts in which only the total SOC stock over the area of interest is needed. This is the frame that will be investigated in the present chapter. Therefore, we propose to first overview the current applications, limitations and challenges of total SOC stock quantification.

5.1.1 Global C budget

Global SOC stock estimation can be part a global C budget. Global C budgets account for all the pools and fluxes of C between the Earth and its atmosphere. They are a valuable resource within a climate policy framework as they provide definitive records of recent trends as well as quantitative updates on permissible emissions for given climate stabilization targets. However, quantifying SOC stock turns out to be one of the most difficult task in assembling an annual carbon budget.

A representative and current example of global carbon budget is produced each year since 2005 by a community of researchers within the Global Carbon Project (Candela and Carlson, 2017). The budget is divided in five components which are annual growth rates of atmospheric CO_2 concentration, emissions from fossil fuels and industry, emissions from land-use change, CO_2 uptake by the ocean and terrestrial uptake. These components are quantified using different possibilities among direct or indirect measurement, national statistics data, biogeochemistry and vegetation models. The terrestrial uptake component, which includes processes regarding both soils and vegetation, is the one with the largest uncertainty (Le Quéré *et al.*, 2017).

The difficulty to quantify global SOC stock is also supported by the variation and large uncertainty in the estimates of global SOC stocks published during the last decades. Comparing 27 studies, Scharlemann *et al.* (2014) observed global SOC estimate ranging roughly from 500 to 3000 Pg C, and

from 900 to 1800 Pg C since the year 2000. The large variation in estimates of global SOC stocks reflects the disparity in sampling period, intensity and spatial resolution of the soil profile databases, and differences in approaches to calculations. Historical estimates of global stocks have been conducted by extrapolation of soil profiles grouped by soil or ecosystem type (e.g. (Batjes, 2014)) but estimates of global SOC stocks based on spatially explicit prediction have become more popular over the twenty last years (e.g. Köchy *et al.*, 2015). However, Heuvelink *et al.* (2017) estimates that, using a design based approach and random sampling, only 2100 points would be necessary to estimate total SOC stock with a standard deviation of 1.75 Pg C. This value is much lower than the divergences between historical estimation studies and could still be reduced if stratification methods were used.

5.1.2 Carbon policies

Measuring SOC stock can also be valuable in the framework of carbon policies related to climate. The European Union has set up a climate action strategy based on different levers. The first lever is the European Emissions Trading Scheme (EU-ETS) which is currently the biggest C market worldwide. It works on the principle of "cap and trade". A cap is set on the total amount of certain greenhouse gases that can be emitted and is voluntarily reduced over time. Then, companies receive or buy emission allowances that can be traded. They can also buy limited amounts of international credits from emission-saving projects around the world. The EU-ETS accounts for CO_2 emissions from power and heat generation, energy-intensive industry sectors and commercial aviation. Besides the EU-ETS there is the Effort Sharing Regulation (ESR) which covers national emissions for transport, building, waste and non-CO_2 emissions from agriculture (methane and nitrous oxide).

Sequestration of C related to agricultural management is not addressed by the EU-ETS nor the ESR but is part of the Land Use, Land Use Change and Forestry (LULUCF) regulation. This includes CO_2 emissions and removals from forest management, afforestation, reforestation, deforestation, cropland and grazing land management, with a no debit rule which implies a commitment to have zero net emissions (IFOAM, 2017). On a voluntary basis, a country can thus decide to measure the additional amount of C sequestered in soil in response to undertaken agricultural management and to include it in the national LULUCF emissions budget. Besides LULUCF regulation, the Common Agricultural Policy (CAP) proposes to reward European farmers with subsidies for the application of agricultural practices that are considered to have a positive impact on environment and/or economy (such as maintaining permanent grassland). Although some of these measures increase the stock of SOC in the soil, no quantification of this stock is required to justify the subsidies within the CAP.

The specificity of the mechanisms of C sequestration in soils represents one of the difficulties for its inclusion in policies. First, there is a limit to the amount of carbon that can be stored in the soil (see the concept of *saturation* explained

in Section 1.1.2). Agricultural soils have lost 50-70% of their original organic C that was present under natural conditions prior to clearing and cultivation (Sanderman *et al.*, 2017). Even with a shift of agricultural practices, it is admitted that only a limited part of this historical losses could be recovered if a field remains cultivated. Secondly, soil carbon sequestration is not permanent. It takes a long time to sequester carbon in soil, but it can very quickly be lost again if farm management changes. This means one needs to ensure that farmers remain committed to keep the carbon in the soil when the full potential is reached (IFOAM, 2017). Finally, it is essential that sequestration policies only reward sequestration that is additional to what would occur in the normal course of business and also consider leakage, i.e. the transfer of emissions to another location, time or form (Thamo and Pannell, 2016).

5.1.3 The challenge of total soil organic carbon stock measurement

Currently, total SOC stock at field or farm scale can be realized using either a design-based or a model-based approach. In the design-based approach, stochasticity is introduced at the stage of sampling as sample locations are selected by a predetermined random selection procedure. The probability of selection of each sample is accounted for within the estimation of SOC stock. In the model-based approach the soil-forming process which has led to the distribution of SOC in the study area is modelled as a stochastic process. The prediction relies on a spatial model whose parameters are fixed quantities inferred from the observations (Brus and De Gruijter, 1997).

In both design-based and model-based approaches, stratification of the study area can be performed in order to cover the study area in a way that tends to maximize the accuracy of prediction. Stratification can be based on the geography (Viscarra Rossel and Brus, 2018), on ancillary variables (Simbahan *et al.*, 2006), on a soil map or on a map of predicted SOC. These supporting datasets are generally already existing or can be produced at limited cost. Besides, sample compositing can be performed to reduce local variability (local compositing) or to reduce laboratory costs (distant compositing, see for example de Gruijter *et al.*, 2016). Different protocols were developed that assemble and describe classical methods for total SOC quantification (e.g. Aynekulu *et al.*, 2011)

However, the major impediment to an effective and valuable reporting of C sequestration in soils is the lack of accurate, affordable and standardized methodologies for SOC quantification. Current methodologies are expensive and cannot easily detect small difference of total stock. Ancev (2011) estimated a high cost of measurement per ton CO_2 potentially sequestered, leading him to the conclusion that in general, mandating agriculture in an ETS – where monitoring stocks are compulsory even in the absence of agricultural actions – would not be socially beneficial. The Food and Agriculture Organization of the United Nations argued that the development of accurate methods for

measuring, monitoring and verifying changes in soil organic C stock is becoming increasingly important to fulfill national and international agreements and initiatives (FAO, 2017).

Recent technological and methodological advances were made in order to improve the cost-effectiveness of total SOC stock measurement. Pallasser *et al.* (2015) developed a method for determination of SOC in whole soil cores. It allows decreasing the analytical errors that arise due to carbon heterogeneity in large soil aliquots. Doetterl *et al.* (2013b) presented a method of SOC measurement by proximal sensing at very high depth resolution. It decreases significantly cost and labor in fine texture soils. Viscarra Rossel and Brus (2018) used proximal sensing to estimate the total SOC stock within a field. This method resulted in a reduction of the costs by a factor 2.1 compared to the traditional techniques and was based on stratification and analysis of composite samples. Besides, the uncertainty was reduced with the sensing method and a map of SOC stocks was produced that could help optimize the sampling design in subsequent rounds. de Gruijter *et al.* (2016) proposed an iterative procedure to optimize the stratification by exploiting a map of SOC stock and its uncertainty. The method aims at maximizing the expected profit for the farmer by taking into account price of sequestered C, sampling cost and risk inherent to uncertainty in the estimated amount of sequestered C.

5.1.4 Objective and plan of this chapter

Innovative and more efficient methodologies and techniques are being developed continuously for data collection, analysis and modelling. Therefore, developing a project of total SOC estimation over an area of interest requires multiple decisions to be taken in order to best allocate the limited amount of funds and efforts available. For a given cost, the best decisions are the ones that minimize the variance of prediction of the total stock. Therefore, it is essential to assess how the spatial distribution of SOC will influence this variance of prediction.

In this chapter, we aim at determining how the prediction of spatial variation is useful for the estimation of the total SOC stock over a study area A. To fulfill this objective, we propose a mathematical formalization of the problem of prediction of the total stock and associated uncertainty. In our framework (that can be qualified as model-based), the SOC stock by unit of surface is represented by a spatial random field which is the sum of the prediction of a model and a spatially dependent random residual (see Section 5.2.1).

In Section 5.2.2 we estimate the variance of prediction of the total stock as a function of both the extent of A and the spatial dependence of the model error. We assume that the predictor variables of the model are available at any point of A so that the model prediction can be computed at any location. In Section 5.2.3, we assume that SOC stock has also been measured at a set of sampling location. We exploit this information to spatially predict model error using the best linear unbiased predictor. Then, we assess how predicting model error with the samples influences the variance of prediction of the total

stock. Finally, in Section 5.2.4 we discuss two cases where the added value of estimating the spatial variation of SOC stock becomes negligible.

Finally, in Section 5.3.1 we propose a flowchart that recommends a methodology of total SOC stock estimation for when a prediction model is available along with covariates and, potentially, a set of SOC samplings. In Section 5.3.2 we analyze which decisions can be taken to maximize a priori, i.e. before data acquisition and model development, the cost-effectiveness of total stock quantification.

5.2 Mathematical formalization

5.2.1 Spatial integration of a random field

We place ourselves in the context of total SOC prediction made by spatially averaging over an area of interest (named A) the prediction $Z_p(\mathbf{u}) = E[Z(\mathbf{u})|\mathbf{Y}(\mathbf{u})]$ provided in any spatial location by a model built on a vector of auxiliary variables $\mathbf{Y}(\mathbf{u})$. We work under the hypotheses that the values of these variables are available in any point of A and that the total stock is obtained by spatial integration of $Z_p(\mathbf{u})$ over A (approximated by a Riemann sum).

Let us consider the spatially distributed random variable $Z(\mathbf{u})$ over a continuous spatial domain $A \in \mathbb{R}^2$, where $\mathbf{u} \in A$. What we seek for is the the mean $\overline{Z}_A$ over the area $|A|$ as given by

$$\overline{Z}_A = \frac{1}{|A|} \int_A Z(\mathbf{u}) d\mathbf{u} \tag{5.1}$$

Typically, this integral is approximated to an arbitrary precision using a Riemann sum based on a fine discretization of the domain, with $A = \cup_{i=1}^{n_A} A_i$ where $A_i \cap A_{i'} = \varnothing \ \forall i, i'$, so that $|A| = \sum_{i=1}^{n_A} |A_i|$ and thus

$$\overline{Z}_A \simeq \frac{1}{|A|} \sum_{i=1}^{n_A} Z(\mathbf{u}_i)|A_i| \tag{5.2}$$

where $Z(\mathbf{u}_i)$ is the value for $Z(\mathbf{u})$ associated with A_i.

Let us consider that the values for $Z(\mathbf{u})$ are predicted based on a non perfect model that makes use of a vector of sampled continuous valued auxiliary variables $\mathbf{Y}(\mathbf{u}) = (Y_1(\mathbf{u}), \ldots, Y_n(\mathbf{u}))'$ at the same location $\mathbf{u}$ and a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_p)'$ of parameters, with

$$Z(\mathbf{u}) = g(\mathbf{Y}(\mathbf{u}); \boldsymbol{\alpha}) + \mathcal{E}(\mathbf{u}) \tag{5.3}$$

where $g(.)$ is an arbitrary (and possibly nonlinear) functional and where $\mathcal{E}(\mathbf{u})$ is a zero-mean second-order stationary gaussian random residual with variance $\sigma^2_{\mathcal{E}}$, that accounts for all possible unidentified variables and measurements errors. From these assumptions, it is clear that so far the only random error to

be accounted for is linked to the spatial error $\mathcal{E}(\mathbf{u})$ that comes from the use of a non perfect prediction model for $Z(\mathbf{u})$, as we assume that $\mathbf{Y}(\mathbf{u})$ is sampled (and so the corresponding values are at hand).

Practically, $g(.)$ can figure any type of model of SOC prediction based on soil and environmental (including anthropic) covariates. The only constraint imposed is that $\mathbf{Y}(\mathbf{u})$ is known exhaustively over the study area – or practically, that we have a value for each nodes of a fine spatial grid – and that no spatial interpolation occurs within g itself. The model g could thus be based on the different frameworks for SOC prediction presented in Section 1.4 such as DSM, statistical modelling or machine learning. Within the framework of DSM, $\mathbf{Y}(\mathbf{u})$ can be a set of spectral information obtained by proximal sensing or remote sensing.

We will consider here that $\mathcal{E}(\mathbf{u})$ can be viewed as a spatial random variable coming from a zero-mean second-order stationary spatial random field $\mathfrak{E} = \{\mathcal{E}(\mathbf{u}) : \mathbf{u} \in A, \mathcal{E}(\mathbf{u}) \in \mathbb{R}\}$ such that $E[\mathcal{E}(\mathbf{u})] = 0$, with a covariance function $C_{\mathcal{E}}(\mathbf{h}) \triangleq Cov[\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u}')]$ where $\mathbf{h} = \mathbf{u}' - \mathbf{u}$. Typically, this covariance function can be inferred from a calibration data set where $(\mathbf{Y}(\mathbf{u}), Z(\mathbf{u}))$ are jointly measured at a set of locations so that the corresponding $\mathcal{E}(\mathbf{u}) = Z(\mathbf{u}) - g(\mathbf{Y}(\mathbf{u}); \boldsymbol{\alpha})$ are at hand for conducting inference about $C_{\mathcal{E}}(\mathbf{h})$.

It is worth noting that, in general, the values for the parameters $\boldsymbol{\alpha}$ are unknown and thus need to be inferred in some way. However, as our concern is not about propagating the uncertainty of the model estimation in the prediction, we will consider them as certain for the subsequent developments. Moreover, the corresponding uncertainties about them are sometimes out of reach, typically for situations where the model is based on previous studies and is given "as is"', without any clear information about the quality of the parameters estimation.

5.2.2 Influence of model error on total stock

In this section we highlight the link between the prediction variance of the total stock over A as obtained by the model and the range of the spatial dependence of the error of this model. We focus on two limit cases. The first case is when the range of spatial dependence is negligible compared to the size of A. The second case is when the range of spatial dependence largely exceeds the size of A.

As our goal is to predict at best $Z(\mathbf{u})$ based on the set of sampled auxiliary variables $\mathbf{Y}(\mathbf{u})$ at the same location, the best predictor in the mean square sense for $Z(\mathbf{u})$ is the conditional expectation, i.e.,

$$Z^p(\mathbf{u}) = E[Z(\mathbf{u})|\mathbf{Y}(\mathbf{u})] = g(\mathbf{Y}(\mathbf{u})); \boldsymbol{\alpha}) \tag{5.4}$$

where from eq. (5.3) the corresponding expected prediction error and prediction variance are given by

$$\begin{aligned} E[Z(\mathbf{u}) - Z^p(\mathbf{u})] &= E[\mathcal{E}(\mathbf{u})] = 0 \\ Var[Z(\mathbf{u}) - Z^p(\mathbf{u})] &= Var[\mathcal{E}(\mathbf{u})] = \sigma^2_{\mathcal{E}} \end{aligned} \tag{5.5}$$

Based on these last results, using now $Z^p(\mathbf{u})$ instead of $Z(\mathbf{u})$ in eq. (5.1) leads to a predictor of $\overline{Z}_A$, with

$$\overline{Z}^p_A = \frac{1}{|A|}\int_A Z^p(\mathbf{u})d\mathbf{u} \simeq \frac{1}{|A|}\sum_{i=1}^{n_A} Z^p(\mathbf{u}_i)|A_i| \tag{5.6}$$

As coming from eqs. (5.1), (5.2), (5.3), (5.4) and (5.6), we have

$$\overline{Z}_A - \overline{Z}^p_A = \frac{1}{|A|}\int_A \mathcal{E}(\mathbf{u})d\mathbf{u} \simeq \frac{1}{|A|}\sum_{i=1}^{n_A} \mathcal{E}(\mathbf{u}_i)|A_i| \tag{5.7}$$

The expected prediction error and the prediction variance are thus given by

$$\begin{aligned} E[\overline{Z}_A - \overline{Z}^p_A] &= \frac{1}{|A|}\int_A E[\mathcal{E}(\mathbf{u})]d\mathbf{u} = 0 \\ Var[\overline{Z}_A - \overline{Z}^p_A] &= \frac{1}{|A|^2}\int_A\int_A Cov[\mathcal{E}(\mathbf{u}), \mathcal{E}(\mathbf{u}')]d\mathbf{u}d\mathbf{u}' \\ &\simeq \frac{1}{|A|^2}\sum_{i=1}^{n_A}\sum_{i'=1}^{n_A} Cov[\mathcal{E}(\mathbf{u}_i), \mathcal{E}(\mathbf{u}_{i'})]|A_i||A_{i'}| \end{aligned} \tag{5.8}$$

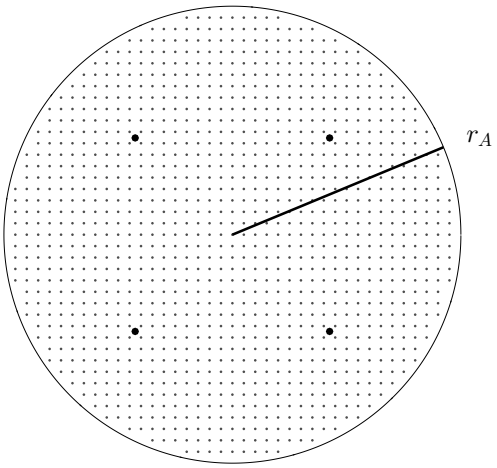
Clearly, the prediction variance only depends here on the spatial dependence for $\mathcal{E}(\mathbf{u})$ and on the domain A. For most practical cases, one expects that $\mathcal{E}(\mathbf{u})$ will be spatially correlated to some extent, so that the question is to evaluate the impact of this correlation as a function of the size of the domain A where $\overline{Z}_A$ is sought for. As the results will depend both on the geometry of A and on the spatial dependence structure that one considers for $\mathcal{E}(\mathbf{u})$, we will make here some choices for the sake of illustration. Let us consider that A is a circular domain of radius r_A and that the covariance function $C_{\mathcal{E}}(\mathbf{h})$ is an exponential model with

$$C_{\mathcal{E}}(\mathbf{h}) = \sigma^2_{\mathcal{E}}\ \exp(-\frac{3||\mathbf{h}||}{r_{exp}}) \tag{5.9}$$

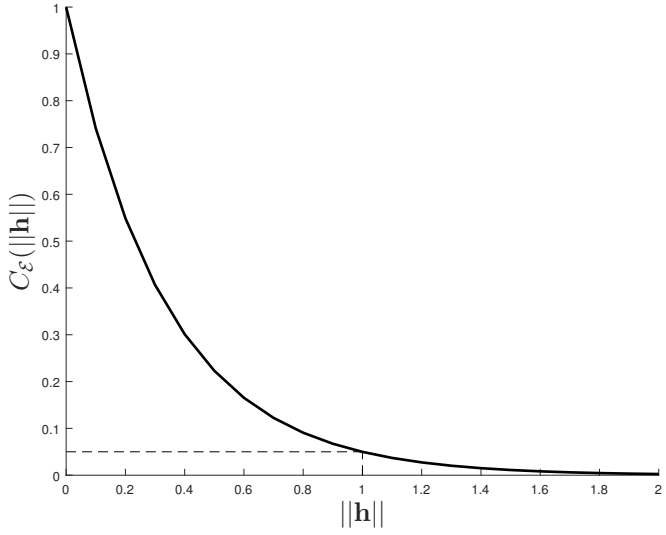
with $\sigma^2_{\mathcal{E}} = 1$ for the sake of simplicity, and where $r_{exp} \geq 0$ corresponds to the practical range of the model, i.e. the distance r_{exp} for which the model is equal to 5% of the σ^2_{exp} value (see Figure 5.1).

We can now easily assess the effect of r_{exp}/r_A on the prediction variance. It is worth considering two limit cases, i.e. when $r_{exp}/r_A \to 0$ and $r_{exp}/r_A \to \infty$, for which we have

$$\begin{aligned} \lim_{r_{exp}/r_A \to 0} C_{\mathcal{E}}(\mathbf{u} - \mathbf{u}') &= \sigma^2_{\mathcal{E}}\ \delta(||\mathbf{u} - \mathbf{u}'||) \\ \lim_{r_{exp}/r_A \to \infty} C_{\mathcal{E}}(\mathbf{u} - \mathbf{u}') &= \sigma^2_{\mathcal{E}} \end{aligned} \tag{5.10}$$



(a)



(b)

Figure 5.1: (a) A disk area A of radius r_A over which $\overline{Z}_A$ is sought for. The regular grid of dots corresponds to the discretization of the disk into small areas A_i, while the four big dots are the locations where a sample is at hand. (b) Exponential covariance model $C_{\mathcal{E}}(||\mathbf{h}||)$ with range r_{exp} equal to 1, so that $C_{\mathcal{E}}(r_{exp}) = 0.05\ \sigma^2_{\mathcal{E}}$ (where $\sigma^2_{\mathcal{E}}$ is chosen equal to 1 here for the sake of simplicity).

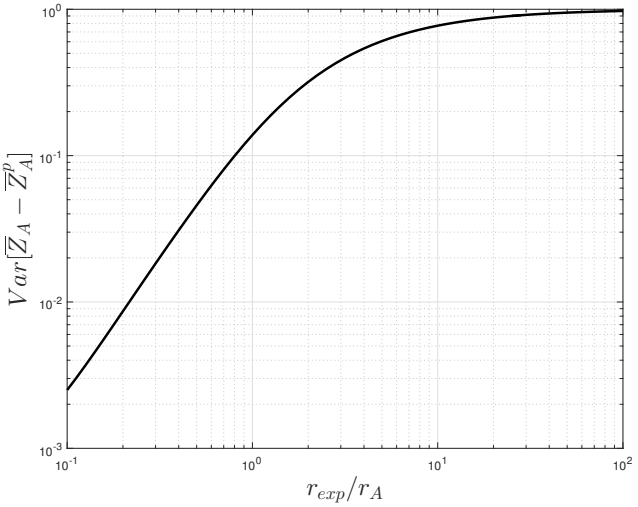


Figure 5.2: Prediction variance as a function of the relative range r_{exp}/r_A for the predictor in eq. (5.4) that neglects the errors.

where $\delta(.)$ is the Kronecker delta function, i.e. the so-called pure nugget effect correlation model. At the scale of A, the first result corresponds to the case of a non correlated process, while the second one is the case of a perfectly correlated process. Clearly, this last case implies that $\mathcal{E}(\mathbf{u}) = \mathcal{E}(\mathbf{u}')$, i.e. the random field behaves as a random constant over A. Accordingly, from eq. (5.8) the corresponding prediction variances are given by

$$\begin{aligned} \lim_{r_{exp}/r_A \to 0} Var[\overline{Z}_A - \overline{Z}_A^p] &= 0 \\ \lim_{r_{exp}/r_A \to \infty} Var[\overline{Z}_A - \overline{Z}_A^p] &= \sigma_{\mathcal{E}}^2 \end{aligned} \tag{5.11}$$

For intermediate situations, the corresponding prediction variances are given in Figure 5.2.

It is worth noting that the previous developments are relying on the fact that the various $\mathbf{Y}(\mathbf{u})$ are sampled everywhere over A, so that $g(\mathbf{Y}(\mathbf{u}); \boldsymbol{\alpha})$ can be computed for any location $\mathbf{u} \in A$.

5.2.3 Spatial prediction of model error using samples

In this second step we generalize previous results in order to improve model prediction by accounting for the spatial dependence of model error. We use the best unbiased linear predictor on a set of points for which the real values of $Z(\mathbf{u})$ are known. This allows us to describe the general formulas of the prediction variance for a predictor that uses simultaneously the prediction of the model and the interpolation of its errors. We show that variance of prediction tends to zero for two limit cases (very short or very long range of spatial dependence at

the scale of A) while it is maximum for ranges of the same order of magnitude as the size of A.

As long as $\mathcal{E}(\mathbf{u})$ is spatially correlated, it is tempting to try reducing the prediction variance by predicting the $\mathcal{E}(\mathbf{u})$ based on a set of locations where the error is at hand. Typically, this is the case for situations where a calibration data set is at hand, as explained before, so that for these sampled locations $\mathbf{v}_1, \ldots, \mathbf{v}_m \in A$ the errors $\mathcal{E}(\mathbf{v}_i) = Z(\mathbf{v}_i) - g(\mathbf{Y}(\mathbf{v}_i); \boldsymbol{\alpha})$ are at hand. Let us define the vector $\mathcal{E}_{\mathbf{v}} = (\mathcal{E}(\mathbf{v}_1), \ldots, \mathcal{E}(\mathbf{v}_m))'$ from which a predictor $\mathcal{E}^p(\mathbf{u})$ is sought for at any arbitrary location $\mathbf{u}$. For this task, minimum variance unbiased linear predictors are routinely obtained by relying on the so-called kriging approach, with

$$\mathcal{E}^p(\mathbf{u}) = \boldsymbol{\lambda}'(\mathbf{u})\mathcal{E}_{\mathbf{v}} \tag{5.12}$$

where $\boldsymbol{\lambda}(\mathbf{u})$ is a vector of optimal weights. Remembering here that $E[\mathcal{E}(\mathbf{u})] = 0$ $\forall \mathbf{u} \in A$, these weights are here the solution of a simple linear equation systems (the so-called simple kriging equations) given by

$$\Sigma_{\mathbf{v}}\boldsymbol{\lambda}(\mathbf{u}) = \boldsymbol{\sigma}(\mathbf{u}) \quad \Longleftrightarrow \quad \boldsymbol{\lambda}(\mathbf{u}) = \Sigma_{\mathbf{v}}^{-1}\boldsymbol{\sigma}(\mathbf{u}) \tag{5.13}$$

where $\Sigma_{\mathbf{v}}$ is the covariance matrix of the vector $\mathcal{E}_{\mathbf{v}}$ as built from the covariance function $C_{\mathcal{E}}(\mathbf{h})$, while $\boldsymbol{\sigma}(\mathbf{u})$ is the vector of covariance between $\mathcal{E}(\mathbf{u})$ and $\mathcal{E}_{\mathbf{v}}$. Accordingly, the expected prediction error and prediction variance are given by

$$\begin{aligned} E[\mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u})] &= 0 \\ Var[\mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u})] &\equiv Var[\mathcal{E}(\mathbf{u})] - 2Cov[\mathcal{E}(\mathbf{u}), \mathcal{E}^p(\mathbf{u})] + Var[\mathcal{E}^p(\mathbf{u})] \\ &= \sigma_{\mathcal{E}}^2 - 2\boldsymbol{\lambda}'(\mathbf{u})\boldsymbol{\sigma}(\mathbf{u}) + \boldsymbol{\lambda}'(\mathbf{u})\Sigma_{\mathbf{v}}\boldsymbol{\lambda}(\mathbf{u}) \end{aligned} \tag{5.14}$$

By comparison with eq. (5.4), a better predictor of $Z(\mathbf{u})$ is now given by

$$Z^p(\mathbf{u}) = g(\mathbf{Y}(\mathbf{u}); \boldsymbol{\alpha}) + \mathcal{E}^p(\mathbf{u}) \tag{5.15}$$

with a prediction error

$$Z(\mathbf{u}) - Z^p(\mathbf{u}) = \mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u}) \tag{5.16}$$

so that the expected prediction error and prediction variance are directly given by eq. (5.14), with

$$\begin{aligned} E[Z(\mathbf{u}) - Z^p(\mathbf{u})] &= 0 \\ Var[Z(\mathbf{u}) - Z^p(\mathbf{u})] &= \sigma_{\mathcal{E}}^2 - 2\boldsymbol{\lambda}'(\mathbf{u})\boldsymbol{\sigma}(\mathbf{u}) + \boldsymbol{\lambda}'(\mathbf{u})\Sigma_{\mathbf{v}}\boldsymbol{\lambda}(\mathbf{u}) \end{aligned} \tag{5.17}$$

and where this new predictor can be used in eq. (5.6). It is worth remarking here that averaging the predictors $Z^p(\mathbf{u})$ over A for predicting $\overline{Z}_A$ is not the usual theoretical approach. Indeed, the best linear minimum variance unbiased predictor of $\overline{Z}_A$ can directly be obtained by solving the so-called block-kriging equations, that directly aim at obtaining a single vector of optimal weights $\boldsymbol{\lambda}$ so that the block-kriging predictor of $\overline{Z}_A$ is directly given by $\boldsymbol{\lambda}'\mathcal{E}_{\mathbf{v}}$. However, it is easily proved that both ways lead to the same final result (see e.g. Olea, 1999).

Resuming from eqs. (5.1), (5.3), (5.6) and (5.15), we now have

$$\overline{Z}_A - \overline{Z}_A^p = \frac{1}{|A|}\int_A \mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u})d\mathbf{u} \simeq \frac{1}{|A|}\sum_{i=1}^{n_A}\big(\mathcal{E}(\mathbf{u}_i) - \mathcal{E}^p(\mathbf{u}_i)\big)|A_i| \quad (5.18)$$

so that the expected prediction error and prediction variance are now

$$E[\overline{Z}_A - \overline{Z}_A^p] = \frac{1}{|A|}\int_A E[\mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u})]d\mathbf{u} = 0$$

$$\begin{aligned} Var[\overline{Z}_A - \overline{Z}_A^p] &= \frac{1}{|A|^2}\int_A\int_A Cov[\mathcal{E}(\mathbf{u}) - \mathcal{E}^p(\mathbf{u}), \mathcal{E}(\mathbf{u}') - \mathcal{E}^p(\mathbf{u}')]d\mathbf{u}d\mathbf{u}' \\ &\simeq \frac{1}{|A|^2}\sum_{i=1}^{n_A}\sum_{i'=1}^{n_A} Cov[\mathcal{E}(\mathbf{u}_i) - \mathcal{E}^p(\mathbf{u}_i), \mathcal{E}(\mathbf{u}_{i'}) - \mathcal{E}^p(\mathbf{u}_{i'})]|A_i||A_{i'}| \end{aligned} \quad (5.19)$$

Computing the prediction variance in eq. (5.19) requires the computation of the covariance between prediction errors for an arbitrary pair of locations $(\mathbf{u}, \mathbf{u}')$. Going back to eq. (5.12), the predictors needed in the approximation of eq. (5.19) can be jointly obtained with

$$\begin{bmatrix} \mathcal{E}^p(\mathbf{u}_1) \\ \vdots \\ \mathcal{E}^p(\mathbf{u}_{n_A}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}'(\mathbf{u}_1) \\ \vdots \\ \boldsymbol{\lambda}'(\mathbf{u}_{n_A}) \end{bmatrix} \mathcal{E}_\mathbf{v} \quad \Longleftrightarrow \quad \mathcal{E}_\mathbf{u}^p = \boldsymbol{\Lambda}'\mathcal{E}_\mathbf{v} \quad (5.20)$$

where from eq. (5.13) we have

$$\Sigma_\mathbf{v}\boldsymbol{\Lambda} = \boldsymbol{\sigma} \quad \Longleftrightarrow \quad \boldsymbol{\Lambda} = \Sigma_\mathbf{v}^{-1}\boldsymbol{\sigma} \quad (5.21)$$

with $\boldsymbol{\sigma} = (\boldsymbol{\sigma}(\mathbf{u}_1), \ldots, \boldsymbol{\sigma}(\mathbf{u}_{n_A}))$, so that the prediction errors are given by

$$\mathcal{E}_\mathbf{u} - \mathcal{E}_\mathbf{u}^p = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Lambda}' \end{bmatrix} \begin{bmatrix} \mathcal{E}_\mathbf{u} \\ \mathcal{E}_\mathbf{v} \end{bmatrix} \quad (5.22)$$

and the corresponding covariance matrix for the prediction errors is

$$\begin{aligned} Cov[\mathcal{E}_\mathbf{u} - \mathcal{E}_\mathbf{u}^p] &\triangleq \begin{bmatrix} Var[\mathcal{E}(\mathbf{u}_1) - \mathcal{E}^p(\mathbf{u}_1)] & \cdots & Cov[\mathcal{E}(\mathbf{u}_1) - \mathcal{E}^p(\mathbf{u}_1), \mathcal{E}(\mathbf{u}_{n_A}) - \mathcal{E}^p(\mathbf{u}_{n_A})] \\ \vdots & \ddots & \vdots \\ (sym) & \cdots & Var[\mathcal{E}(\mathbf{u}_{n_A}) - \mathcal{E}^p(\mathbf{u}_{n_A})] \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Lambda}' \end{bmatrix} \begin{bmatrix} \Sigma_\mathbf{u} & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \Sigma_\mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\boldsymbol{\Lambda} \end{bmatrix} = \Sigma_\mathbf{u} - 2\boldsymbol{\Lambda}'\boldsymbol{\sigma} + \boldsymbol{\Lambda}'\Sigma_\mathbf{v}\boldsymbol{\Lambda} \end{aligned} \quad (5.23)$$

where $\Sigma_\mathbf{u}$ is the covariance matrix of $\mathcal{E}_\mathbf{u} = (\mathcal{E}(\mathbf{u}_1), \ldots, \mathcal{E}(\mathbf{u}_{n_A}))'$.

5.2.4 Simplification of the predictor in the limit cases

In this section we compare the prediction variance of the optimal linear predictor with the prediction variance corresponding of two specific cases. The first specific case is when only $E[Z(\mathbf{u})|\mathbf{Y}(\mathbf{u})]$ is used as predictor, either because the user neglects spatial information brought by the residuals or because these residuals are not accessible in the area of interest.

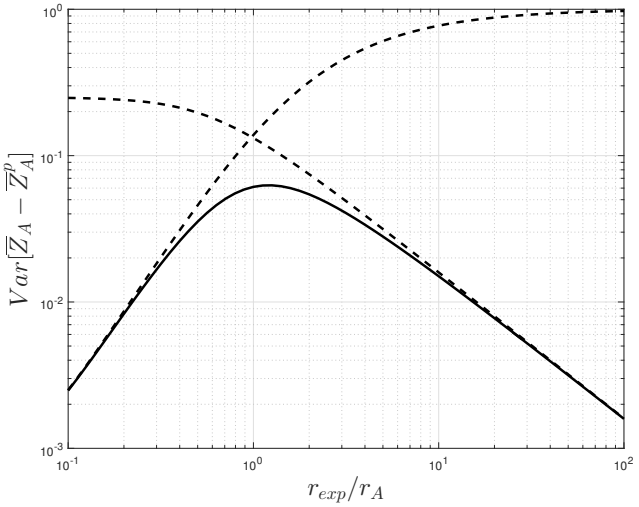


Figure 5.3: Prediction variance for the optimal predictor in eq. (5.15) (plain curve), along with the prediction variances for the predictor that neglects the errors (increasing dashed curve) and for the predictor that corresponds to an arithmetic average of the four samples (decreasing dashed curve).

The second specific case is when the spatial prediction of the residuals is approximated by their simple arithmetic average.

Beyond the optimal estimation of the kriging weights $\boldsymbol{\lambda}(\mathbf{u})$ as provided by eq. (5.13), the formula for the prediction variance in eq. (5.19) allows us to assess the impact of any specific choice for these weights. As a first possible choice, let us consider the prediction variance that one obtains by neglecting the errors as in eq. (5.4), which is a special case of eq. (5.15) by considering that $\mathcal{E}^p(\mathbf{u}) = 0 \ \forall \mathbf{u}$, or equivalently by setting $\boldsymbol{\lambda}(\mathbf{u}) = \mathbf{0}$ in eq. (5.12). This leads to $\boldsymbol{\Lambda} = \mathbf{0}$, so that $Cov[\mathcal{E}_{\mathbf{u}} - \mathcal{E}^p_{\mathbf{u}}] = \Sigma_{\mathbf{u}}$ and eq. (5.19) yields the same result as eq. (5.8)). Another possible choice is to neglect the computation of optimal weights by considering a naive arithmetic average of the sample at hand, so that the predictor for the error becomes

$$\mathcal{E}^p(\mathbf{u}) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{E}(\mathbf{v}_i) = \frac{1}{m}\mathbf{1}'\mathcal{E}_{\mathbf{v}} \qquad \forall \mathbf{u} \tag{5.24}$$

which is equivalent to considering that $\boldsymbol{\lambda}(\mathbf{u}) = (1/m)\mathbf{1}$ in eq. (5.15), leading to $\boldsymbol{\Lambda} = (1/m)\mathbf{1}\mathbf{1}'$.

Obviously, these two possible choices will lead to corresponding prediction variances that are higher than the prediction variance for the optimal weights (as these weights precisely correspond to those that minimize the prediction variance). However, it can be shown too that the differences are negligible in some circumstances. For the sake of illustration, let us use the same covariance model as in eq. (5.9) in order to see how the prediction variance of $\overline{Z}_A^p$ is changing as a function of r_{exp}/r_A for the predictor given in eq. (5.15). We will assume that the same four sampled locations are at hand in order to conduct the prediction (so $m = 4$). The corresponding results are given in Figure 5.3 (plain line).

For the prediction variance of the predictor that makes use of optimal weights, one can see that it tends to 0 both when $r_{exp}/r_A \to 0$ and when $r_{exp}/r_A \to \infty$ (the formal proof is omitted here for the sake of brevity) and is maximum for a value of r_{exp}/r_A which is close to 1, i.e. when the extent of A as measured by r_A is in the same order of magnitude than the extent of the spatial dependence as measured by r_{exp}. By comparison, the prediction variances for the predictors $\mathcal{E}^p(\mathbf{u}) = 0$ and $\mathcal{E}^p(\mathbf{u}) = (1/m)\mathbf{1}'\mathcal{E}_{\mathbf{v}}$ exhibit different behaviours, where the first one simply corresponds to the result of Figure 5.2, while the second one is decreasing down to 0 as $r_{exp}/r_A \to \infty$.

When $r_{exp}/r_A \to 0$, it can be seen that neglecting the errors does not lead to any significant increase for the prediction variance of $\overline{Z}_A^p$. In contrast, using an arithmetic average is clearly a bad idea as the four points are not sufficient to make a precise estimation of $\frac{1}{|A|}\sum_{i=1}^{n_A}\mathcal{E}(\mathbf{u}_i)|A_i|$. On the opposite, when $r_{exp}/r_A \to \infty$, the arithmetic average yields the same result as for the optimal predictor. In this case, neglecting the error becomes clearly a bad idea.

It is also worth noting that two of the three curves of Figure 5.3 depend on the number of samples m (which is here equal to four). Indeed, sampling density influences both the best linear predictor (plain curve) and the arithmetic average of the samples (decreasing dashed curve). In the theoretical limit case where the sampling density was increased towards infinite value, the prediction variance would converge to 0. In contrast, the predictor that neglects the errors does not, by definition, rely on the sample values. Apart from this limit case it is likely that the general shape of the curves would stay the same with a change of sampling density.

Practically, there is no benefit in accounting for the errors instead of just neglecting them when the correlation range of these errors is much lower than the size of the domain A. Indeed, when $r_{exp}/r_A \to 0$, these errors are averaging out and disappear from the areal average $\overline{Z}_A$. When the correlation range is much higher than the size of the domain A, however, these errors start to behave like a single random constant that can easily be estimated from a sample over A whatever the specific weights that one consider, so that a crude arithmetic average obtained from a very limited sample size will do the job. For both limit cases, the optimal predictor is thus reduced to a predictor that yields a constant value over A. This constant value is equal to 0 when $r_{exp}/r_A \to 0$, while it is a random constant when $r_{exp}/r_A \to \infty$.

One could also question the absence of a nugget in the covariance model (eq. 5.9). Indeed, it is realistic to assume some non spatially dependent variability caused by the uneven structure of soil and organic materials at the extent of the sample. Besides, a nugget effect is also often included in the covariance model when the sampling design does not allow to estimate the covariance at very short distance. In this chapter, it was decided for the sake of simplification to use a model free of nugget component. This choice does not impact the fundamental results presented in this chapter as, whatever the approach selected (kriging, neglecting or averaging model errors), non spatially dependent spatial errors would average out in the spatial integration (eq. 5.1).

5.3 General Discussion

In the two following sections, we discuss how the findings of Section 5.2 can be exploited by surveyors to increase cost-effectiveness associated with total stock quantification. Section 5.3.1 recommends some simplification in the methodology of total stock estimation that can be done when the model as well as data on SOC and covariates are already at hand. Section 5.3.2 discusses the perspective of maximizing a priori, i.e. before data acquisition and model development, the cost-effectiveness of total stock quantification.

5.3.1 Choice of the methodology of prediction

As it has been presented in Section 5.2.4, spatially interpolating the errors of the model is not always necessary as simplification can be made without any serious loss of precision. Figure 5.4 proposes a decision support tool for total SOC stock quantification when a spatially exhaustive prediction model (i.e. $g(\mathbf{Y}(\mathbf{u});\boldsymbol{\alpha})$ in eq (5.3) is available along with covariates to feed the model and, potentially, SOC point measurements. The flowchart summarizes key elements that can be faced by the land surveyor and that will influence the choice of the method. For each cases it recommends the best way to manage model errors.

The first way to avoid spatial interpolation is to quantify total stock by simply integrating over A the spatial prediction of the model. This is what we refer to as the "predictor that neglects the errors". The second way is to approximate the model error by a single constant

over the study area. Its value is the mean over sampling locations of model errors. The total stock is calculated as the spatial integration of the model prediction plus this constant. This is what we refer to as the "predictor that averages the errors". Finally, if spatially interpolation of model error is done then kriging is used as it is the predictor that minimizes variance of prediction. This is what we refer to as the "best linear predictor". Incidentally, we point out that if samples are available, they can be used to calibrate or validate the model but they can also be used to estimate the parameters of the random error component $\mathcal{E}(\mathbf{u})$. On the contrary, if no samples are available, this absolutely means that the model was calibrated on another study area.

In the simplest case (upper left branch), the number of samples is sufficient to assess the spatial dependence of model error in a robust way. A model of spatial autocorrelation is fitted using preferentially a REML approach (see Appendix B) and a range parameter is derived. This range is then compared to the size of the study area. If it is much larger than the size of the study area then we recommend using the predictor that averages the errors. If, on the contrary, this range is much smaller than the size of the study area then we recommend using the predictor that neglects the errors. If none of these cases are encountered, the best linear predictor should be used.

Things become more complex when the available point samples do not allow to derive a reasonably good estimate of spatial dependence for the error (upper right branch). This may occur if points are not in sufficient number or are not distributed conveniently over the study area. Two alternatives are possible.

The first alternative is that the model was calibrated locally or on an area of similar soil and environmental properties. In that case we may assume that we the model prediction is unbiased on the study area. Again, we divide the problem in two cases according to the fact that the models accounts for influencing variables in an exhaustive way or not.

If all soil and environmental factors that are expected to influence SOC at the scale of the study area are efficiently accounted for in the model, we may assume that the range of spatial dependence of model error is much smaller that the size of the study area. Indeed, model error will mainly be composed of variation due to stochastic soil processes occurring at a scale smaller that the global, the national or even the field scale. For example, for a model based on remotely sensed hyperspectral Vis-NIR reflectance spectroscopy, it is admitted that a rigorous procedure of calibration (typically using data covering the whole range of variation of spectral features in the study area) will generate model error that is at most lowly spatially dependent (Hbirkou *et al.*, 2012b). In real situation, error could also arise due to imprecision in covariates and to the fact that the exact value of the covariates cannot be practically known in any point of the space. For example, using a soil map where soil properties are given by class intervals instead of exact numeric values could generate spatially dependent error at the border of soil map units if soil properties were actually smoothly varying over the study area. Similarly, the prediction of a model based on spectral information remotely sensed over a pixel grid would be imprecise at a scale finer than the corresponding pixel size. In these two cases, if the spatial units or the mesh size are small at the scale of the study area, which is generally the case, the same will be true for the range of spatial dependence of model error. Therefore, we recommend using the predictor that neglects the errors when all influencing factors are accounted for in the model.

If, on the contrary, important factors were not included in model covariates (and are not fully dependent on present model covariates), then the recommendation will depend on the expected range of variation of these factors. If this range is much smaller than the size of the study area, the we would again recommend using the predictor that neglects the errors. Otherwise, the best linear predictor should be used to minimize prediction error by exploiting point SOC information.

The second alternative appears when the model has been calibrated in another area than the study area and in a different state regarding one or multiple soil and environmental factors. In general, due to the high spatial and temporal variability of SOC and to the large number of related processes, most SOC model calibrations are not expected to be valid out of the area on which they were achieved. Even spectral models relying on large spectral libraries require local re-calibration on a small number of samples to ensure accurate prediction (Nocita *et al.*, 2015). More precisely, when a model whose parameters were derived on a given area

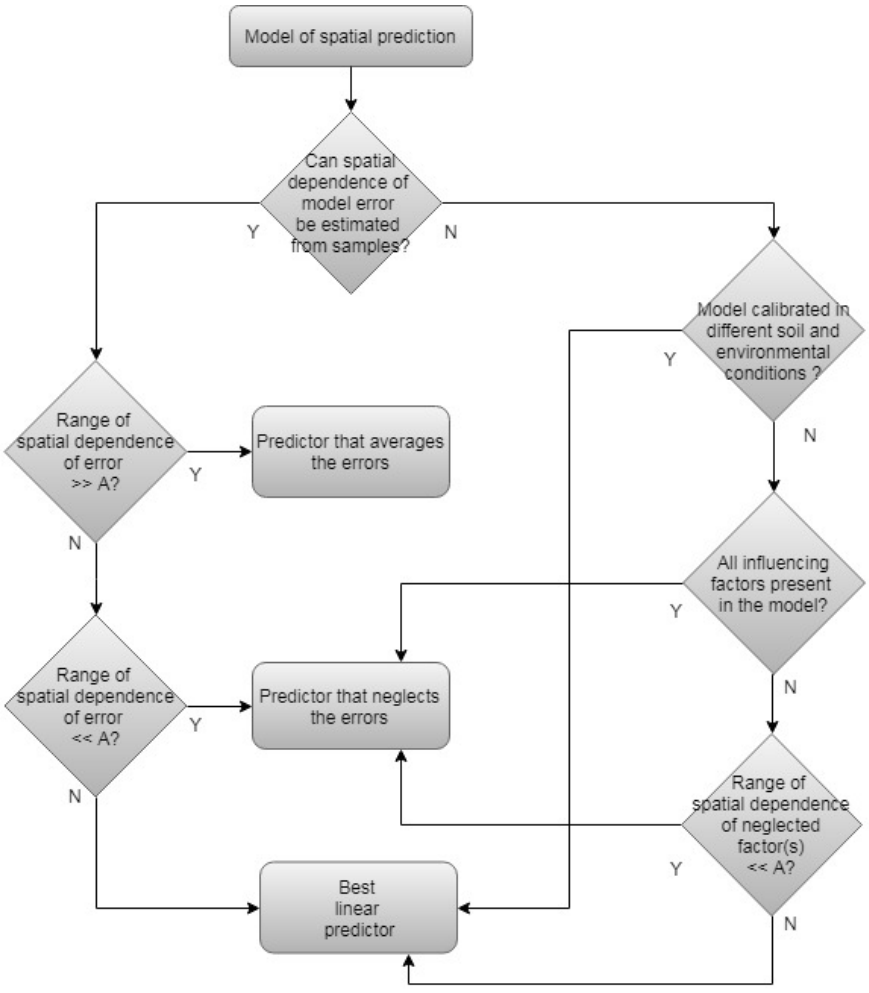


Figure 5.4: This flowchart is intended to serve as decision support tool for total SOC stock quantification when a spatially exhaustive prediction model is available along with covariates to feed the model and, potentially, SOC point measurements. In the simplest case (upper left branch), the number of samples is sufficient to assess the spatial dependence of model error in a robust way. Things become more complex when the available point samples do not allow to derive a reasonably good estimate of spatial dependence for the error (upper right branch). More details on the recommendations are given in Section 5.3.1)

is used on another, not similar study area, it is likely that only covariates varying within the calibration area were considered. Applying such a model on the study area, where the values of these neglected factors are different, would thus most likely lead to model error. For example, using a model calibrated in a region under a given climate and which accounts not explicitly for the climate in its covariates would likely generate serious model error if it is used in a region of different climate. In this case the best linear predictor should be used to maximize the precision on the total stock, but the results should be taken cautiously.

5.3.2 Implications for cost-effectiveness

The previous section indicated which methodology the land surveyor can preferentially use according to the knowledge he can gather on model properties and scale of model error. In contrast, in this section we aim at analyzing the question in the perspective of maximizing a priori, i.e. before data acquisition and model development, the cost-effectiveness of total stock quantification. Therefore, we determine which simplification to the full methodology can be done in order to reduce costs without notable loss of precision.

Firstly, not all the factors that influence SOC have to be measured and included in the model. It is generally not necessary to account in the model for factors that influence SOC at ranges that are much smaller than the size of the study area. Indeed, omitting these factors will most likely generate model error at the same scale and this error will be averaging out in the estimation of the total stock. For example, in Chapter 4 a large effect on SOC content related to the field boundaries was highlighted. Variation related to field was likely due to differences in current land management or in LULM history. Present results allow us to state that explicitly accounting for field limits in the model in not useful when the goal is to estimate total SOC stock at regional or larger scale, i.e. at a spatial scale which is several orders of magnitude largr than the one related to field boundaries. This is noteworthy as information over current and past LULM is generally sparse. More generally, this result provides insight to spare data acquisition or purchase costs.

Secondly, under some circumstances, the number of point samplings may be limited to minimum requirements. As an example, let us assume that one wants to estimate the evolution of total stock following the introduction of land management change in an undulating field. A process-based model that accounts for sediment redistribution and land management practices is at hand. Let us also assume that this model has been tested and validated over similar soil and environmental conditions and that the only error that has been observed is associated with a very small spatial range. In this case, we can consider that it is not necessary to collect point samples in order to validate model prediction, estimate the range of model error or perform kriging prediction. This again allows us to reduce cost and time requirements.

Thirdly, there is no need to perform model prediction at the finest spatial resolution. Spectral models that are used in remote sensing are almost free of spatial dependence in the error on a pixel basis when they are thoroughly calibrated (Hbirkou *et al.*, 2012b). Besides, in remote sensing, there is a well-known trade-off between the pixel resolution and the spatial coverage. Typically, for airplane remote sensing, higher flight altitude results in larger coverage but larger pixel size. When only total SOC is aimed at, using larger pixels is therefore a way to decrease the cost of measurement by unit of surface covered.Two restriction should be considered however. On the one hand, increasing pixel size is interesting if the performance of prediction of the model with the larger pixel size is not lessened relatively to the model with a smaller pixel size. On the other hand, as the size of pixels increases, the proportion of pixels that are only partially contained in the study area also increases and this generates uncertainty. In the same way, When covariates are soil and environmental factors, spatial prediction does not absolutely have to be performed at the finest resolution but are better applied at the resolution which is the most convenient for the model. Indeed, many of the models used in the environmental sciences are scale-specific and should not be applied uncritically to scales for which they were not developed (Heuvelink, 1998).

Fourthly, we want to go beyond the case where the random effect was assumed to present a single spatial scale of variation, with a single exponential function used as covariance function. Let us now assume that the random effect presents a multiscale variation and that the best

covariance model results in a sum of covariance function with contrasting range parameters. The idea is thus the same that the one which was developed in Chapter 3, where it is also shown that, because of the linear properties of the covariance, a random field with multiple scales of variation can be viewed as as a constant plus the random fields corresponding to these ranges. In Figure 5.3, we observe that the interest of using the kriging predictor instead of the predictor that neglects the error or the arithmetic average of the samples decreases as $r_{exp} \gg r_A$ or $r_{exp} \ll r_A$. Building on this fact, one may infer that in the case of multiscale variation, a kriging prediction does not need to be based on the complete covariance function. Specifically, there is no need to account for error component(s) with range of variation that are much smaller than the size of the study area. This means that kriging can be performed using a covariance model from which the covariance functions related to these ranges are neglected. In the context of improving cost-effectiveness, this also means that it is not necessary to estimate the variogram at distances that are very short compared to r_A, as this would not lead to any noticeable gain in the quality of the estimation of the total stock. It is therefore not necessary to take point samples at short distance within the study area in these instances.

Lastly, we consider the situation where no model are at hand to predict the spatial variation of SOC. Simply said, total stock estimation relies only on point samples. In this case, one may ask if design-based approach, where sampling locations are selected by probability sampling, would bring better results than model-based approach. Some comparison between both approaches with equal numbers of data points highlighted that design-based approach performed better where there were abrupt changes, when an estimate of the global mean had to be done with a small sample size or when the best confidence interval was wanted. In contrast, model-based approach performed better where the soil varied gradually or when an estimate of the local mean had to be used (Brus and De Gruijter, 1997). For the case of total stock prediction (i.e. when only "how much?" matters and not "where?"), these authors recommend the use of model-based upon design-based approach only when objectivity and strict validity are not required and when SOC values at points are autocorrelated. The lack of objectivity of the model-based approach lies in the selection of the sampling locations and the choice of the model. Strict validity refers to the estimates of the estimation variance and confidence intervals. In the model-based approach several assumptions need to be made, e.g. about stationarity, isotropy and function type which prevents a formal guarantee for the validity of the results.

The question that arises is how the developments presented here would help in estimating total SOC stock change over the area. We consider the case where soil conservation management practices have been implemented in the area and we want to assess their effect. We also assume that the model g is valid at the two times at which the total stock is estimated. However, we do not consider land management as predictor variable in g as we aim at verifying the influence of land management and not at modelling it. This means that observed change would depend only on the samples. Therefore, it cannot be envisaged here to neglect model errors. The arithmetic average of model error measured at sampling locations would be convenient if the range of the variogram is largely larger than the size of A. Otherwise, predicting model error would be necessary.

5.4 Conclusion

In this chapter, we focused on the generic problem which consists in estimating at best the total stock of SOC (or equivalently the stock of SOC per surface unit) over a continuous spatial domain of interest, where SOC stock is viewed as a two-dimensional geographically distributed variable. Total SOC stock was estimated by integration over an area of interest of the prediction provided in any spatial location by a model built on a vector of auxiliary variables. This allowed us to highlight the link between the prediction variance of the total stock over A as obtained by the model and the range of the spatial dependence of the error of this model. In a next step, previous results were generalized in order to improve model prediction by accounting for the spatial dependence of model error. We compared the prediction variance of the kriging predictor with the prediction variance corresponding of two specific cases. Theoretical results indicate that, if kriging shows the lowest variance

of prediction for the total stock (as it is the optimal linear predictor), there are situations in which spatial interpolation of model error is not necessary and simplification can be made without any serious loss of precision.

In a next step, a flowchart diagram was presented that summarizes all situations that can be met in real conditions, with a recommended methodology for each case. The recommendation depends on the possibility to assess spatial dependence of model error, on the exhaustivity of factors considered in the model and on the range of model error in comparison to the size of the study area. Besides, recommendations were also made in order to improve cost-effectiveness for the frequent cases where SOC and auxiliary variables have not been measured yet before total stock quantification is decided. These recommendations allow the land surveyor to allocate efficiently measurement and computation efforts by focusing on elements that will significantly decrease the prediction variance on the total SOC stock and neglecting factors and computational steps that will bring only a negligible added value.

Reliable and low-cost methodologies for measuring total SOC stock are required in order to integrate efficiently SOC in global inventory and in national and international sequestration policies and trading schemes. Although recent advances have been proposed these last years, there is still a lack of broadly accepted methodological references able to estimate total stock at different spatial scales, in various soil and environmental contexts and at affordable price. With this chapter, our hope is to help developing and promoting such standard, flexible and cost-effective methodologies.

Chapter 6

General conclusion

6.1 Main findings

During this thesis, we have examined the spatial variability of SOC content in cropland from different angles. This allowed us to acquire a deeper knowledge of the factors controlling this spatial variability and the scale at which they exert their influence. Besides, datasets of various format, size and quality were used and this provided us insight about their specific potential, limitations and pitfalls. Finally, Chapter 5 explored how predicting spatial variation of SOC is useful for the estimation of the total SOC stock. In the following paragraphs we propose to review how we met the objectives.

6.1.1 Identify and quantify the effect of the factors controlling the spatial variability of SOC concentration and stock

The large number of local applications of digital soil mapping have exposed the control of different factors, their importance varying widely with the regional context and the spatial scale considered. Here we summarize and discuss the dominant links between SOC, soil and environmental variables that were observed.

Soil type

Soil type is a very generic term. A multitude of soil classification at multiple geographic levels exist. Soil classes or types are defined according to one or an ensemble of features such as mechanism of soil formation, parent material, profile development, texture, drainage, land use or chemical composition. For this reason, instead of considering soil type, we propose to analyze some of these properties separately.

Texture

Texture showed a significant influence both in the loam belt and over the Grand-Duchy of Luxembourg.

In the loam belt, for a fixed class of drainage or profile development, soil profiles with a larger proportion of silt (loam and clay-loam) contained on average a larger concentration of SOC, at every depth layer considered, than soil profiles with, on contrary, a larger proportion

of sand (light and heavy sandy loam). This is generally attributed to physical protection of SOM through association with clay and silt particles. The differences were however limited, likely due to the relative textural homogeneity of the region. Results also showed that the relative differences of SOC concentration between texture classes increased with soil depth.

In Luxembourg, the average SOC surface concentration by texture class could be derived from a map of soil textural units and an airborne image of SOC. The differences between classes were large. For example, the expected concentration in loamy locations was more than twice the expected concentration in sandy-loam locations, with high significance. Such a large difference is explained by the fact that loamy soils compose the northern region of Oesling. Particularities of climate, geology and agriculture are therefore expected to have a large influence on soil. In the southern region of Gutland, most represented textures are clay and sandy-loam. Clay soils are still expected to contain 60 % more SOC at the surface than sandy-loamy soils.

Drainage

In the loam belt, well drained soil profiles displayed higher concentrations of SOC than poorly drained soil profiles of the same profile development class. On the one hand, water saturation can inhibit root and plant development due to the lack of oxygen in soil, therefore reducing photosynthesis and biomass input. On the other hand, the activity of microorganisms also decreases below a threshold due to non optimal wetness conditions, leading to a lower decomposition of SOC. The lower SOC content associated with poorly drained soils seems thus to indicate that the combination of reduced biomass input (unfavourable for SOC concentration) along with reduced decomposition from microorganisms (favourable for SOC concentration) had on average a detrimental effect on SOC content within the context of our analysis. Besides, we also observed a positive correlation between finer textural classes and well drained soils in the region, with thicker layers of loess deposits associated with deeper water tables. This means that the larger SOC concentration observed with a better drainage was partly an effect of the texture too.

Profile development and topography

In the loam belt, colluvial profiles displayed on average a higher concentration of SOC than non colluvial profiles of the same texture or drainage class. The difference was small in the top layer and became larger with depth.

Soil in colluvial profiles can be seen as former topsoil that has been transported as sediment from upper slope. According to current knowledge, the higher SOC content below the top layer was attributed to the decreased mineralization rate of buried sediment through physical protection mechanisms. However, SOC content in the subsoil of colluvial profiles was still on average largely inferior to SOC content in the topsoil of both colluvial and non-colluvial profiles. This indicates that a large part of SOM contained in sediments is mineralized during transport or/and after burial at the colluvium.

In the loam belt, because of the quality of available dataset, topography was analyzed for a smaller area of 40 $\times$ 40 km with a more homogeneous texture. The effect of topography was observed mainly for layers in the subsoil. The third of the profiles with highest slope gradient values displayed on average lower SOC content than the others in subsoil because of a most frequent occurrence of shallow soils. In contrast, the third of the profiles with highest wetness index displayed higher content of SOC, because wetness index is related to sediment deposition and colluvial soils.

A linear model based on topographic indices alone was able to select the best combination of predictors along with terms of interactions. It predicted a higher SOC content in concave parts of the fields, where deposition of sediments occurs. Below the topsoil layer, the stock predicted in depositional locations could reach approximately twice of the stock estimated in convex locations.

Agricultural practices

The effect of agricultural practices on landscape scale SOC variability was rigorously identified and quantified in Luxembourg. In small regions of homogeneous texture, variation of surface SOC content associated with the pattern of fields represented typically between a quarter and half of total variation.

In the loam belt, we also expect agricultural management to cause variation from a field to another. However, available data did not allow more in-depth analyzes. While the R_AW dataset was sampled by toposequence, the number of samples by field was not sufficient to estimate a "field effect" in a robust way. In other words, incorporating the field ID as random variable in the prediction model did not allow a robust estimation of the effect of other predictors (topographic indices) and prediction of the SOC content.

A dataset on current and historical agricultural management (like a coefficient of cultivation intensity or tillage intensity) for each field of the area (or at least with a fair spatial coverage) would have been a precious variable for the prediction model. However, if various surveys based on questionnaires have been conducted in the past, no consistent spatialized dataset was found.

In the loam belt, the effect of agricultural management on SOC content acted thus as a perturbation for the model and this can contribute to explain the relatively weak coefficient of determination that was obtained.

6.1.2 Identify the influence of the spatial scale at which the different factors control the spatial distribution of SOC concentration and stock.

Each chapter is an attempt to bring answers to the questions raised by the preceding ones. In particular, since different factors are acting at different scales, we had to focus on the scale at which variation of SOC remains unexplained in order to identify the missing factors of influence. In the following paragraphs we explain why the second objective (identify the influence of spatial scale) arises from an iterative questioning and continuous advancement towards the first objective (identify and quantify controlling factors).

The starting hypothesis was that sediment redistribution had a critical influence on the spatial distribution of SOC stocks in sloping landscapes. This hypothesis was supported by the multiple evidences of the effect of soil redistribution on processes related to SOC (see section 1.2.3). The objective of Chapter 2 was therefore to assess the importance of the topographic control on SOC spatial distribution over cropland. The loam belt was chosen because of its undulating topography, its intensive cultivation and the availability of a large number of soil samples. A link between topography and SOC was eventually highlighted that became stronger with increasing depth. However, the part of explained variation was substantially less than values reported for studies conducted in single fields or micro-catchments. We concluded that important explaining factors were missing in the predictor variables or that meaningful relationships were partially masked by short range variability acting as noise. This raised the need to study the problem separately over multiple spatial scales in order to identify scales at which control on SOC was well addressed and, on the other hand, scales at which important explaining factors were possibly neglected.

The objective of Chapter 3 was therefore to separate scales of variation of SOC and controlling factors in order to characterize the scale dependency of their relationships. Eventually, three specific ranges of spatial dependence were identified in SOC variation (referred to as short, middle and long range and corresponding respectively to 720 m, 5.5 km and 28 km). We observed that the part of SOC variation characterized by a long range was relatively larger for the the top layer than for deeper layers of soil. Besides, the fine textural fraction varied relatively more at long range, compared to SOC. Correlation analyses by scale component showed that the correlation between SOC and the fine textural fraction appeared only between their long range component of variation. Besides, the correlation between topsoil and subsoil layers for a given factor was also stronger as the scale increased, for

both SOC and textural fractions. These results indicated that the influence of texture was mainly occurring at long range and on topsoil while other processes had a large influence at short range and particularly on SOC located below the topsoil. This short range corresponds approximately to the size of a field and could be related to factors such as erosion processes and land management. Further research was therefore needed to estimate the relative effect of these factors.

The objective of Chapter 4 was to identify the missing factor(s) preventing us to predict SOC with high precision at fine spatial resolution in Chapter 2, and likely to occur at small scale (cf Chapter 3). In particular, we wanted to quantify the relative part of variation at the landscape scale related to the pattern of fields in order to check the possibility that the absence of consideration of factors related to fields explained the low coefficients of determination in Chapter 2. It should be pointed out that this study was conducted in Luxembourg due to the availability of remote sensing data and that the hierarchy of factors may be partly different than in the loam belt. The variation related to field boundaries was not significant everywhere but could locally explain up to 68.8 % of the spatial variation in the landscape. This confirmed that land management and/or historical land use change played a major control on SOC at the landscape scale.

If Chapters 2 to 4 contributed to better explaining the variability of SOC over spatial scale, Chapter 5 proposed to put these results into perspective and ask if predicting the variability over scales and with all contributing factors was always needed. A theoretical analysis of total stock estimation indicated that, in function of the spatial dependence and the extent of the area, the added value of predicting the spatial variation could be negligible.

6.1.3 Analyze the potential, limitations and pitfalls of the different formats of data used in this study.

In this section we review the main format of data that were used and the most instructive conclusions.

Legacy and recent datasets of soil profiles

Legacy soil data have been produced over 70 years in many countries, generally in the form of choropleth maps and soil profile descriptions (Arrouays *et al.*, 2017; Odgers *et al.*, 2012). Currently, such data are being rescued through digitization and are exploited within the framework of DSM, typically for the creation of digital soil maps following GlobalSoilMap specifications (Arrouays *et al.*, 2017). Although legacy data can be a valuable source of information in DSM, they are also associated with different types of problems and limitations (Mayr *et al.*, 2008).

We used –a subset of– the Aardewerk legacy dataset in Chapters 2 and 3. It contains soil profiles described during the Belgian National Soil Survey (1947–1962). Soil profile location, horizon limits, physical and chemical properties and classification were recorded on paper. In 1993, these records were digitized, creating the digital soil database Aardewerk. The complete dataset Aardewerk contains 8962 profiles, north of the river axis Sambre and Meuse, representing 48 311 horizons. Due to this large number of points, Aardewerk provides precious data for the study of SOC processes or the estimation of change in total SOC stock. Multiple limitations and pitfalls were however highlighted.

Information gaps It is likely that some horizons were lost during the digitization. Because of that, we had to discard around 18% of the profiles

Loss of precision Because of storage constraints, data description was simplified and standardized and, as a result, part of the information was lost. A critical simplification was the coding of the horizon boundaries as intervals instead of exact values. A method was developed in Chapter 2 to retrieve the limits of the horizons. It exploits simultaneously the information on all horizons of the profile. However, even with this method

some uncertainty remained on the position of the horizon and thus on the SOC vertical concentration, in particular at the lower boundary of the plough layer.

Method of analysis The horizon samples were analyzed with the method of Walkley and Black. This imposes the use of a correction factor which may induce decrease in precision or even a systematic error (see Section 1.1.4)

Error of geolocation The correlation between SOC and TA put in evidence a lack of precision and/or accuracy in the location of the profiles (Chapitre 2). This was confirmed by a sensitivity analysis using the resampled dataset (that has precise coordinates). The possible sources of error are positioning errors on paper maps during the original survey and transformation between different geoid.

Inconsistencies within data The methodology applied in Chapter 4 highlighted the presence of a probable bias in the SOC observed for a group of profiles sampled by a specific surveyor.

Here we underline the importance of a partial resampling to compare the results, like the R_AW dataset that we exploited in Chapter 2. This can help assessing the importance of error in geolocation or the precision and accuracy of sampled properties and correcting anomalies.

Airborne imagery

Airborne and satellite spectral imaging has the potential to cover large extents with fair spatial resolution. Besides, the absence of spatial interpolation eliminates a common source of spatial dependence in the error and allows the detection of small-scale spatial variability of SOC content. In Chapter 4, an image acquired by airborne hyperspectral Vis-NIR reflectance spectroscopy was exploited to quantify variation of SOC in cropland. The fine resolution (5 m) and the spatial exhaustivity of the information over multiple neighboring field allowed us to separate variation related to the limits of fields with random Gaussian variation at roughly the same scale. Airborne imagery can therefore help in the identification, quantification and modeling of SOC processes that plays a role at the field and landscape scales.

6.1.4 Determine how the prediction of spatial variation is useful for the estimation of the total SOC stock within a study area.

In Chapter 5, we focused on the generic problem which consists in estimating at best the total stock of SOC within an area of interest. This stock was estimated by integration over an area of interest of the prediction provided in any spatial location, by a model built on auxiliary variables as well as the prediction of model errors based on available SOC sampling points. This allowed us to highlight the link between the prediction variance of the total stock, the size of the study area and the range of spatial dependence of model error. We concluded that practically, there is no benefit in accounting for the errors instead of just neglecting them when the range of spatial dependence of these errors is much lower than the size of the domain A. However, When this range is much higher than the size of the domain A, these errors start to behave like a single random constant so that a crude arithmetic average obtained from a very limited sample size will do the job. Between these limit cases, the best estimate of total stock is obtained by spatially predicting SOC with the best linear unbiased predictor (i.e. by kriging model errors).

A fundamental result is that it is not necessary to include all the controlling factors in the model. Factors influencing SOC at a spatial scale largely inferior to the the size of the study area can be neglected without significant impact. Similarly, collecting soil samples and using them to spatially predict model error is not necessary when the model accounts for all sources of SOC variability (at a spatial scale that is not largely inferior to the the size of the study area).

6.2 Perspectives

Shift agricultural practices for climate change mitigation

Management of SOC is seen as a means to mitigate climate change. However, the implementation of soil-based greenhouse gas mitigation activities are at an early stage and accurately quantifying emissions and reductions remains a substantial challenge (Paustian *et al.*, 2016). Here we describe what is needed, at our opinion, to support the implementation and increase the efficiency of climate change mitigating practices in cropland.

On the one hand, we believe that it is essential to assess the effect of management practices in the long run and in realistic environmental conditions. In this perspective, it should be focused on the interactions of practices with climate change and with landscape scale processes. For example, Nadeu *et al.* (2015) used a process-based model at the regional extent and at 20 m resolution that accounted for erosion and land management. This allowed them to conclude that practices targeting C sequestration would be most effective when accompanied by measures that reduce soil erosion given that erosion loss can balance potential C uptake, particularly in sloping areas. Besides, if land management generally has a direct impact on topsoil, when erosion occurs the effect of practices can also influence subsoil SOC stock. Therefore, efforts should continue in order to reduce the uncertainty on the long-term fate on this buried SOC (Trumbore and Czimczik, 2008).

On the other hand, DSM of SOC and others soil properties could help targeting management practices. Spatial prediction of variables such as the texture, clay type, level of aggregation or C/N ratio could allow estimating the level of saturation of the different functional pools of SOC sequestration. This would help identifying the fields that are the most depleted in SOC due to past intensive management and where the potential for sequestration is the highest. From this, information about the potential sequestration at short or long term, in the form of potential saturation curve, achievable curve and critical curve could be derived (Stockmann *et al.*, 2013). The application presented in Chapter 4, which filters SOC variation associated to the fields, could already help identifying the fields with the highest sequestration potential, as long as a high resolution SOC image is available.

Predict soil properties at global scale

There is a high demand for spatial information on soil properties at the global scale. Such data are essential to spatially predict SOC stocks. Such spatial information is also useful for the prediction of total stock by design-based approach as they can be used for stratification. For a given threshold of detectable change, appropriate stratification can indeed significantly reduce the number of required samples compared to simple random sampling (Heuvelink *et al.*, 2017).

In the frame of the development of the harmonized global soil database GlobalSoilMap (GSM, Arrouays *et al.* (2014b)), the opportunity and suitability of incorporating legacy datasets is widely discussed (e.g. Legrain *et al.* (2014)). In the SoilGrids250m project, a large number of datasets from various sources and formats, including legacy datasets of soil profiles, are exploited by machine learning algorithms to predict multiple soil properties at global scale, at a resolution of 250 m and at different soil depths (Hengl *et al.*, 2017). An important challenge is the harmonization of these multiple dataset of soil profiles having different codifications and definitions. The issues faced with Aardewerk in our analysis could potentially be met with any legacy dataset of soil profiles. Based on our findings, different strategies could be applied to assess their general quality, detect potential errors, correct them and reconstruct missing information. Filtered kriging as a method to detect anomalies inside a dataset could be beneficial in solving current issues. Besides, due to the differences of instruments, methods of analysis and procedures, systematic differences between datasets may exist. Our method allows to highlight these differences through the selection of a specific spatial scale at which they appear more obviously. Finally, we believe that the method applied to reconstruct the vertical distribution of horizons in Chapter 2 provides potentially useful insight for equivalent problems of uncertain information in soil profiles.

Appendix A

Multi-scale decomposition and filtering of spatial variables

The central idea of filtering is to separate the spatial variation in multiple orthogonal scale components. For this purpose, we make the assumption that the variation of a soil variable is the sum of a spatially dependent trend along with independent stochastic processes acting at different spatial scales.

$$z(\mathbf{x}) = \mu_Z(\mathbf{x}) + \boldsymbol{\epsilon}(\mathbf{x}) = \mu_Z(\mathbf{x}) + \boldsymbol{\epsilon}_1(\mathbf{x}) + \ldots + \boldsymbol{\epsilon}_m(\mathbf{x}) \tag{A.1}$$

The spatially dependent trend is equal to the expectation value of $Z(\mathbf{x})$

$$\mu_Z(\mathbf{x}) = E[Z(\mathbf{x})] \tag{A.2}$$

Thanks to the linear property of the variance, the semi-variogram model of $z(\mathbf{x})$ in eq. (A.1) is the sum of n functions characterized by different range and sills. Each function represents the part of the total variation at the corresponding spatial scale with

$$\gamma(||\mathbf{h}||) = \gamma_1(||\mathbf{h}||) + \ldots + \gamma_m(||\mathbf{h}||) \tag{A.3}$$

Assuming further a second-order stationary processes, the covariance function for a distance interval $||\mathbf{h}||$ is directly related to the semi-variogram with

$$C(||\mathbf{h}||) = C(0) - \gamma(||\mathbf{h}||) \tag{A.4}$$

and

$$C_i(||\mathbf{h}||) = C_i(0) - \gamma_i(||\mathbf{h}||) \tag{A.5}$$

Therefore, the covariance of the $\boldsymbol{\epsilon}_i$ sums up with

$$C(||\mathbf{h}||) = C_1(||\mathbf{h}||) + \ldots + C_m(||\mathbf{h}||) \tag{A.6}$$

Eq. (A.1) decomposes Z in a spatial trend $\mu_Z(\mathbf{x})$ and a random component $\boldsymbol{\epsilon}(\mathbf{x})$ that can itself be further decomposed in multiple scale components. Assuming a linear model, the spatial trend is defined as

$$\mu_Z(\mathbf{x}) = \begin{pmatrix} 1 & f_1(\mathbf{x}) & \cdots & f_k(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \mathbf{f}(\mathbf{x}) \tag{A.7}$$

where the $f_j(\mathbf{x}), j = 1, \ldots, k$ are linear or non-linear functionals of the spatial coordinates and $\boldsymbol{\beta}$ is a vector of parameters that can be estimated using generalized least squares regression with

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{z} \tag{A.8}$$

where $\mathbf{X}$ is a $p \times (k+1)$ matrix whose ith row is $\mathbf{f}(\mathbf{x}_i)$ and $\boldsymbol{\Sigma}$ is the matrix of covariance of $\mathbf{Z} = [Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_p)]$.

The random component of Eq. (A.1) is predicted at location $\mathbf{x}_0$ directly or separately component by component with

$$\hat{\boldsymbol{\epsilon}}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad ; \qquad \hat{\boldsymbol{\epsilon}}_i(\mathbf{x}_0) = \boldsymbol{\lambda}_i^\top (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{A.9}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_i$ are vectors of simple kriging weights. These are given by

$$\boldsymbol{\lambda}^\top = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma})^\top \qquad ; \qquad \boldsymbol{\lambda}_i^\top = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_i)^\top \tag{A.10}$$

where $\boldsymbol{\sigma}$ is the vector of covariances between $\boldsymbol{\epsilon}(\mathbf{x}_0)$ and $[\boldsymbol{\epsilon}(\mathbf{x}_1), \ldots, \boldsymbol{\epsilon}(\mathbf{x}_p)]$ and $\boldsymbol{\sigma}_i$ is the vector of covariances between $\boldsymbol{\epsilon}_i(\mathbf{x}_0)$ and $[\boldsymbol{\epsilon}_i(\mathbf{x}_1), \ldots, \boldsymbol{\epsilon}_i(\mathbf{x}_p)]$. Since we assume that the covariances depend only on distances, $\boldsymbol{\Sigma}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}_i$ are estimated from the covariances functions (Eq. (A.6)). Besides, as the filtering is linear, $\boldsymbol{\sigma}_i$ sum up to $\boldsymbol{\sigma}$ and the scale components of the residue sum up to the residue according to

$$\sum_i \hat{\boldsymbol{\epsilon}}_i(\mathbf{x}_0) = [\boldsymbol{\Sigma}^{-1}(\sum_i \boldsymbol{\sigma}_i)]^\top (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma})^\top (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\epsilon}}(\mathbf{x}_0) \tag{A.11}$$

Appendix B

Method REML-EBLUP

Let us consider the linear spatial model

$$\mathbf{z}(\mathbf{x}) = \mathbf{M}(\mathbf{x})\boldsymbol{\beta} + \boldsymbol{\epsilon}(\mathbf{x}) \tag{B.1}$$

where $\mathbf{x}$ is the vector of spatial coordinates, $\mathbf{z}$ is the $n \times 1$ dependent variable, $\mathbf{M}$ is the $n \times p$ design matrix for the independent variables (fixed effects) with $\boldsymbol{\beta}$ the vector of parameters and $\boldsymbol{\epsilon}$ a zero-mean normally distributed and spatially correlated random effect. If $\mathbf{z}$ is a vector of values at sampled locations, the best linear unbiased predictor of $\mathbf{z}(\mathbf{x})$ at unsampled location $\mathbf{x}_0$ is given by

$$z(\mathbf{x}_0) = \mu_Z(\mathbf{x})(\mathbf{x}_0)^\top \hat{\boldsymbol{\beta}} + \mathbf{k}^\top \mathbf{K}^{-1}(\mathbf{z} - \mathbf{M}\boldsymbol{\beta}) \tag{B.2}$$

where $\mathbf{K} = \text{cov}(\boldsymbol{\epsilon}(\mathbf{x}), \boldsymbol{\epsilon}^\top(\mathbf{x}))$ and $\mathbf{k} = \text{cov}(\boldsymbol{\epsilon}(\mathbf{x}), \boldsymbol{\epsilon}(\mathbf{x}_0))$. $\mathbf{K}$ and $\mathbf{k}$ are not known but are modeled by a covariance function with parameter vector $\boldsymbol{\theta}$. Assuming that $\boldsymbol{\epsilon}$ is normally distributed, an unbiased estimation of $\boldsymbol{\theta}$ is obtained by the REML method (Kitanidis and Vomvoris, 1983). This method implies transforming $\mathbf{z}$ into

$$\mathbf{y} = \mathbf{T}\mathbf{z} = (\boldsymbol{I} - \mathbf{M}(\mathbf{M}^\top\mathbf{M})^{-1}\mathbf{M}^\top)\mathbf{z} = \mathbf{z} - \mathbf{M}\hat{\boldsymbol{\beta}}' \tag{B.3}$$

where $\hat{\boldsymbol{\beta}}'$ is the ordinary least squares estimator of $\boldsymbol{\beta}$. $\hat{\boldsymbol{\theta}}$ is then obtained by maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\log|\boldsymbol{W}| - \frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{Q}y \tag{B.4}$$

where $\boldsymbol{W} = \mathbf{M}^\top\mathbf{K}^{-1}\mathbf{M}$ and $\mathbf{Q} = \boldsymbol{I} - \mathbf{M}\boldsymbol{W}^{-1}\mathbf{M}^\top\mathbf{K}^{-1}$. Finally, using the generalized least squares method, an estimate for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{W}^{-1}\mathbf{M}^\top\mathbf{K}^{-1}\mathbf{z} \tag{B.5}$$

and an estimate $\hat{\mathbf{z}}(\mathbf{x}_0)$ is computed from Equation (B.1) and from the estimates $\hat{\boldsymbol{\beta}}$. The residuals are then obtained from

$$\hat{\boldsymbol{\epsilon}} = \mathbf{z} - \mathbf{M}\hat{\boldsymbol{\beta}} \tag{B.6}$$

The covariance of the parameters $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(M^\top K^{-1} M\right)^{-1} \tag{B.7}$$

Bibliography

Abdollahi, L., Schjønning, P., Elmholt, S., and Munkholm, L. J. 2014. The effects of organic matter application and intensive tillage and traffic on soil structure formation and stability. Soil and Tillage Research, 136:28–37.

AGIV. 2006. DHM-Vlaanderen, raster, 5 m.

Alexander, P., Paustian, K., Smith, P., and Moran, D. 2015. The economics of soil C sequestration and agricultural emissions abatement. SOIL, 1(1):331–339.

Ancev, T. 2011. Policy Considerations for Mandating Agriculture in a Greenhouse Gas Emissions Trading Scheme. Applied Economic Perspectives and Policy, 33(1):99–115.

Angers, D. A., Arrouays, D., Saby, N. P. A., and Walter, C. 2011. Estimating and mapping the carbon saturation deficit of French agricultural topsoils. Soil Use and Management, 27(4):448–452.

Arrouays, D., Deslais, W., and Badeau, V. 2001. The carbon content of topsoil and its geographical distribution in France. Soil use and Management, 17(1):7–11.

Arrouays, D., Marchant, B. P., Saby, N. P. A., *et al.*. 2012. Generic Issues on Broad-Scale Soil Monitoring Schemes: A Review. Pedosphere, 22(4):456–469.

Arrouays, D., Grundy, M. G. M. M. G., Hartemink, A. A. E., *et al.*. 2014a. Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. volume 125 of *Advances in Agronomy*, page 93. Academic Press.

Arrouays, D., McKenzie, N., Hempel, J., Forges, A. R. d., and McBratney, A. B. 2014b. *GlobalSoilMap: Basis of the global spatial soil information system*. CRC Press.

Arrouays, D., Leenaars, J. G. B., Richer-de Forges, A. C., *et al.*. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ, 14:1–19.

Aynekulu, B. E., Vagen, T., Shephard, K., and Winowiecki, L. 2011. A Protocol for Modeling, Measurement and Monitoring Soil Carbon Stocks in Agricultural Landscapes. Version 1.1. Technical report, World Agroforestry Centre (ICRAF).

Ayoubi, S. and Karchegani, P. M. 2012. Determination the factors explaining variability of physical soil organic carbon fractions using Artificial Neural Network. Int. J. Soil Sci., 7(1):1–14.

Barraclough, D., Smith, P., Worrall, F., Black, H. I. J., and Bhogal, A. 2015. Is there an impact of climate change on soil carbon contents in England and Wales? European Journal of Soil Science, 66(3):451–462.

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

Batjes, N. H. 2014. Batjes, N. H. 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science, 47, 151–163. Reflections by N.H. Batjes. European Journal of Soil Science, 65(1):2–3.

Baveye, P. C. and Laba, M. 2015. Moving away from the geostatistical lamppost: Why, where, and how does the spatial heterogeneity of soils matter? Ecological Modelling, 298:24–38.

Bellamy, P. H., Loveland, P. J., Bradley, R. I., Lark, R. M., and Kirk, G. J. D. 2005. Carbon losses from all soils across England and Wales 1978–2003. Nature, 437(7056):245–248.

Berhe, A. A. and Kleber, M. 2013. Erosion, deposition, and the persistence of soil organic matter: mechanistic considerations and problems with terminology. Earth Surface Processes and Landforms, 38(8):908–912.

Berhe, A. A., Harden, J. W., Torn, M. S., and Harte, J. 2008. Linking soil organic matter dynamics and erosion-induced terrestrial carbon sequestration at different landform positions. Journal of Geophysical Research: Biogeosciences, 113(4):4039.

Beven, K. J. and Kirkby, M. J. 1979. A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrological Sciences Bulletin, 24(1):43–69.

Bilgili, A. V., Akbas, F., and Es, H. M. 2010. Combined use of hyperspectral VNIR reflectance spectroscopy and kriging to predict soil variables spatially. Precision Agriculture, 12(3):395–420.

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

Biswas, A. and Si, B. C. 2011. Revealing the Controls of Soil Water Storage at Different Scales in a Hummocky Landscape. Soil Science Society of America Journal, 75(4):1295.

Biswas, A., Cresswell, H. P., Chau, H. W., Viscarra Rossel, R. A., and Si, B. C. 2013. Separating scale-specific soil spatial variability: A comparison of multi-resolution analysis and empirical mode decomposition. Geoderma, 209-210:57–64.

Boucneau, G., Van Meirvenne, M., and Hofman, G. 1998. Comparing pedotransfer functions to estimate soil bulk density in northern Belgium. Pedologie - Themata, 5:67–70.

Bourennane, H., Nicoullaud, B., Couturier, A., *et al.*. 2012. Geostatistical filtering for improved soil water content estimation from electrical resistivity data. Geoderma, 183–184:32–40.

Brus, D. and De Gruijter, J. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma, 80(1-2):1–44.

Burnham, K. P. and Anderson, D. R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media.

Butz, T. 2006. *Fourier transformation for pedestrians*. Springer, Berlin; New York.

Candela, J. P. and Carlson, D. 2017. The Annual Global Carbon Budget. World Meteorological Organization Bulletin., 66(1).

Cardinael, R., Chevallier, T., Barthès, B. G., *et al.*. 2015. Impact of alley cropping agroforestry on stocks, forms and spatial distribution of soil organic carbon — A case study in a Mediterranean context. Geoderma, 259-260(Supplement C):288–299.

Casa, R., Castaldi, F., Pascucci, S., and Pignatti, S. 2012. Potential of hyperspectral remote sensing for field scale soil mapping and precision agriculture applications. Italian Journal of Agronomy, 7(4):e43.

Casa, R., Castaldi, F., Pascucci, S., Basso, B., and Pignatti, S. 2013. Geophysical and hyperspectral data fusion techniques for in-field estimation of soil properties. Vadose Zone Journal, 12(4).

Chai, X., Shen, C., Yuan, X., and Huang, Y. 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma, 148(2):159–166.

Chambers, A., Lal, R., and Paustian, K. 2016. Soil carbon sequestration potential of US croplands and grasslands: Implementing the 4 per Thousand Initiative. Journal of Soil and Water Conservation, 71(3):68A–74A.

Chaopricha, N. T. and Marín-Spiotta, E. 2014. Soil burial contributes to deep soil organic carbon storage. Soil Biology and Biochemistry, 69:251–264.

Chartin, C., Evrard, O., Salvador-Blanes, S., *et al.*. 2013. Quantifying and modelling the impact of land consolidation and field borders on soil redistribution in agricultural landscapes (1954–2009). CATENA, 110:184–195.

Chartin, C., Stevens, A., Goidts, E., *et al.*. 2017. Mapping Soil Organic Carbon stocks and estimating uncertainties at the regional scale following a legacy sampling strategy (Southern Belgium, Wallonia). Geoderma Regional, 9:73–86.

Chilès, J. and Delfiner, P. 1999. *Statistics for spatial data*. Wiley, New York.

Ciais, P., Borges, A. V., Abril, G., *et al.*. 2008. The impact of lateral carbon fluxes on the European carbon balance. Biogeosciences, 5(5):1259–1271.

Coleman, K. and Jenkinson, D. S. 1999. ROTHC-26.3. A model for the turnover of carbon in soils. Herts, Rothamsted Research, Harpenden, Hertfordshire, UK.

Corstanje, R., Schulin, R., and Lark, R. M. 2007. Scale-dependent relationships between soil organic carbon and urease activity. European Journal of Soil Science, 58(5):1087–1095.

Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Denef, K., and Paul, E. 2013. The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: do labile plant inputs form stable soil organic matter? Global Change Biology, 19(4):988–995.

Davidson, E. A. E. A., Janssens, I. A., Marks, D., *et al.*. 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature, 440(7081):165–73.

de Brogniez, D., Ballabio, C., Stevens, A., *et al.*. 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. European Journal of Soil Science, 66(1):121–134.

de Gruijter, J. J., McBratney, A. B., Minasny, B., *et al.*. 2016. Farm-scale soil carbon auditing. Geoderma.

De Gryze, S., Six, J., Bossuyt, H., Van Oost, K., and Merckx, R. 2008. The relationship between landform and the distribution of soil C, N and P under conventional and minimum tillage. Geoderma, 144(1-2):180–188.

Demarcin, P., Degre, A., Smoos, A., and Dautrebande, S. 2009. Cartographie numérique des zones à risque de ruissellement et d'érosion des sols en Région wallonne. Technical report, Service Public de Wallonie.

Department of Sustainable Natural Resources. Unknown year. Soil survey standard test method organic carbon. Technical report, New South Wales.

Dexter, A. R. 2003. Soil physical quality Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth. Geoderma, 120:201–214.

Dlugoß, V., Fiener, P., and Schneider, K. 2010. Layer-Specific Analysis and Spatial Prediction of Soil Organic Carbon Using Terrain Attributes and Erosion Modeling. Soil Science Society of America Journal, 74(3):922.

Dlugoß, V., Fiener, P., Van Oost, K., and Schneider, K. 2012. Model based analysis of lateral and vertical soil carbon fluxes induced by soil redistribution processes in a small agricultural catchment. Earth Surface Processes and Landforms, 37(2):193–208.

Doetterl, S., Six, J., Van Wesemael, B., and Van Oost, K. 2012. Carbon cycling in eroding landscapes: Geomorphic controls on soil organic C pool composition and C stabilization. Global Change Biol., 18(7):2218–2232.

Doetterl, S., Stevens, A., van Oost, K., Quine, T. A., and van Wesemael, B. 2013a. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma, 204-205:31–42.

Doetterl, S., Stevens, A., Van Oost, K., and Van Wesemael, B. 2013b. Soil organic carbon assessment at high vertical resolution using closed-tube sampling and Vis-NIR spectroscopy. Soil Science Society of America Journal, 77(4):1430–1435.

Dungan, J. L., Perry, J. N., T Dale, M. R., *et al.*. 2002. A balanced view of scale in spatial statistical analysis. Ecography, 25:242–244.

Easter, M., Paustian, K., Killian, K., *et al.*. 2007. The GEFSOC soil carbon modelling system: A tool for conducting regional-scale soil carbon inventories and assessing the impacts of land use change on soil carbon. Agric. Ecosyst. Environ., 122(1):13–25.

Ellert, B. and Bettany, J. 1995. Calculation of organic matter and nutrients stored in soils under contrasting management regimes. Canadian Journal of Soil Science, 75(4):529–538.

Falloon, P. and Smith, P. 2002. Simulating SOC changes in long-term experiments with RothC and CENTURY: model evaluation for a regional scale application. Soil Use and Management, 18(2):101–111.

Falloon, P., Jones, C. D., Ades, M., and Paul, K. 2011. Direct soil moisture controls of future global soil carbon changes: An important source of uncertainty. Global Biogeochem. Cycles, 25(3):1–14.

FAO. 2017. Unlocking the potential of soil organic carbon. Technical report, Food and Agriculture Organization.

Fiener, P., Auerswald, K., and Van Oost, K. 2011. Spatio-temporal patterns in land use and management affecting surface runoff response of agricultural catchments—A review. Earth-Science Reviews, 106(1-2):92–104.

Florence, R. and Mengel, D. 2013. Evaluation of Walkley-Black, Loss-on-Ignition, and Dry Combustion for Estimating Soil Organic Matter in Kansas. Technical report, Kansas State Department of Agronomy Soil Testing Lab, Kansas.

Florinsky, I. 2016. *Digital terrain analysis in soil science and geology*. Academic Press.

Florinsky, I. V., Eilers, R. G., Manning, G. R., and Fuller, L. G. 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software, 17(3):295–311.

Fontaine, S., Barot, S., Barré, P., *et al.*. 2007. Stability of organic carbon in deep soil layers controlled by fresh carbon supply. Nature, 450(7167):277–280.

Foster, D., Swanson, F., Aber, J., *et al.*. 2003. The Importance of Land-Use Legacies to Ecology and Conservation. BioScience, 53(1):77.

Galli, A., Gerdil-Neuillet, F., and Dadou, C. 1984. Factorial Kriging Analysis: A Substitute to Spectral Analysis of Magnetic Data. In Verly, G., David, M., Journel, A. G., and Marechal, A., editors, *Geostatistics for Natural Resources Characterization*, pages 543–557. Springer Netherlands.

Gessler, P. E., Moore, I. D., McKENZIE, N. J., and Ryan, P. J. 1995. Soil-landscape modelling and spatial prediction of soil attributes. International journal of geographical information systems, 9(4):421–432.

Ghariani, K., Chehata, N., Le Bris, A., and Lagacherie, P. 2014. Agricultural field delimitation using active learning and random forests margin. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 1717–1720.

Goidts, E. 2009. *Soil Organic carbon evolution at the regional scale : overcoming uncertainties & quantifying driving forces*. Ph.D. thesis, Université catholique de {L}ouvain.

Goidts, E. and van Wesemael, B. 2007. Regional assessment of soil organic carbon changes under agriculture in Southern Belgium (1955–2005). Geoderma, 141(3-4):341–354.

Goidts, E., van Wesemael, B., and Crucifix, M. 2009a. Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. European Journal of Soil Science, 60(5):723–739.

Goidts, E., Wesemael, B. V., and Van Oost, K. 2009b. Driving forces of soil organic carbon evolution at the landscape and regional scale using data from a stratified soil monitoring. Global Change Biology, 15(12):2981–3000.

Gomez, C., Viscarra Rossel, R. A., and McBratney, A. B. 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma, 146(3-4):403–411.

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

Gomez, C., Drost, A. P. A., and Roger, J. M. 2015. Analysis of the uncertainties affecting predictions of clay contents from VNIR/SWIR hyperspectral data. Remote Sensing of Environment, 156:58–70.

Goossens, D. 1997. Long-term aeolian loess accumulation modelled in the wind tunnel: the Molenberg case (central loess belt, Belgium). Zeitschrift fur geomorphologie, 41(1):115–129.

Goovaerts, P. 1992. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. Journal of Soil Science, 43(4):597–619.

Goovaerts, P. 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma, 89(1-2):1–45.

Goovaerts, P. and Webster, R. 1994. Scale dependent correlation between topsoil copper and cobalt concentrations in Scotland. European Journal of Soil Science, 45(1):79–95.

Goovaerts, P., Sonnet, P., and Navarre, A. 1993. Factorial kriging analysis of springwater contents in the Dyle River Basin, Belgium. Water Resources Research, 29(7):2115–2125.

Govers, G., Quine, T. A., Desmet, P. J. J., and Walling, D. E. 1996. The Relative Contribution of Soil Tillage and Overland Flow Erosion to Soil Redistribution on Agricultural Land. Earth Surface Processes and Landforms, 21(10):929–946.

Grace, J. 2004. Understanding and managing the global carbon cycle. Journal of Ecology, 92(2):189–202.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma, 146(1-2):102–113.

Haase, D., Fink, J., Haase, G., *et al.*. 2007. Loess in Europe—its spatial distribution based on a European Loess Map, scale 1:2,500,000. Quaternary Science Reviews, 26(9-10):1301–1312.

Harden, J. W., Sharpe, J. M., Parton, W. J., *et al.*. 1999. Dynamic replacement and loss of soil carbon on eroding cropland. Glob. Biogeochem. Cycles, 13(4):885–901.

Harden, J. W., Berhe, A. A., Torn, M., *et al.*. 2008. Soil erosion: Data say C sink. Science, 320(5873):178–179.

Harrison, R. B., Footen, P. W., and Strahm, B. D. 2011. Deep soil horizons: Contribution and importance to soil carbon pools and in assessing whole-ecosystem response to management and global change. Forest Science, 57(1):67–76.

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.

Hastie, T. J. 2017. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Hbirkou, C., Pätzold, S., Mahlein, A.-K., and Welp, G. 2012a. Airborne hyperspectral imaging of spatial soil organic carbon heterogeneity at the field-scale. Geoderma, 175-176:21–28.

Hbirkou, C., Pätzold, S., Mahlein, A.-K., and Welp, G. 2012b. Airborne hyperspectral imaging of spatial soil organic carbon heterogeneity at the field-scale. Geoderma, 175-176:21–28.

Hengl, T., Mendes De Jesus, J., Macmillan, R. A., *et al.*. 2014. SoilGrids1km-Global Soil Information Based on Automated Mapping. PLOS One, 9(8):1–17.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., *et al.*. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE.

Herbst, M., Hellebrand, H. J., Bauer, J., *et al.*. 2008. Multiyear heterotrophic soil respiration: Evaluation of a coupled CO2 transport and carbon turnover model. Ecological Modelling, 214(2-4):271–283.

Heuvelink, G., Batjes, N., Brus, D., *et al.*. 2017. Towards a sampling design for monitoring global soil organic carbon stocks. In *Book of Abstracts Wageningen Soil Conference 2017*, pages 74–74.

Heuvelink, G. B. M. 1998. Uncertainty analysis in environmental modelling under a change of spatial scale. Nutrient Cycling in Agroecosystems, 50(1-3):255–264.

Hillel, D. and Hatfield, J. L. 2005. *Encyclopedia of Soils in the Environment*, volume 3. Elsevier Amsterdam.

Hoffmann, T., Schlummer, M., Notebaert, B., Verstraeten, G., and Korup, O. 2013. Carbon burial in soil sediments from Holocene agricultural erosion, Central Europe. Global Biogeochemical Cycles, 27(3):828–835.

IFOAM, E. G. 2017. IFOAM EU position paper on the Effort Sharing Regulation and Land Use, Land Use Change and Forestry (LULUCF) regulation. Technical report, IFOAM.

Janzen, H. H., Campbell, C. A., Izaurralde, R. C., *et al.*. 1998. Management effects on soil C storage on the Canadian prairies. Soil&TillageResearch, 47:181–195.

Jenny, H. 1994. *Factors of soil formation: a system of quantitative pedology*. Courier Corporation.

Johannes, A., Matter, A., Schulin, R., *et al.*. 2017. Optimal organic carbon values for soil structure quality of arable soils. Does clay content matter? Geoderma, 302:14–21.

Kempen, B., Heuvelink, G. B. M., Brus, D. J., and Stoorvogel, J. J. 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. European Journal of Soil Science, 61(3):333–347.

Kempen, B., Brus, D. J., and Stoorvogel, J. J. 2011. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. Geoderma, 162(1-2):107–123.

Kimble, J. M., Stewart, B. A., and Follett, R. F. 2000. *Assessment methods for soil carbon*. CRC press.

Kirk, G. J. D. and Bellamy, P. H. 2010. Analysis of changes in organic carbon in mineral soils across England and Wales using a simple single-pool model. European journal of soil science, 61(3):406–411.

Kirkels, F. M. S. A., Cammeraat, L. H., and Kuhn, N. J. 2014. The fate of soil organic carbon upon erosion, transport and deposition in agricultural landscapes — A review of different concepts. Geomorphology, 226:94–105.

Kitanidis, P. and Vomvoris, E. 1983. A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. Water Resources Research, 19(3):677–690.

Klotzbücher, T., Kaiser, K., Guggenberger, G., Gatzek, C., and Kalbitz, K. 2011. A new conceptual model for the fate of lignin in decomposing plant litter. Ecology, 92(5):1052–1062.

Köchy, M., Hiederer, R., and Freibauer, a. 2015. Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. SOIL, 1(1):351–365.

Lacoste, M., Minasny, B., McBratney, A., *et al.*. 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma, 213:296–311.

Lacoste, M., Viaud, V., Michot, D., and Walter, C. 2015. Landscape-scale modelling of erosion processes and soil carbon dynamics under land-use and climate change in agroecosystems. European Journal of Soil Science, 66(4):780–791.

Lacoste, M., Michot, D., Viaud, V., *et al.*. 2016. High resolution 3D mapping for soil organic carbon assessment in a rural landscape. Computing Ethics: A Multicultural Approach, 341:341–345.

Ladoni, M., Bahrami, H. A., Alavipanah, S. K., and Norouzi, A. A. 2009. Estimating soil organic carbon from soil reflectance: a review. Precision Agriculture, 11(1):82–99.

Lagacherie, P. and Gomez, C. 2014. What can GlobalSoilMap expect from Vis-NIR HyperSpectral imagery in the near future. GlobalSoilMap: Basis of the global spatial soil information system, pages 387–392.

Lagacherie, P. and McBratney, A. B. 2006. Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. Developments in soil science, 31:3–22.

Lagacherie, P., Sneep, A.-R., Gomez, C., *et al.*. 2013. Combining Vis–NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). Geoderma, 209-210:168–176.

Lal, R. 2004. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science, 304(5677):1623–1627.

Lal, R. 2016. Beyond COP 21: potential and challenges of the "4 per Thousand" initiative. Journal of Soil and Water Conservation, 71(1):20A–25A.

Lal, R. and Pimentel, D. 2008. Soil Erosion: A Carbon Sink or Source? Science, 319(5866):1040–1042.

Lark, R. M., Cullis, B. R., and Welham, S. J. 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science, 57(6):787–799.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., *et al.*. 2017. Global carbon budget 2017. Earth System Science Data Discussions, pages 1–79.

Legrain, X., Bock, L., and Colinet, G. 2014. Suitability of the soil map and legacy data in Wallonia (BE) to support the GSM initiative. In McBratney, A., editor, *GlobalSoilMap*, pages 99–102. CRC Press.

Lehmann, J. and Kleber, M. 2015. The contentious nature of soil organic matter. Nature, 528(7580):60.

Lettens, S., Van Orshoven, J., van Wesemael, B., and Muys, B. 2004. Soil organic and inorganic carbon contents of landscape units in Belgium derived using data from 1950 to 1970. Soil Use and Management, 20(1):40–47.

Lettens, S., De Vos, B., Quataert, P., *et al.*. 2007. Variable carbon recovery of Walkley-Black analysis and implications for national soil organic carbon accounting. European Journal of Soil Science, 58(6):1244–1253.

Lin, H., Wheeler, D., Bell, J., and Wilding, L. 2005. Assessment of soil spatial variability at multiple scales. Ecological Modelling, 182(3-4):271–290.

Liu, S., Bliss, N., Sundquist, E., and Huntington, T. G. 2003. Modeling carbon dynamics in vegetation and soil under the impact of soil erosion and deposition. Global Biogeochemical Cycles, 17(2).

Loveland, P. and Webb, J. 2003. Is there a critical level of organic matter in the agricultural soils of temperate regions: a review. Soil and Tillage Research, 70(1):1–18.

Lugato, E., Bampa, F., Panagos, P., Montanarella, L., and 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, Z., Wang, E., Sun, O. J., Smith, C. J., and Probert, M. E. 2011. Modeling long-term soil carbon dynamics and sequestration potential in semi-arid agro-ecosystems. Agricultural and Forest Meteorology, 151(12):1529–1544.

MacMillan, R. A., Jones, R., and McNabb, D. H. 2004. Defining a hierarchy of spatial entities for environmental analysis and modeling using digital elevation models (DEMs). Computers, Environment and Urban Systems, 28(3):175–200.

Malamoud, K., McBratney, A., Minasny, B., and Field, D. J. 2009. Modelling how carbon affects soil structure. Geoderma, 149(1-2):19–26.

Mallat, S. 2009. *A Wavelet Tour of Signal Processing*. Academic Press.

Malone, B. P., McBratney, A. B., Minasny, B., and Laslett, G. M. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma, 154(1-2):138–152.

Manrique, L. A. and Jones, C. A. 1991. Bulk density of soils in relation to soil physical and chemical properties. Soil Sci. Soc. Am. J., 55(2):476–481.

Marchant, B. P., Villanneau, E. J., Arrouays, D., Saby, N. P. A., and Rawlins, B. G. 2015. Quantifying and mapping topsoil inorganic carbon concentrations and stocks: approaches tested in France. Soil Use and Management, 31(1):29–38.

Martin, M. P., Wattenbach, M., Smith, P., *et al.*. 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences, 8(5):1053–1065.

Martin, M. P., Orton, T. G., Lacarce, E., *et al.*. 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma.

Martinez, C., Hancock, G. R., and Kalma, J. D. 2010. Relationships between 137Cs and soil organic carbon (SOC) in cultivated and never-cultivated soils: An Australian example. Geoderma, 158(3-4):137–147.

Marx, S. 2013. *La cartographie des sols au Grand-Duché de Luxembourg*. Ph.D. thesis, Ministère de l'agriculture, de la viticulture et du développement rural, Ettelbruck.

Matheron, G. 1963. Principles of geostatistics. Economic geology, 58(8):1246–1266.

Matheron, G. 1982. Pour une analyse krigeante des données régionalisées. Centre de Géostatistique, Report N-732, Fontainebleau.

Mayr, T. R., Palmer, R. C., and Cooke, H. J. 2008. Digital Soil Mapping Using Legacy Data in the Eden Valley, UK. In Hartemink, A. E., McBratney, A., and Mendonça-Santos, M. d. L., editors, *Digital Soil Mapping with Limited Data*, pages 291–301. Springer Netherlands, Dordrecht.

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

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., and Van Molle, M. 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, 143(1-2):1–13.

Meersmans, J., van Wesemael, B., De Ridder, F., and Van Molle, M. 2009a. Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, 152(1-2):43–52.

Meersmans, J., van Wesemael, B., and Van Molle, M. 2009b. Determining soil organic carbon for agricultural soils: a comparison between the Walkley & Black and the dry combustion methods (north Belgium). Soil Use and Management, 25(4):346–353.

Meersmans, J., Martin, M. P., Lacarce, E., *et al.*. 2012a. A high resolution map of French soil organic carbon. Agronomy for Sustainable Development, 32(4):841–851.

Meersmans, J., Martin, M. P., Ridder, F. D., *et al.*. 2012b. A novel soil organic C model using climate, soil type and management data at the national scale in France. Agronomy for Sustainable Development, 32(4):873–888.

Milne, E., Adamat, R. A., Batjes, N. H., *et al.*. 2007. National and sub-national assessments of soil organic carbon stocks and changes: The GEFSOC modelling system. Agriculture, Ecosystems & Environment, 122(1):3–12.

Minasny, B. 2007. Predicting soil properties. Jurnal Ilmu Tanah dan Lingkungan, 7(1):54–67.

Minasny, B. 2013. Digital mapping of soil carbon. Advances in Agronomy, 1:1–47.

Minasny, B. and McBratney, A. B. 2005. The Matérn function as a general model for soil variograms. Geoderma, 128(3–4):192–207.

Minasny, B. and McBratney, A. B. 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma, 140(4):324–336.

Minasny, B., McBratney, A. B., Mendonça-Santos, M. L., Odeh, I. O. A., and Guyon, B. 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. Aust. J. Soil Res., 44(3):233–244.

Minasny, B., McBratney, A. B., Malone, B. P., and Wheeler, I. 2013. Digital Mapping of Soil Carbon. In *Advances in Agronomy*, volume 118, pages 1–47. Elsevier.

Moore, I. D., Grayson, R., and Ladson, A. 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological processes, 5(1):3–30.

Motaghian, H. R. and Mohammadi, J. 2011. Spatial estimation of saturated hydraulic conductivity from terrain attributes using regression, kriging, and artificial neural networks. Pedosphere, 21(2):170–177.

Mulder, V. L., de Bruin, S., Schaepman, M. E., and Mayr, T. R. 2011. The use of remote sensing in soil and terrain mapping — A review. Geoderma, 162(1-2):1–19.

Muñoz, J. D. and Kravchenko, A. 2011. Soil carbon mapping using on-the-go near infrared spectroscopy, topography and aerial photographs. Geoderma, 166(1):102–110.

Myers, D. B., Kitchen, N. R., Sudduth, K. A., *et al.*. 2011. Peak functions for modeling high resolution soil profile data. Geoderma, 166(1):74–83.

Nadeu, E., Gobin, A., Fiener, P., Van Wesemael, B., and Van Oost, K. 2015. Modelling the impact of agricultural management on soil carbon stocks at the regional scale: the role of lateral fluxes. Global change biology, 21(8):3181–3192.

Nocita, M., Stevens, A., Noon, C., and van Wesemael, B. 2013. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. Geoderma, 199:37–42.

Nocita, M., Stevens, A., Toth, G., *et al.*. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology and Biochemistry, 68:337–347.

Nocita, M., Stevens, A., van Wesemael, B., *et al.*. 2015. Soil spectroscopy: an opportunity to be seized. Global Change Biology, 21(1):10–11.

Nourzadeh, M., Mahdian, M. H., Malakouti, M. J., and Khavazi, K. 2012. Investigation and prediction spatial variability in chemical properties of agricultural soil using geostatistics. Archives of Agronomy and Soil Science, 58(5):461–475.

Odeh, I., McBratney, A., and Chittleborough, D. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma, 67(3-4):215–226.

Odgers, N. P., Libohova, Z., and Thompson, J. A. 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma, 189-190:153–163.

Odum, E. P. 1969. The Strategy of Ecosystem Development. Science (New York, N.Y.), 164:262–270.

Oliver, M. A. and Webster, R. 1986. Combining Nested and Linear Sampling for Determining the Scale and Form of Spatial Variation of Regionalized Variables. Geographical Analysis, 18(3):227–242.

Oliver, M. A., Webster, R., and Slocum, K. 2000. Filtering SPOT imagery by kriging analysis. International Journal of Remote Sensing, 21(4):735–752.

Pallasser, R., Minasny, B., and McBratney, A. B. 2015. Carbon Determination System for Whole Soil Cores. Communications in Soil Science and Plant Analysis, 46(2):221–234.

Parton, W. J. 1996. The CENTURY model. In Powlson, D. S., Smith, P., and Smith, J. U., editors, *Evaluation of Soil Organic Matter Models: Using Existing Long-Term Datasets*, pages 283–291. Springer Berlin Heidelberg, Berlin, Heidelberg.

Patterson, H. D. and Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika, 58(3):545–554.

Paustian, K., Lehmann, J., Ogle, S., *et al.*. 2016. Climate-smart soils. Nature, 532(7597):49.

Piccini, C., Marchetti, A., and Francaviglia, R. 2014. Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. Ecological Indicators, 36:301–314.

Pike, R. J., Evans, I. S., and Hengl, T. 2009. Geomorphometry: A brief guide. Developments in Soil Science, 33:3–30.

Ping, J. and Dobermann, A. 2006. Variation in the precision of soil organic carbon maps due to different laboratory and spatial prediction methods. Soil Science, 171(5):374–387.

Poeplau, C. and Don, A. 2015. Carbon sequestration in agricultural soils via cultivation of cover crops–A meta-analysis. Agriculture, Ecosystems & Environment, 200:33–41.

Ponce-Hernandez, R., Marriott, F. H. C., and Beckett, P. H. T. 1986. An improved method for reconstructing a soil profile from analyses of a small number of samples. Journal of Soil Science, 37(3):455–467.

Post, W. M., Izaurralde, R. C., Mann, L. K., and Bliss, N. 2001. Monitoring and verifying changes of organic carbon in soil. In *Storing Carbon in Agricultural Soils: A Multi-Purpose Environmental Strategy*, pages 73–99. Springer.

Powlson, D., Bhogal, A., Chambers, B., *et al.*. 2012. The potential to increase soil carbon stocks through reduced tillage or organic material additions in England and Wales: A case study. Agriculture, Ecosystems & Environment, 146(1):23–33.

Pribyl, D. W. 2010. A critical review of the conventional SOC to SOM conversion factor. Geoderma, 156(3):75–83.

Quinton, J. N., Govers, G., Van Oost, K., and Bardgett, R. D. 2010. The impact of agricultural soil erosion on biogeochemical cycling. Nature Geoscience, 3(5):311–314.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., *et al.*. 2013. The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma, 195–196:268–279.

Rawlins, B. G., Marchant, B. P., Smyth, D., *et al.*. 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. European Journal of Soil Science, 60(1):44–54.

Rosenbloom, N. A., Harden, J. W., Neff, J. C., and Schimel, D. S. 2006. Geomorphic control of landscape carbon accumulation. Journal of Geophysical Research: Biogeosciences, 111(G1).

Saby, N. P. A., Bellamy, P. H., Morvan, X., *et al.*. 2008. Will European soil-monitoring networks be able to detect changes in topsoil organic carbon content? Global Change Biology, 14(10):2432–2442.

Sanderman, J. and Chappell, A. 2013. Uncertainty in soil carbon accounting due to unrecognized soil erosion. Global Change Biology, 19(1):264–272.

Sanderman, J., Hengl, T., and Fiske, G. J. 2017. Soil carbon debt of 12,000 years of human land use. Proceedings of the National Academy of Sciences, 114(36):9575–9580.

Santra, P., Das, B., and Chakravarty, D. 2012. Spatial prediction of soil properties in a watershed scale through maximum likelihood approach. Environmental Earth Sciences, 65(7):2051–2061.

Sarker, J. R., Singh, B. P., Dougherty, W. J., *et al.*. 2018. Impact of agricultural management practices on the nutrient supply potential of soil organic matter under long-term farming systems. Soil and Tillage Research, 175:71–81.

Scharlemann, J., Tanner, E., Hiederer, R., and Kapos, V. 2014. Global soil carbon: Understanding and managing the largest terrestrial carbon pool. Carbon Management, 5(1):81–91.

Scheunert, I., Mansour, M., and Andreux, F. 1992. Binding of Organic Pollutants to Soil Organic Matter. International Journal of Environmental Analytical Chemistry, 46(1-3):189–199.

Schmidt, M. W. I., Torn, M. S., Abiven, S., *et al.*. 2011. Persistence of soil organic matter as an ecosystem property. Nature, 478(7367):49–56.

Schrumpf, M., Schulze, E. D., Kaiser, K., and Schumacher, J. 2011. How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories? Biogeosciences Discussions, 8(5):723–769.

Schumacher, B. A. 2002. Methods for the determination of total organic carbon (TOC) in soils and sediments. Technical report, United States Environmental Protection Agency.

Schwanghart, W. and Kuhn, N. J. 2010. TopoToolbox: A set of Matlab functions for topographic analysis. Environmental Modelling and Software, 25(6):770–781.

Selige, T., Böhner, J., and Schmidhalter, U. 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. Geoderma, 136(1-2):235–244.

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., and Zhou, Y. 2015. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis–NIR spectral library. European Journal of Soil Science, 66(4):679–687.

Simbahan, G. C., Dobermann, A., Goovaerts, P., Ping, J., and Haddix, M. L. 2006. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. Geoderma, 132(3-4):471–489.

Six, J., Conant, R., Paul, E. A., and Paustian, K. 2002. Stabilization mechanisms of soil organic matter: implications for C-saturation of soils. Plant and soil, 241(2):155–176.

Six, J., Bossuyt, H., Degryze, S., and Denef, K. 2004. A history of research on the link between (micro)aggregates, soil biota, and soil organic matter dynamics. Soil and Tillage Research, 79(1):7–31.

Sleutel, S., Neve, S., and Hofman, G. 2007. Assessing causes of recent organic carbon losses from cropland soils by means of regional-scaled input balances for the case of Flanders (Belgium). Nutrient Cycling in Agroecosystems, 78(3):265–278.

Smith, J., Smith, P., Wattenbach, M., *et al.*. 2005. Projected changes in mineral soil carbon of European croplands and grasslands, 1990–2080. Global Change Biology, 11(12):2141–2152.

Smith, P. 2012. Agricultural greenhouse gas mitigation potential globally, in Europe and in the UK: what have we learnt in the last 20 years? Global Change Biology, 18(1):35–43.

Smith, P., Smith, J. U., Powlson, D. S., *et al.*. 1997. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma, 81(1):153–225.

Smith, P., Chapman, S. J., Scott, W. A., *et al.*. 2007. Climate change cannot be entirely responsible for soil carbon loss observed in England and Wales, 1978–2003. Global Change Biology, 13(12):2605–2609.

Smith, P., Davies, C. A., Ogle, S., *et al.*. 2012. Towards an integrated global framework to assess the impacts of land use and management change on soil carbon: current capability and future vision. Global Change Biology, 18(7):2089–2101.

Stevens, A. and van Wesemael, B. 2008. Soil organic carbon stock in the Belgian Ardennes as affected by afforestation and deforestation from 1868 to 2005. Forest Ecology and Management, 256(8):1527–1539.

Stevens, A., Van Wesemael, B., Vandenschrick, G., Touré, S., and Tychon, B. 2006. Detection of carbon stock change in agricultural soils using spectroscopic techniques. Soil Science Society of America Journal, 70(3):844–850.

Stevens, A., Udelhoven, T., Denis, A., *et al.*. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma, 158(1-2):32–45.

Stevens, A., Miralles, I., and van Wesemael, B. 2012. Soil Organic Carbon Predictions by Airborne Imaging Spectroscopy: Comparing Cross-Validation and Validation. Soil Science Society of America Journal, 76(6):2174.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B. 2013. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLoS ONE, 8(6):e66409.

Stevens, F., Bogaert, P., Van Oost, K., Doetterl, S., and Van Wesemael, B. 2014. Regional-scale characterization of the geomorphic control of the spatial distribution of soil organic carbon in cropland. European Journal of Soil Science, 65(4):539–552.

Stevens, F., Bogaert, P., and van Wesemael, B. 2015a. Detecting and quantifying field-related spatial variation of soil organic carbon using mixed-effect models and airborne imagery. Geoderma, 259–260:93–103.

Stevens, F., Bogaert, P., and van Wesemael, B. 2015b. Spatial filtering of a legacy dataset to characterize relationships between soil organic carbon and soil texture. Geoderma, 237–238:224–236.

Stevenson, F. J. 1994. *Humus chemistry: genesis, composition, reactions*. John Wiley & Sons.

Stewart, C. E., Paustian, K., Conant, R. T., Plante, A. F., and Six, J. 2007. Soil carbon saturation: concept, evidence and evaluation. Biogeochemistry, 86(1):19–31.

Stewart, C. E., Plante, A. F., Paustian, K., Conant, R. T., and Six, J. 2008. Soil carbon saturation: linking concept and measurable carbon pools. Soil Science Society of America Journal, 72(2):379–392.

Stewart, C. E., Paustian, K., Conant, R. T., Plante, A. F., and Six, J. 2009. Soil carbon saturation: Implications for measurable carbon pool dynamics in long-term incubations. Soil Biology and Biochemistry.

Stocker, T. F., Qin, D., Plattner, G.-K., *et al.*. 2013. Climate Change 2013 The Physical Science Basis Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, International Panel for Climate Change (IPCC), New York.

Stockmann, U., Adams, M. A., Crawford, J. W., *et al.*. 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., *et al.*. 2015. Global soil organic carbon assessment.

Suuster, E., Ritz, C., Roostalu, H., Kolli, R., and Astover, a. 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. European Journal of Soil Science, 63(3):351–359.

Tan, Z. X., Lal, R., Smeck, N. E., and Calhoun, F. G. 2004. Relationships between surface soil organic carbon pool and site variables. Geoderma, 121(3-4):187–195.

Tekin, Y., Tumsavas, Z., and Mouazen, A. 2012. Effect of moisture content on prediction of organic carbon and pH using visible and near-infrared spectroscopy. Soil Science Society of America Journal, 76(1):188–198.

Temme, A. J. a. M., Peeters, I., Buis, E., Veldkamp, A., and Govers, G. 2011. Comparing landscape evolution models with quantitative field data at the millennial time scale in the Belgian loess belt. Earth Surface Processes and Landforms, 36(10):1300–1312.

Thamo, T. and Pannell, D. J. 2016. Challenges in developing effective policy for soil carbon sequestration: perspectives on additionality, leakage, and permanence. Climate Policy.

Thompson, J. A. and Kolka, R. K. 2005. Soil Carbon Storage Estimation in a Forested Watershed using Quantitative Soil-Landscape Modeling. Soil Science Society of America Journal.

Trumbore, S. E. and Czimczik, C. I. 2008. An uncertain future for soil carbon. Science, 321(5895):1455–1456.

Uno, Y., Prasher, S. O., Patel, R. M., *et al.*. 2005. Development of field-scale soil organic matter content estimation models in Eastern Canada using airborne hyperspectral imagery. Canadian Biosystems Engineering / Le Genie des biosystems au Canada, 47:09–1.

Van De Vreken, P., Beckers, V., Jacxsens, P., Van Meirvenne, M., and Van Orshoven, J. 2011. Historiek van de bodemdatabank Aardewerk-Vlaanderen-2010.

Van Hemelryck, H., Govers, G., Van Oost, K., and Merckx, R. 2011. Evaluating the impact of soil redistribution on the in situ mineralization of soil organic carbon. Earth Surf. Process. Landforms, 36(4):427–438.

van Lützow, M., Kögel-Knabner, I., Ekschmitt, K., *et al.*. 2006. Stabilization of organic matter in temperate soils: mechanisms and their relevance under different soil conditions–a review. European Journal of Soil Science, 57(4):426–445.

Van Oost, K., Govers, G., Quine, T. A., *et al.*. 2005. Landscape-scale modeling of carbon cycling under the impact of soil redistribution: The role of tillage erosion. Global Biogeochemical Cycles, 19(4).

Van Oost, K., Quine, T. A., Govers, G., *et al.*. 2007. The Impact of Agricultural Soil Erosion on the Global Carbon Cycle. Science, 318(5850):626–629.

Van Orshoven, J. and Vandenbroucke, D. 1993. Handleiding bij AARDEWERK, databank van profielgegevens in Belgie. Rapport nr 18A (Nederlands) en 18B (Frans).

Van Orshoven, J., Maes, J., Vereecken, H., Feyen, J., and Dudal, R. 1988. A structured database of Belgian soil profile data. Pedologie: bulletin van de Belgische bodemkundige vereniging, 38(2):191–206.

van Wesemael, B., Paustian, K., Andrén, O., *et al.*. 2010a. How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO2 fluxes in agricultural soils? Plant and Soil, 338(1-2):247–259.

van Wesemael, B., Paustian, K., Meersmans, J., *et al.*. 2010b. Agricultural management explains historic changes in regional soil carbon stocks. Proceedings of the National Academy of Sciences of the United States of America, 107(33):14926–14930.

VandenBygaart, A., Gregorich, E., and Angers, D. 2003a. Influence of agricultural management on soil organic carbon: A compendium and assessment of Canadian studies. Canadian Journal of Soil Science, 83(4):363–380.

VandenBygaart, A. J., Gregorich, E. G., and Angers, D. A. 2003b. Influence of agricultural management on soil organic carbon: A compendium and assessment of Canadian studies. Can. J. Soil Sci., 83(4):363–380.

VandenBygaart, A. J., Gregorich, E. G., Angers, D. A., and McConkey, B. G. 2007. Assessment of the lateral and vertical variability of soil organic carbon. Can. J. Soil Sci., 87(4):433–444.

Vandenbygaart, A. J., Kroetsch, D., Gregorich, E. G., and Lobb, D. 2012. Soil C erosion and burial in cropland. Global Change Biology, 18(4):1441–1452.

VandenBygaart, A. J., Kroetsch, D., Gregorich, E. G., and Lobb, D. 2012. Soil C erosion and burial in cropland. Global Change Biology, 18(4):1441–1452.

Vanwalleghem, T., Poesen, J., McBratney, A., and Deckers, J. 2010. Spatial variability of soil horizon depth in natural loess-derived soils. Geoderma, 157(1-2):37–45.

Vasques, G. M., Grunwald, S., Comerford, N. B., and Sickman, J. O. 2010. Regional modelling of soil carbon at multiple depths within a subtropical watershed. Geoderma, 156(3-4):326–336.

Vasques, G. M., Grunwald, S., and Myers, D. B. 2012a. Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. Journal of Geophysical Research, 117(G4).

Vasques, G. M., Grunwald, S., and Myers, D. B. 2012b. Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. Journal of Geophysical Research, 117(G4).

Viaud, V., Angers, D. A., and Walter, C. 2010a. Toward Landscape-Scale Modeling of Soil Organic Matter Dynamics in Agroecosystems. Soil Sci. Soc. Am. J., 74(6):1847–1860.

Viaud, V., Angers, D. A., and Walter, C. 2010b. Toward Landscape-Scale Modeling of Soil Organic Matter Dynamics in Agroecosystems. Soil Science Society of America Journal, 74(6):1847.

Viscarra Rossel, R. A. and Brus, D. J. 2018. The cost-efficiency and reliability of two methods for soil organic C accounting. Land Degradation & Development.

Walkley, A. and Black, I. A. 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Science, 37(1):29–38.

Wang, X., Wang, J., and Zhang, J. 2012. Comparisons of Three Methods for Organic and Inorganic Carbon in Calcareous Soils of Northwestern China. PLoS ONE.

Wang, X., Cammeraat, E. L. H., Cerli, C., and Kalbitz, K. 2014a. Soil aggregation and the stabilization of organic carbon as affected by erosion and deposition. Soil Biology and Biochemistry, 72:55–65.

Wang, X., Cammeraat, E. L. H., Cerli, C., and Kalbitz, K. 2014b. Soil aggregation and the stabilization of organic carbon as affected by erosion and deposition. Soil Biology and Biochemistry, 72:55–65.

Weiss, A. D. 2001. Topographic Position and Landforms Analysis. poster.

Wen, R. and Sinding-Larsen, R. 1997. Image Filtering by Factorial Kriging—Sensitivity analysis and application to Gloria side-scan sonar images. Mathematical Geology, 29(4):433–468.

Whelan, B. M. and McBratney, A. B. 2000. The "Null Hypothesis" of Precision Agriculture Management. Precision Agriculture, 2(3):265–279.

White, R. E., Davidson, B., Kee Lam, S., and Chen, D. 2018. A critique of the paper 'Soil carbon 4 per mille' by Minasny et al. (2017). Geofisica Internacional, 309:115–117.

Wiaux, F., Cornelis, J.-T., Cao, W., Vanclooster, M., and Van Oost, K. 2014. Combined effect of geomorphic and pedogenic processes on the distribution of soil organic carbon quality along an eroding hillslope on loess soil. Geoderma, 216:36–47.

Wickham, H. 2009. *Ggplot2 : elegant graphics for data analysis*. Springer.

Wieland, R., Dalchow, C., Sommer, M., and Fukuda, K. 2011. Multi-Scale Landscape Analysis (MSLA) — A method to identify correlation of relief with ecological point data. Ecological Informatics, 6(2):164–169.

Yang, Y., Fang, J., Tang, Y., *et al.*. 2008. Storage, patterns and controls of soil organic carbon in the Tibetan grasslands. Global Change Biology.

Yoo, K., Amundson, R., Heimsath, A. M., and Dietrich, W. E. 2006. Spatial patterns of soil organic carbon on hillslopes: Integrating geomorphic processes and the biological C cycle. Geoderma, 130(1-2):47–65.

Yu, D.-S., Shi, X.-Z., Wang, H.-J., *et al.*. 2007. National scale analysis of soil organic carbon storage in China based on Chinese soil taxonomy. PEDOSPHERE, 17(1):11–18.

Zádorová, T., Penížek, V., Vašát, R., *et al.*. 2015. Colluvial soils as a soil organic carbon pool in different soil regions. Geoderma, 253–254:122–134.

Zhang, C. and Kovacs, J. M. 2012. The application of small unmanned aerial systems for precision agriculture: a review. Precision Agriculture, 13(6):693–712.

Zhang, S., Huang, Y., Shen, C., Ye, H., and Du, Y. 2012. Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. Geoderma, 171-172:35–43.