Environmental factors related to soil organic carbon in Swedish forests at different spatial scales.

Robin Pfannendörfer

Degree project• 30 credits
Swedish University of Agricultural Sciences, SLU
Faculty of Natural Resources and Agricultural Sciences
Department of Soil and Environment
EnvEuro Environmental Science
Part number 2024:13
Uppsala 2024

Environmental factors related to soil organic carbon in Swedish forests at different spatial scales.

Robin Pfannendörfer

Supervisor:	**Marie Spohn, Swedish University of Agricultural Sciences, Department of Soil and Environment**
Co-supervisor:	Klaus Katzensteiner, BOKU – University of Natural Resources and Life Science, Department of Forest- and Soil Sciences
Examiner:	Björn Lindahl, Swedish University of Agricultural Sciences, Department of Soil and Environment

Credits:	30 credits
Level:	Master’s level (A2E)
Course title:	Master thesis in Environmental science
Course code:	EX0897
Programme/education:	EnvEuro
Course coordinating dept:	Department of Aquatic Science and Assessment
Place of publication:	Uppsala
Year of publication:	2024
Keywords:	Soil organic carbon, scale, random forest, climate, topography, soil texture, soil chemistry

Swedish University of Agricultural Sciences
Faculty of Natural Resources and Agricultural Sciences
Department of Soil and Environment
Biogeochemistry of Forest Soils

Abstract

Soil organic carbon (SOC) is an important pool in the carbon cycle that is dependent on a wide array of environmental variables. These are known as the soil forming factors: Parent material, climate, biota, topography and time. Which of those are best suited to predict SOC for a given area depends on the environmental conditions in that area (e.g. arid or humid, vegetation type), but is also understood to be dependent on spatial scale. The aim of this thesis was to investigate this scale dependency of the relations between SOC and environmental variables, using the example of Swedish forest soils. I hypothesized that climate variables would be most important at the national scale, texture would not depend on scale and topographic and chemical variables are the most important at smaller scales.

Data on soil properties, soil chemistry and vegetation from the Swedish forest soil inventory was combined with topographic and climatic data to train random forest models of soil organic carbon concentrations in different depths and soil organic carbon stocks in the organic layer. These models were created for the whole of Sweden and for parts of Sweden in the same way, enabling the analysis of how the spatial scale influenced variable importances. The nature of the relationships between SOC and environmental variables was investigated with linear regressions.

The main finding was that carbon concentrations are in general best predicted by chemical parameters and that the exact parameters depend more on sample depth than scale. Climate variables were more important at national scale as hypothesized, while the hypotheses regarding chemistry, topography and texture could not be confirmed. For carbon stocks, chemical variables were in total slightly less important and lead by Mn. The lower importance of chemical variables in comparison to carbon concentration models is mainly balanced by the higher importance of topographic variables. Overall, scale dependencies of relationships were observable only for climate, and overall, the variable importances were dominated by chemical variables.

Table of contents

List of tables

List of figures

Abbreviations

IQR	Interquartile range
MAP	Mean annual precipitation
MAT	Mean annual temperature
RCV	Robust coefficient of variation
RMSE	Root mean square error
SFSI	Swedish forest soil inventory
SOC	Soil organic carbon
WRB	World Reference Base for soil resources

1. Introduction

Soil organic carbon (SOC) is one of the biggest carbon pools in the global carbon cycle, surpassing both the atmospheric and vegetation pool in size (Intergovernmental Panel On Climate Change 2023). The uncertainty related to SOC stocks and their changes in response to global climate change in earth system models is considerable. In order to better predict future climate change, it is therefore necessary to improve our understanding of environmental controls on SOC. Furthermore, various other soil properties like soil structure, water balance and cation exchange capacity are influenced by the amount of soil organic matter (SOM) present (Blume et al. 2016). Identifying suitable environmental factors for SOC prediction can thus not only benefit climate modelling but also decision making in land management (Wiesmeier et al. 2019). Lastly, information on such environmental factors is sometimes more readily available than SOC data or can easily be obtained by remote sensing techniques. In these cases, being able to predict SOC instead of measuring it directly can save costs and time.

Based on the classic concept of soil forming factors, SOC is understood to be controlled by climate, vegetation, topography, parental material and other soil properties (Jenny 1941; Wiesmeier et al. 2019). In modern digital soil mapping approaches the same concept is expresses as SCORPAN: soil, climate, organisms, topography, parent material, age and space (McBratney et al. 2003). It is generally accepted that the spatial scales at which these factors operate or work best as predictors of SOC can differ substantially (Miller et al. 2015; Guo et al. 2019; Wiesmeier et al. 2019; Adhikari et al. 2020).

One way to conceptualise and understand this finding is to think of a “hierarchy of controls” (Wiesmeier et al. 2019). Climate controls for example carbon inputs to soil by affecting plant growth and, in conjunction with parent material, soil properties by affecting weathering processes and therefore sits on top of a hierarchy of controls (Manning et al. 2015). This is why SOC variability/distribution at broad scales can be explained by climatic conditions. At finer scales, its variability is no longer as pronounced, so that other environmental factors are becoming more important in relative terms (Wiesmeier et al. 2019). On pit and field scale for example, the variability in climate, land-use and hydrology diminishes, allowing researchers to focus on the effect of soil chemical and mineralogical properties (Moni et al. 2010).

The influence of climate on SOC is twofold and commonly seen as the strongest environmental control at the largest spatial scales (Wiesmeier et al. 2019). Net primary production (NPP), the net carbon fixation by plants at ecosystem scale, is mainly controlled by climatic variables. It strongly increases with increasing annual precipitation up to roughly 2000-3000 mm year^{-1} and also shows a positive relationship with temperature when excluding the driest ecosystems (Chapin et al. 2011). Microbial activity and thus decomposition of SOM is strongly promoted by increasing temperatures but can be inhibited under high soil moisture that creates anaerobic conditions (Chapin et al. 2011; Deluca & Boisvenue 2012). Put together, these climatic influences on biota lead to a gradient of increasing SOC stocks from warm and dry to cool and humid biomes (Wiesmeier et al. 2019).

This general pattern has been confirmed in several studies that modelled soil carbon at the continental scale in North America and Europe (De Brogniez et al. 2015; De Vos et al. 2015; Adhikari et al. 2020). Climatic variables are important predictors of different soil related carbon stocks in these studies, but there is no agreement whether precipitation or temperature is more relevant. In the study modelling carbon in the US, climatic variables exhibit lower relative importances at smaller spatial scales while land use and topographic variables gain in relative importance (Adhikari et al. 2020). Although I am not aware of a similar publication that specifically compares different scales in Europe, a lower importance of climatic variables for soil carbon at smaller scales is also evident in other studies (González-Domínguez et al. 2019; Wiesmeier et al. 2019). An important deviation from the general rule of highest SOC under cool and humid conditions can be found in European boreal forests where SOC in forest soils is positively related to temperature, but not with precipitation (De Brogniez et al. 2015; Spohn & Stendahl 2022). In this biome, evaporative demand is low and soil moisture is high. Therefore, tree growth is mainly controlled by temperature and not by precipitation (Osman 2013). Together with high soil moisture that slows down decomposition, this strong influence of temperature on tree growth might explain why in this biome, SOC increases with increasing temperature, contrary to the global trend.

Lastly, climatic conditions also influence weathering and soil formation, which means they can have an indirect effect on SOC stabilisation via soil geochemistry (Doetterl et al. 2015). This type of effect was shown in studies on a continental scale in line with the concept of “hierarchy of controls” (Doetterl et al. 2015; Rasmussen et al. 2018).

The type of vegetation and land use is another important factor that affects carbon input into soils and can lead to vast differences in SOC even under the same climatic conditions. Croplands are clearly the land use type with the lowest SOC stocks which can be explained by the change of soil physical properties and the cultivation of annual crops which allocate most of their carbon to aboveground biomass that gets removed (Wiesmeier et al. 2019; Beillouin et al. 2023). Forests, shrublands and grasslands all have higher SOC stocks than cropland, but there is no general order for these vegetation types (De Brogniez et al. 2015; Wiesmeier et al. 2019; Beillouin et al. 2023). Wetlands present an extreme case with SOC stocks far surpassing the other discussed vegetation types. This is due to anaerobic conditions caused by flooding that slow down decomposition of organic matter and therefore can result in the build-up of peat (Osman 2013). Histosols, the soils often found in wetlands, are precisely defined by thick layers of material rich in organic carbon (Blume et al. 2016). In one soil survey covering the whole EU, the median topsoil carbon content of wetlands was more than ten times that of woodlands and grasslands (De Brogniez et al. 2015).

Understanding how vegetation also causes SOC variation within woodlands has remained a challenging task and it is unclear whether tree species affect the SOC carbon stock over the whole soil profile (Wiesmeier et al. 2019). The methodological challenge is to disentangle the effect of tree species from site conditions, like climate, soil type nutrient status that influence both the SOC stock and tree species (Vesterdal et al. 2013). Tree species are however one of the most important factors for predicting forest floor SOC stocks in Europe (De Vos et al. 2015). Findings about this effect are consistent between both experimental and soil mapping approaches, showing higher SOC stocks under coniferous species (Oostra et al. 2006; Vesterdal et al. 2013; Heitkamp et al. 2021). Since the amount of litterfall is only weakly related to tree species, explanations for these differences focus on factors influencing the decomposition processes, like chemical properties of the plat biomass (Vesterdal et al. 2013). Coniferous species have more acidic litter and a study in Sweden reported that the C:N:P stoichiometry of the organic layer and even the mineral layer is influenced by tree species (Wiesmeier et al. 2019; Spohn & Stendahl 2022).

Topography impacts soil organic carbon by its effect on hydrological processes. Water erosion transports soil material and thus also SOC away from hillslopes and deposits it in lower, flatter areas (Doetterl et al. 2016). Additionally, water accumulation is favoured in “sites with a high contributing area, low inclination and concave curvature” (Wiesmeier et al. 2019) leading to high soil moisture. As previously discussed in relation to climate, high soil moisture inhibits decomposition and favours SOC accumulation.

Topographic features tend to be more useful predictors of SOC with decreasing scale, although there is notable variation in their usefulness between regions (Guo et al. 2019; Adhikari et al. 2020; Hounkpatin et al. 2021).

Soil types categorize soil by their properties that arise as a result of soil formation processes and thus contain information about soil forming factors that also control SOC. SOC content itself is also a criterion in many soil classification systems. Therefore soil type influences SOC stocks at different spatial scales and was identified as the most important predictor of SOC in mineral soils at the European level (De Vos et al. 2015; Wiesmeier et al. 2019). Soil type was also found to strongly influence the depth distribution of SOC in forest soils in Germany and Alaska (Deluca & Boisvenue 2012; Heitkamp et al. 2021).

In contrast to climatic conditions that control C input, soil chemical and textural properties control SOM stabilisation. While it was accepted for a long time that the chemical composition of SOM was a decisive factor for the rate of mineralisation, a new understanding of SOM dynamics has emerged recently (Schmidt et al. 2011). More emphasis is now put on SOM in the context of its abiotic and biotic environment. SOM that is bound to mineral surfaces for example is somewhat protected against decomposition. Fine particles of clay minerals or Fe and Al oxides and hydroxides can serve as sorption sites for SOM due to their high surface area and charge. Clay content or soil texture, for example, have been shown to correlate with SOC storage and storage potential on different spatial scales, in numerous climate zones and under diverse land uses (Wiesmeier et al. 2019; Heitkamp et al. 2021; Spohn & Stendahl 2024). Polyvalent metal cations, like Ca2+, Mg2+, Al3+ play a crucial role in SOM-clay interaction. SOM and clay by forming cation bridges between negatively charged clay minerals and SOM (Wiesmeier et al. 2019; Spohn & Stendahl 2024). Several studies have successfully used metal cations for predicting SOC contents (Rasmussen et al. 2018; Wiesmeier et al. 2019; Spohn & Stendahl 2024). Interestingly, these studies cover very large spatial scales (from national to continental) as well as very fine spatial scales (soil horizons in a profile). The scale dependency of the effect of metal cations on SOC is thus still an open question.

Soil pH is a key variable for sorption processes and therefore modulates the interaction of SOM with mineral surfaces (Rasmussen et al. 2018). Additionally, SOC concentrations in Swedish forest soils are higher in more acidic soils which can be explained by lower microbial respiration (Spohn & Stendahl 2024).

While the influence of all these environmental factors is well documented, the knowledge about their impact at different scales is often derived from reviewing multiple studies whereas studies that explicitly consider different spatial scales are rarer (Moni et al. 2010; Zhou et al. 2016; Guo et al. 2019; Wiesmeier et al. 2019; Adhikari et al. 2020). This thesis makes use of the Swedish forest soil inventory (SFSI), a national survey of forest soils including chemical and physical properties, covering all of Sweden (except agricultural and urban areas) to do so. Together with data on tree species (the forest soil inventory is linked to a forest inventory) and climatic and topographic data which can easily be combined with the SFSI via plot coordinates this allows me to cover a wide array of the soil forming factors and study how they together relate to SOC stocks. In order to evaluate the scale-dependence of these relationships, the national dataset is divided into four regional subsets which are further divided into four sub-regional subsets each. The aim of this thesis is to construct models predicting SOC concentrations and stocks on the basis of the environmental variables at different scales and use their results to analyse to analyse how the relationships between SOC and environment variables change with spatial scale. This investigation is, though not limited to, guided by three hypotheses.

Mean annual temperature (MAT) in Sweden is strongly correlated with latitude and thus varies mostly from north to south. Dividing Sweden into smaller areas by latitude should reduce climatic variability and therefore I hypothesize that climatic variables, are less important predictors for SOC at smaller spatial scales.

Soil texture has been found to be a good predictor of SOC in multiple studies, conducted at different scales. I hypothesize that the importance of texture for predicting SOC in Sweden will be similar across scales.

Although geochemistry was found to be related to SOC at very large and very small scales, I hypothesize that the importance of chemical variables for predicting SOC will be relatively higher on smaller scales, owing to the reduced relative importance of climatic variables.

Similarly, I hypothesize that the importance of topographic variables, with the exception of Elevation since it can sometimes serve as a proxy for climate, will be relatively higher on smaller scales.

2. Materials and Methods

2.1 Materials

2.1.1 Swedish Forest Soil Inventory

All soil data used in this study is taken from the Swedish Forest Soil Inventory. This Inventory collects data on vegetation and soil properties in all of Sweden, except for arable land, urban areas, and the high mountains. Combined sampling of vegetation and soils arose out of the National Forest Inventory in the early 1980s (Ranneby et al. 1987). The inventory comprises a total of around 23,500 permanent plots which are organized in quadratic clusters (tracts) on a triangular grid. Each tract is made up of four (in the southwestern part of Sweden) or eight circular plots (Figure 1). The plots are circular with a radius of 10m. To reflect the more fragmented landscape and diverse geology in the south of Sweden, the grid is denser there and the side length of clusters decreases from north to south.

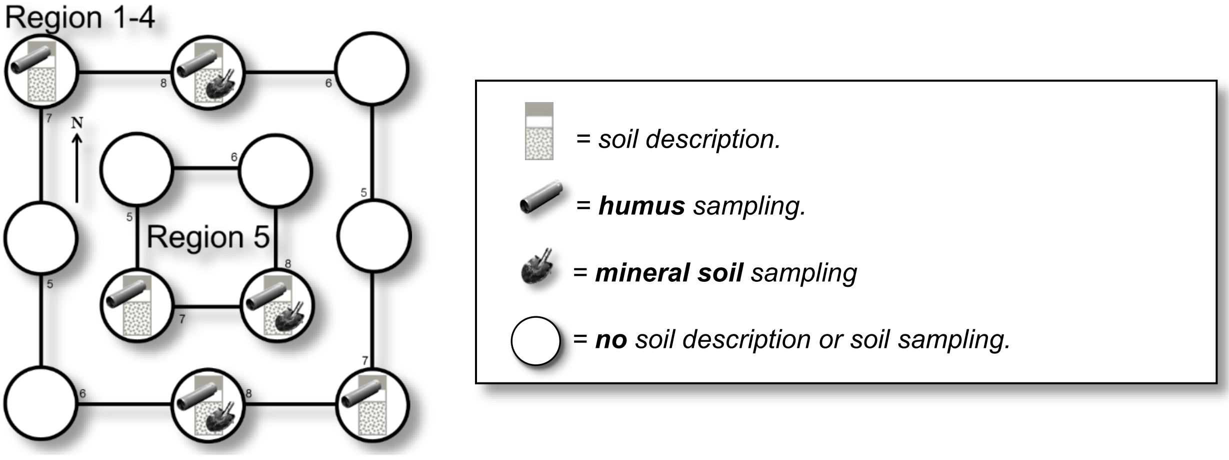
Region 1-4
N
Region 5
8
6
7
5
6
5
5
8
7
7
6
8
= soil description.
= **humus** sampling.
= ***mineral soil*** sampling
= ***no*** soil description or soil sampling.

Figure 1: Schematic of tract design. The eight or four circles represent the 10m radius circular and the symbols inside them indicate the type of sampling.

Field workers collect data on vegetation (dominant tree species, tree diameters, stand age etc) on all plots, while organic layer samples are collected at ca. 10.000 of these permanent plots and mineral soil samples are collected on ca. 4500 plots. Every year, around a tenth of plots from all over Sweden are visited, resulting in 10-year sampling periods, the latest of which started in 2023. Due to time-consuming laboratory work, not all the data from recent years is available yet so this study uses data from 2011 to 2020.

When a plot is visited, a subplot with a radius of 1m is created inside of it in which all soil data is collected. Between 1 and 9 organic layer samples of the entire humus layer, up to 30 cm for practical reasons, are taken with a 10cm diameter drill and combined into a single sample called H30 (Institutionen för skoglig resurshushållning och Institutionen för mark och miljö (SLU) 2022). The number of samples is chosen so that their combined volume is at least 1.5 litres. In some plots with humus layers exceeding 40cm thickness, an additional singular organic layer sample is taken from the depth 30-50cm. For mineral soil sampling and soil description, a pit is dug in the subplot and mineral soil sample are collected at predefined depth intervals into the mineral soil: 0-10cm, 10-20cm and sometimes 55-65cm. These samples are called M10, M20, M65. If the 55-65cm sample is taken, the soil type and texture, which are usually determined for the top 20 cm of the mineral soil, are in addition determined for the 55-65cm interval. The soil texture and WRB soil type are both determined at the same depth on site: In the presence of an E-horizon the assessment is done 10cm deep into the B horizon, in the absence it is done 20cm deep into the mineral soil. If the 55-65cm sample for soil chemistry is taken, the texture is additionally determined at this depth (Institutionen för skoglig resurshushållning och Institutionen för mark och miljö (SLU) 2022).

The tree species composition and stand age are determined for the 10m radius plots. In young conifer stands the age of trees is determined visually, by counting the annual shoots. Where this is not possible, trees are bored, and their age is determined by counting growth rings. Tree species are recorded as one of twelve classes which are simplified into five classes for this thesis (Table A 2).

The soil samples dried, weighed, sieved (<2mm) and living and dead roots >1mm are removed. All chemical analyses are carried out on the fine fraction <2mm: pH is measured in water with a solid to liquid ratio of 2.5 using a Pt electrode (Aquatrode Plus Pt1000, Metrohm). Calcium, Magnesium and Manganese are extracted in 1M ammonium acetate buffered at pH 7, while Aluminium is extracted using 1M potassium chloride buffered at pH 7. All cations are measured by inductively coupled plasma optical emission spectroscopy (ICP-OES; Avivo 200, Perkin Elmer).

2.1.2 Additional data

The climate variables mean annual temperature (MAT) and mean annual precipitation (MAP) are extracted from a 1km^2 resolution map of monthly historical climate data covering the years 1970-2000 from the WorldClim database (Fick & Hijmans 2017), based on the coordinates of plots.

Topographic variables are calculated based on a DEM of Sweden with a resolution of 2m which is based on measurements from airborne laser scanning (Seibert et al. 2007). They are elevation, aspect, downslope index (DSI), topographic wetness index (TWI) and accumulated flow. For their computation, smaller, 10km by 10km DEMs centred on the respective plot were created from the national DEM. The DSI as proposed by (Hjerdt et al. 2004) serves as a substitute for slope that also takes downslope conditions into account to better serve as an indicator for drainage conditions . It can be defined as the horizontal distance $L_d$in m that “a parcel of water has to travel along its flow path to lose a given head potential, d in m”(Hjerdt et al. 2004) or as the dimensionless gradient

$$\tan(\alpha_d) = \frac{L_d}{d} \quad (1)$$

In this study, the latter is used. The TWI is defined using the specific upslope area a (i.e. the upslope area per contour length) and the local slope tan(ß):

$$TWI = \ln(\frac{a}{\tan(\beta)}) \quad (2)$$

Although a unit can be assigned to the TWI, it is commonly and throughout this thesis regarded as a dimensionless number. Accumulated flow which can be understood as the number of cells or area that is drained through a cell is calculated as laid out in the paper by (Seibert et al. (2007).

Table 1. Environmental variables to be used as predictors of SOC

Variable	Unit/levels	Data source
Soil variables		
Carbon content (target variable)	in g/kg dry weight	elemental analyser (TruMac CN, LECO)
pH	-	soil:water ratio of 1:2.5, Pt electrode
Al concentration	mg/kg dry weight	1 M KCl extraction, ICP OES
Ca concentration	mg/kg dry weight	1 M sodium acetate extraction, ICP OES
Mg concentration	mg/kg dry weight	1 M sodium acetate extraction, ICP OES
Mn concentration	mg/kg dry weight	1 M sodium acetate extraction, ICP OES
texture	**levels**: cobble, gravel, coarse sand, medium sand, fine sand, coarse silt, fine silt, clay	assessed on site by hand
WRB soil type	**levels**: arenosol, cambisol, gleysol, histosol, leptosol, podzol, regosol, ubrisol	assessed on site
Climate variables		
mean annual temperature	°C	WorldClim dataset
mean annual precipitation	mm	WorldClim dataset
Vegetation variables		
stand age	years	assessed by visual inspection or boring
Tree species	**levels**: pine, spruce, mixed conifers, mixed, deciduous	assessed on site
Topographic variables		
downslope index (DSI)	-	derived from 2m DEM
elevation	m	derived from 2m DEM
aspect	°	derived from 2m DEM
topographic wetness index (TWI)	commonly used without	derived from 2m DEM
accumulated flow	-	derived from 2m DEM

2.1.3 Plot selection and subset definition

Not all plots included in the survey are used for analysis, due to disturbances, particularities in the sampling design, or missing data. If plots have a disturbed soil profile (e.g. old charcoal pits) or show signs of past ploughing this is recorded during data collection. Plots with this property are excluded from the analysis in this thesis. Among the remaining plots, only productive forests (in the SFSI defined as: productivity > $1m^3*ha^{-1}*year^{-1}$) with the humus form mor or moder are studied. The focus on productive forest land serves to exclude wetlands and grasslands that are part of the SFSI due to the way that land use classes are defined in Sweden. Plots with the humus form mull are sampled differently from those with other humus forms, namely to a fixed depth of 10cms, prohibiting comparisons of these humus samples with those for other humus forms. At plots with an organic layer thickness of more than 40cm, two separate samples are taken at different depths, which raises questions about how to combine the results of the chemical analysis from those two depths or whether to train separate models for them. Since outside of the humus form peat only very few plots are affected by this problem, they are also excluded. Plots with the humus form peat are completely exempt from the analysis as they by definition have high carbon concentrations and having this confirmed by models would not provide any new insights. Lastly, plots with the texture class peat are also excluded following the same reasoning. In the SFSI, this texture is assigned to soils with an organic layer exceeding 50cm thickness (which are not considered in this study), soils with a layer of charred plant material, and quagmires and flarks, two types of wetlands.

To study the effect of scale on the relationship between carbon concentrations and stocks and the predictor variables, the original national dataset is split into four regional subsets which are further split into four subregional subsets each. These divisions are done based on the coordinates of the plots before eliminating plots with disturbances, more than 40cm organic layer, etc. Therefore, the number of samples used in the analysis is not the same for all regions (and subregions), although they were originally of equal sizes. Regions are defined purely by latitude, with region 1 containing the southernmost quartile of plots (Figure 2). Subregions are defined by first splitting a region along its median longitude into two subsets and then splitting those two subsets by their respective median latitudes into two subsets each, resulting in four subsets per region (Figure 2).

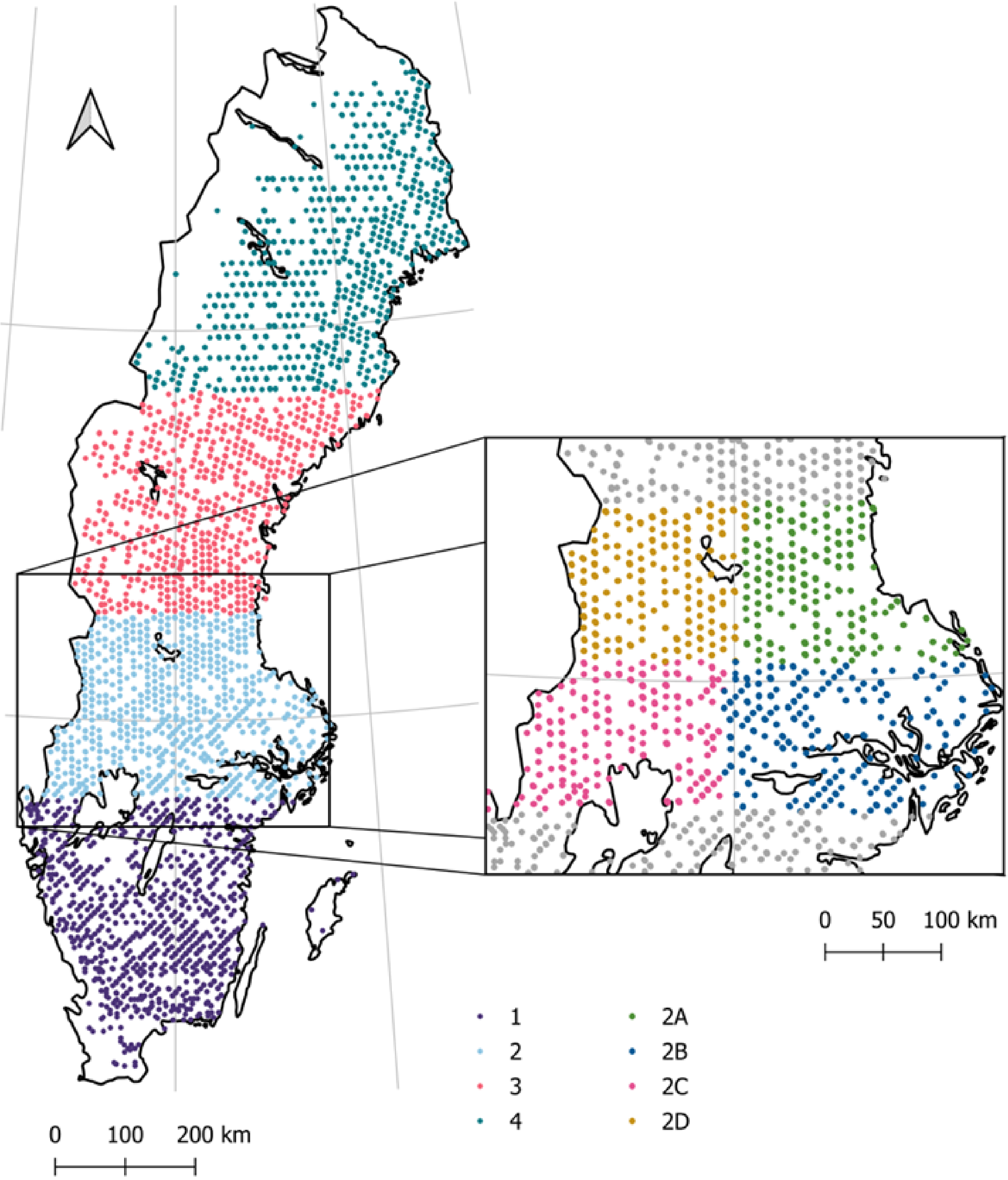

Figure 2. Map of the plots with organic layer sampling used in this thesis. The colours indicate the regions and subregion. Subregions 2A, 2B, 2C, 2D were chosen as an example, subregional delineations work the same way for the region 1,3 and 4.

2.2 Methods

2.2.1 Distribution and correlations of environmental variables

The frequency distributions of the environmental variables are analysed for samples from all depth increments and all areas separately. In the case of numerical predictors and the target variables carbon concentration and carbon stock this was done by calculating the minimum, mean, maximum, median, 1st and 3rd quartile. Carbon stocks are calculated from SFSI data, namely the product of the organic layer stock and the carbon concentration in the organic layer. To allow for comparability across scales as well as between different variables, a dimensionless measure of dispersion is used. To deal with the many skewed frequency distributions, the robust coefficient of variation (RCV) is chosen. This measure of dispersion is a more robust alternative to the coefficient of variation and calculated as follows:

$$RCV = 0.75 * \frac{Q_3 - Q_1}{Q_2}$$

where Q_3 is the third quartile, Q_1 the first quartile and Q_2 the median (Shapiro 2003; Arachchige et al. 2022). In this thesis the RCV is expressed in terms of percentages. For the distribution of MAT, the interquartile range (IQR: Q_3-Q_1) is given alternatively, since negative temperatures lead to negative RCVs which are not interpretable. The distributions of the three categorical variables WRB soil type, texture class and tree species were described by the relative frequency of their levels given in percentages.

The frequency distributions of variables that refer to plots as a whole and not to specific depth increments, namely the WRB soil type, the two vegetation variables, the two climatic ones, the five topographic ones are presented here for the dataset of organic layer samples, as it contains the most plots among the datasets (n = 2585). Since the soil texture is separately determined in 55-65cm depth in case the mineral layer is sampled at that depth interval, its frequency distribution is given based on the data from organic layer samples and mineral layer samples in 55-65 cm depth.

Half of all soil are classified as podzols (50 % of all plots) which are more common in the north. Together with regosols (32%), they make up at least 70% of all plots in every region and subregion (Table 2). The dominant tree species are mostly spruce and pine with more pines towards the north and more spruces towards the south (Table 3Table 3). The median stand age increases with latitude from 44 years in the southernmost region to 66 in the northernmost region (Table 4). Both MAP and MAT decrease with increasing latitude (Table 5) and their associated measures of variability (RCV for MAP and IQR for MAT) decrease from the national to the regional and subregional scale, when averaged over the respective datasets (data not shown). The frequency distributions of the soil texture class at 10/20 and 55-65cm depth are similar, with fine sand being the most common texture classes in most datasets. An important difference is the fact that the texture class "cobble" is not determined for the deeper samples (Table 6). Table 7 shows median values and RCV for the five topographic variables at the national scale. For elevation, the RCV decreases markedly on average from the national to the regional and subregional scale, while for the other four variables there is no such trend (data not shown).

Table 2. Frequency distribution of WRB soil types at national and regional scale

Area	Arenosol (%)	Cambisol (%)	Gleysol (%)	Histosol (%)	Leptosol (%)	Podzol (%)	Regosol (%)	Ubrisol (%)
National	2,6	0,9	3,9	1,9	8,9	49,5	32,1	0,3
Region 1	2,3	1,1	3,6	3,3	13,1	44,6	31,2	0,9
Region 2	2,4	1,9	4,4	2,5	11,4	46,2	31,1	0,0
Region 3	1,7	0,5	3,1	0,9	7,0	52,2	34,4	0,2
Region 4	3,9	0,0	4,3	0,6	4,2	55,0	31,9	0,2

Table 3. Frequency distribution of dominant tree species at national and regional scale

Area	Deciduous (%)	mixed (%)	mixed conifers (%)	pine (%)	spruce (%)
National	5,1	6,6	13,9	43,8	30,6
Region 1	7,7	6,9	12,3	26,6	46,5
Region 2	4,0	4,8	17,6	44,4	29,2
Region 3	3,9	6,9	11,6	50,0	27,7
Region 4	4,6	7,9	14,3	54,5	18,6

Table 4. Distribution of stand age at national and regional scale

Area	National	Region 1	Region 2	Region 3	Region 4
median stand age (years)	51	44	47	53	66
RCV (%)	93	85	87	106	101

Table 5. Distribution of climatic variables at national and regional scale

Area	National	Region 1	Region 2	Region 3	Region 4
median MAP (mm)	651	704	704	651	582
RCV of MAP (%)	15.4	24.2	15.1	9.5	9.7
median MAT (°C)	3,38	6,27	4,57	2,21	0,68
IQR of MAT (°C)	4,33	0,60	1,74	1,22	1,57

Table 6. Frequency distribution of texture classes at national scale

texture class	clay	fine silt	coarse silt	fine sand	medium sand	coarse sand	gravel	cobble
relative frequency at 10/20cm depth (%)	2,6	4,0	16,4	37,0	23,4	7,5	1,1	8,2
relative frequency at 55-65cm depth (%)	3,3	4,8	15,9	34,2	28,1	11,1	2,6	0

Table 7. Distributions of topographic variables at national scale

Variable	Elevation (m)	Aspect (°)	DSI (-)	TWI (-)	accumulated flow (-)
median	226	175	$69.8*10^{-3}$	4.19	4.24
RCV (%)	76	75	94	37	184

The five remaining variables (pH, Al, Ca, Mg, Mn) have vastly different values for the different depth increments, but there are distinct patterns across the four datasets. All five variables decrease with depth as can be seen in the tables below. The variability on the other hand only decreases with depth for pH while it increases for Mn and Mg. There is no clear pattern for Al and Ca. In most cases, the RCVs decrease from the national to the regional and subregional scale, but these differences are less pronounced than with depth or between variables. The RCVs of the pH distributions are the lowest, followed by Al, Mg, Ca and finally Mn.

Table 8. Distribution of chemical parameters in the organic layer at national level

Variable	pH	Al	Ca	Mg	Mn
median (- or $mg*kg^{-1}$)	3.74	211.5	1876.6	357.4	124.9
RCV (%)	9.0	91	61	48	129

Table 9. Distribution of chemical parameters in the mineral layer (0-10cm)at national level

Variable	pH	Al	Ca	Mg	Mn
median (- or mg*kg^{-1})	4.35	198.7	43.7	13.6	1.43
RCV (%)	7.6	71	154	103	262

Table 10. Distribution of chemical parameters in the mineral layer (10-20cm) at national level

Variable	pH	Al	Ca	Mg	Mn
median (- or mg*kg^{-1})	4.66	134.9	29.1	7.4	0.93
RCV (%)	6.1	86	222	122	257

Table 11. Distribution of chemical parameters in the mineral layer (55-65cm) at national level

Variable	pH	Al	Ca	Mg	Mn
median (- or mg*kg^{-1})	4.94	43.0	18.0	3.0	0.49
RCV (%)	6.1	89	177	166	175

2.2.2 Carbon prediction models: random forest with random feature elimination

Random forest (RF) models with recursive feature elimination were trained to predict carbon concentrations of all four sample types and carbon stocks of the organic layer for all datasets using the R package caret (Kuhn & Max 2008). Random forest is an algorithm for classifications and regressions that combines the results of numerous decision trees into one final output by voting or averaging over the results of the prediction trees. In general, the input to these models consists of the 16 environmental variables in Table 1 with the models for the sample type M65 using the texture class determined at 55-65cm depth into the mineral soil and the models for other sample types all using the texture class determined higher up in the mineral layer.

However, the preprocessing tool in the caret package (Kuhn & Max 2008) was used to filter out highly correlated variables. If the correlation between two predictors exceeds 0.8, the one with the higher mean absolute correlation, i.e. the average correlation to all other predictors, is removed. Across all areas and depth intervals this affects five variables: MAT, MAP, Elevation and the concentrations of calcium and magnesium. Elevation and MAT are highly correlated in most regional and subregional datasets, but never on the national scale. Elevation is more often excluded than MAT from the affected models. In the subregion 2B, MAP is additionally correlated to both MAT and temperature. This results in the exclusion of MAP from the models for subregion 2B. The chemical variables only exceed the correlation threshold in the mineral layer and more frequently in the lower depth increments within it. The concentrations of calcium and magnesium are highly correlated in roughly half the datasets for the mineral layer, but just like elevation and MAT never on the national scale. Calcium is excluded more often than magnesium in these cases.

Among the remaining variables, recursive feature elimination (RFE) is used to identify the optimal set of predictors, which works as follows: (1) a random forest model is trained using all remaining predictors, (2) its performance is evaluated, (3) the importances of all included predictors are calculated and (4) the least important predictor is removed. This procedure is repeated until only one predictor remains. Finally, the best performing model and its associated predictors are selected.

The RF models were trained with 10-fold cross-validation, meaning the data for the according region is split into ten evenly sized subsets (folds) of which nine are used as calibration data and one as validation data. This is then repeated 10 times so that every fold is used once as the validation data. Model performance is evaluated via the metric root mean square error (RMSE) which is defined as:

$$RMSE = (\frac{1}{n} * \sum\nolimits_{i=1}^{n} (C_{obs-}C_{pred})^2)^{1/2}$$

where n is the number of samples, C_{obs} is the observed carbon concentration or stock and C_{pred} is the predicted carbon concentration or stock. Variable importance is evaluated via the metric mean decrease in accuracy. This metric is calculated by comparing the accuracy of the RF model at hand with the accuracy of the same model for which the values of the predictor variable in question have been permutated. Permuting the predictors values is meant to break the causal link between this predictor and the target variable (carbon concentration/stock) and the difference in model accuracy is then seen as a measure for how much the predictor contributes to the accuracy model.

2.2.3 Model evaluation

The random forest models' performances are evaluated with the three measures coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). R^2 is chosen due to its wide usage and good interpretability, RMSE because it is the measure that the RFE uses to pick the best model and MAE offers information about the direction of errors, which the other measures can't. RMSE is defined as above, the other two measures as follows:

$$MAE = \frac{1}{n} * \sum\nolimits_{i=1}^{n} (C_{pred,i} - C_{obs,i})$$

$$R^2 = \frac{\sum_{i=1}^{n} (C_{pred,i} - \overline{C_{obs}})^2}{\sum_{i=1}^{n} (C_{obs,i} - \overline{C_{obs}})^2}$$

Where n denotes the number of samples, $C_{pred,i}$ the predicted carbon concentration or stock of the sample i, $C_{obs,i}$ the observed carbon concentration or stock of the sample i and $\overline{C_{obs}}$ the mean observed carbon concentration or stock. All three measures are given with standard which are deviations calculated based on the results for each fold in the 10-fold cross-validation of the final model with the optimal set of predictor variables. Variable importance of the variables included in the final model for a given area and sample type is calculated as the MDA just like for the selection of the optimal subset of variables. The resulting values are subsequently scaled, so that their sum equals 100% i.e. 1 to enable comparisons of variable importance between different models. Scaled down importances are referred to as relative importances from here on.

2.2.4 Linear regressions

Given the low interpretability of RF models, linear regressions (MLR) are calculated to gain a better understanding about the nature of the connection between important predictors and carbon concentrations and stocks. Linear regressions rely on linear relationships between predictor and predicted variables which may not always apply to the data used in this thesis. In cases where linear regressions are a suitable choice, they offer better interpretability than RF models due to their easy visualisation and by expressing the connection between predictors and carbon as a simple, linear function.

Linear regressions are calculated using the three most important predictors in the final RF models for carbon concentration and carbon stocks in the organic layer. For the purpose of this analysis, some variables were log transformed, to achieve linearity. DSI and carbon stocks are transformed by directly applying the decimal log, metal cation concentrations are transformed by adding one and then applying the decimal log. This is necessary since these concentrations can sometimes be below the detection limit (ergo 0 in the database) and therefore their logarithms are not defined. Results of the regression are only reported if the coefficient of the environmental variable is significantly different form 0 ($p < 0.05$). The normality of residuals is assessed with the Shapiro-Wilk normality test and models that pass ($p > 0.05$) are labelled.

3. Results

3.1 RF model evaluation and predictor importances

3.1.1 Organic layer

The RF models for the carbon concentrations in the organic layer performed better than those for carbon stocks across scales and regardless of the metric used to evaluate model performance (Figure 3, Table A 3, Table A 4. Cross-validation results of the RF models for carbon stock in the organic layer). For both model groups, the performance tended to be better on larger scales, but there was considerable variation, especially on the subregional scale. Carbon stock models used on average more variables than carbon concentration models (9.9 compared to 4.2), complicating comparisons of single variable importances between these groups.

Carbon concentration models almost exclusively relied on chemical predictor variables, with pH being the most important, independent of scale. Next to chemical variables, only vegetation and climate variables were used. At the national scale, only MAT and the five chemical variables were used, at the regional scale both climatic variables and stand age were used in addition to chemistry and at the subregional scale, stand age and tree species were the only non-chemical variables. The combined importance of chemical predictor variables peaked at the subregional scale (96.3%), while the combined importance of MAT and MAP declined from 8.8% at the national scale to 3.4% at the regional and 0% at the subregional scale.

For the carbon stock models, the most important predictor variable was always a chemical one, Mn in 17 and Ca in 4 models. The total importance of chemical variables was one third to one half lower than for carbon concentration models at the same scale, but chemical variables were still the most important group while topographic variables accounted for around half the importance of the non-chemical variables. In contrast to the carbon concentration models, there was no environmental variable that wasn’t used in at least one carbon stock model.

Climatic and topographic variables reached their highest total importance at the national scale, while the highest total importance of chemical variables was recorded at the subregional scale. The total importance of climatic variables was 11.3% at the national scale, decreasing to 6.5% at regional and 4.7% at subregional scale. The total importance of topographic variables was 27.2% at the national scale, decreasing to 19.1% at regional and 19.0% at subregional scale. The total importance of chemical variables was 46.8% at the national scale, increasing to 62.3% at regional and 65.6% at subregional scale. Texture was rarely used at all and its importance in a single model peaked at 5.1% resulting in low (<3%) average importances at all three scales.

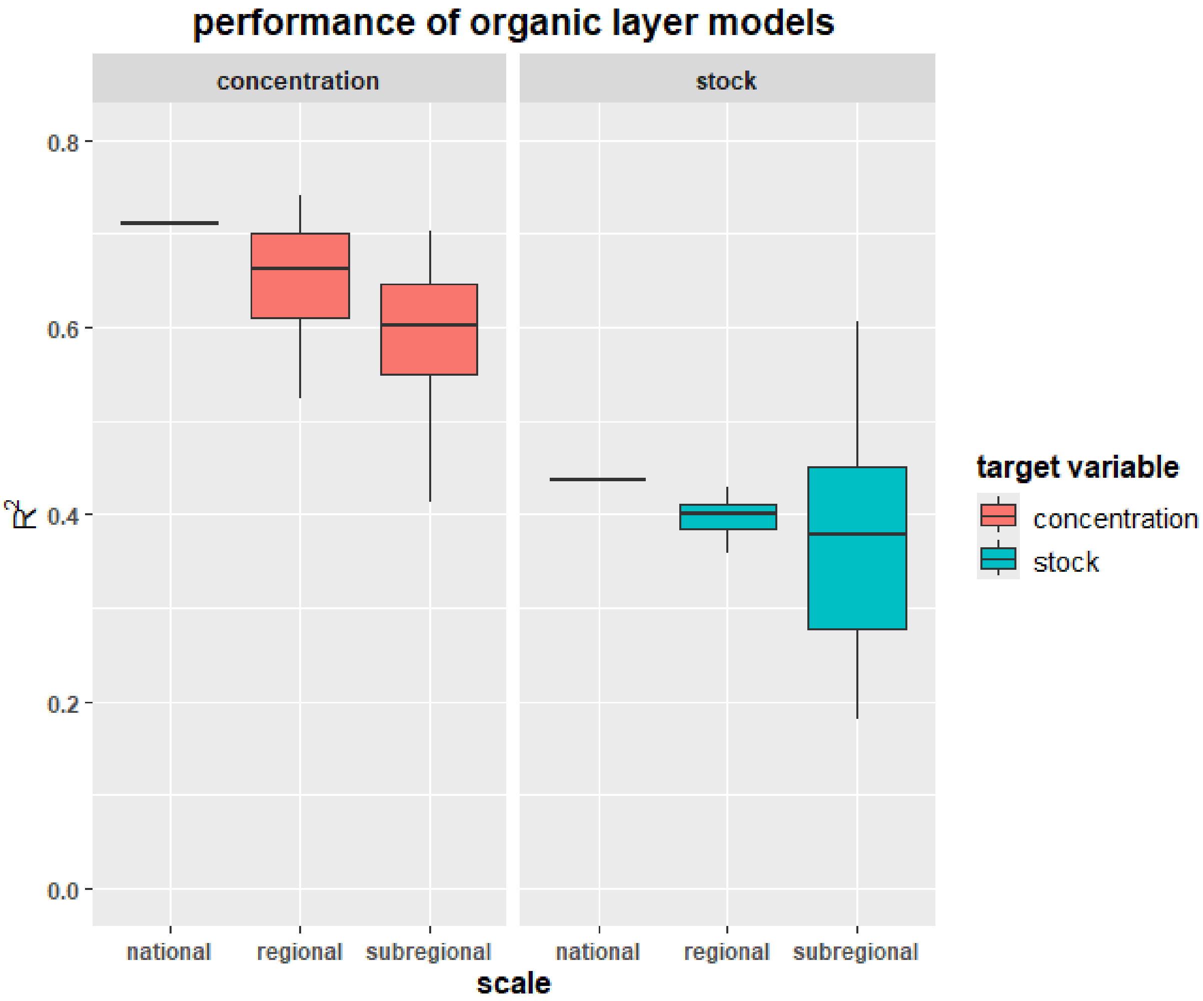

Figure 3: Performance of carbon concentration and carbon stock models for the organic layer in comparison.

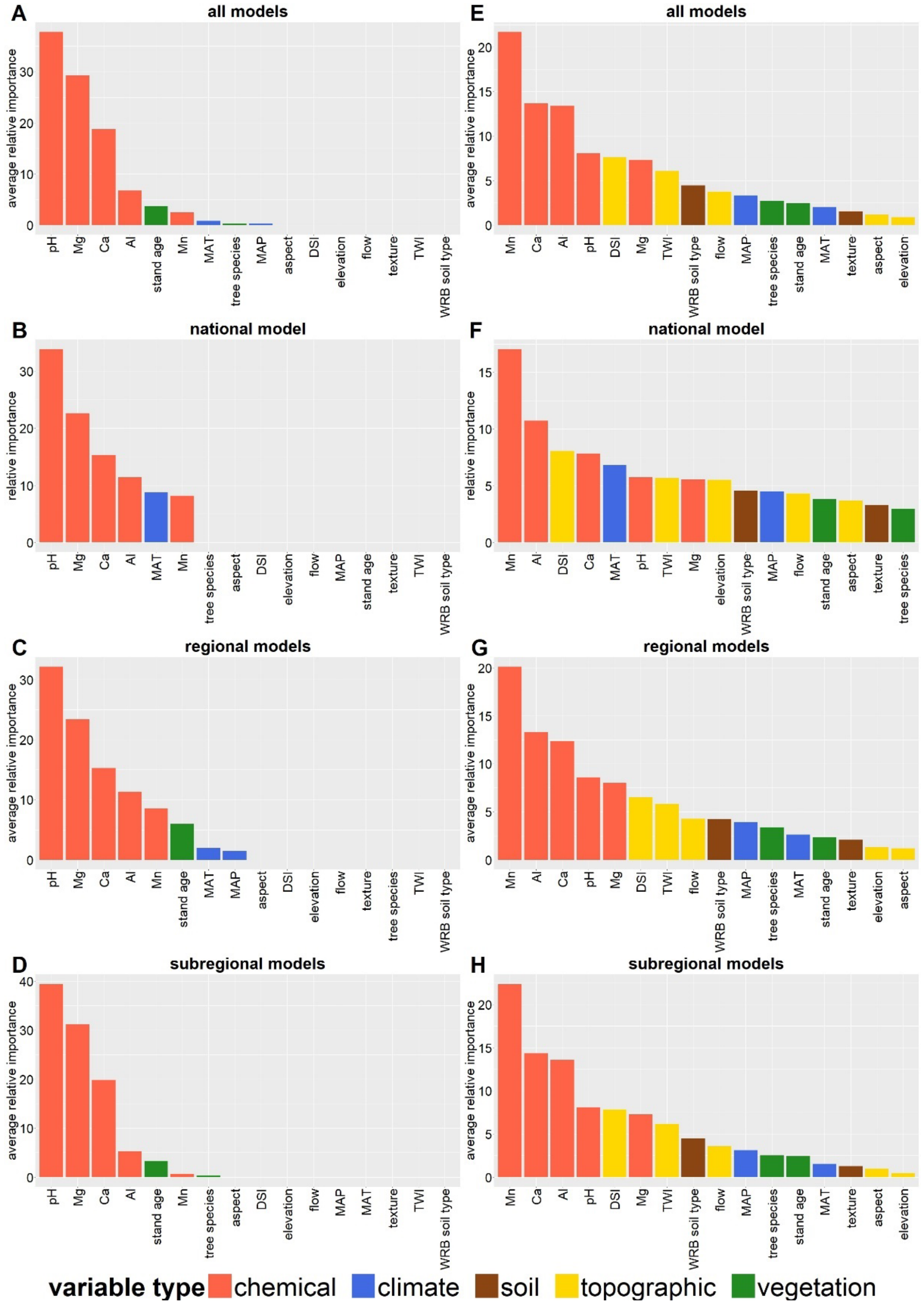

Figure 4: Relative variable importances of final random forest models for the organic layer. Panels A to D refer to models for carbon concentration, panels E to H to models of carbon stocks. Panels A and E show the average varaible importance across all 21 models, B and F only the variable importances in the respective national models, C and G the average across the four regional models and D and H the average across the 16 subregional models.

3.1.2 Mineral layer (M10, M20 and M65 samples)

In contrast to the organic layer models, the RF models for the different mineral layer samples did not show clear, consistent patterns for their performance with scale or sample type (Table A 5, Table A 6, Table A 7). Looking at R^2, the only measure that is not influenced by the scale of the carbon concentration and therefore comparable across sample types, the values for the national models and the median and mean values of all regional subregional models across the three mineral sample types were withing a range of 0.41 to 0.51 (Figure 5). Models for the 0-10cm depth interval (M10) used on average 7.9 variables, models for 10-20cms (M20) 8.1 and models for 55-65cm (M65) used 6.8 variables on average.

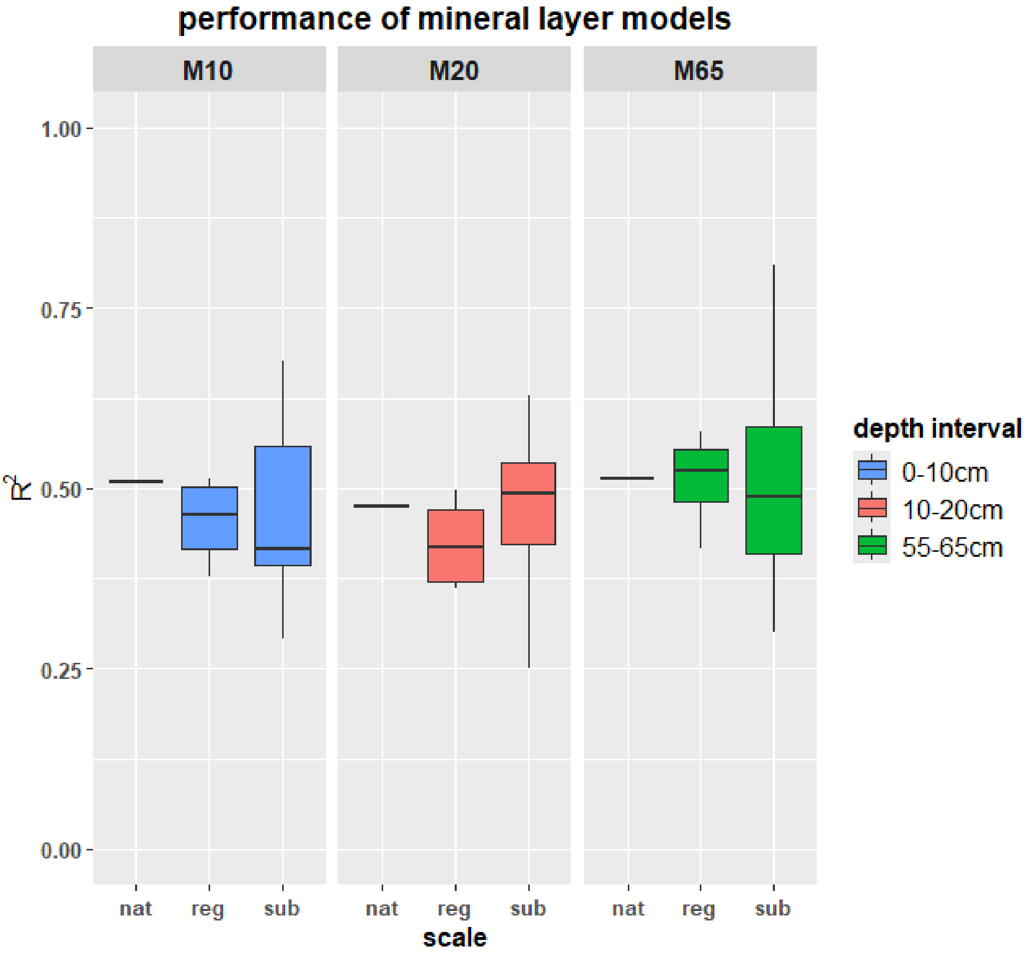

Figure 5: Performance of carbon concentration and carbon stock models for the mineral layer in comparison. Sample type refers to the depth interval, M10 is 0-10cm, M20 10-20cm and M65 55-65cm.

At all scales and for all depths, Al was rated as the most important predictor averaged over the respective models and reached its highest importance of for the deepest samples. The total importance of the chemical predictor variables was higher than for the carbon stock models, but lower than for the carbon concentration models in the organic layer, with values ranging from 57% to 78%.

In M10 models the total average importance of chemical variables peaked at the subregional scale, for M20 models at the national scale and for M65 models at the regional scale. The average combined relative importance of the climatic variables MAT and MAP decreased from the national to the regional and from the regional to the subregional scale for all depths and averaged across scales, it was highest for the M20 models, with 7.5%. Across all depths and scales, topographic variables were of similar importance to the climatic ones but did not exhibit patterns like the latter. Like for the chemical variables, the average combined importance of topographic variables was highest at different scales for different depths. The texture class was used in about a third of mineral layer models as a predictor, with a maximum importance of 19.1% in a single subregional model for 55-65cm. The average importance of texture was low at all scales and for all depths though, at most 3.9%.

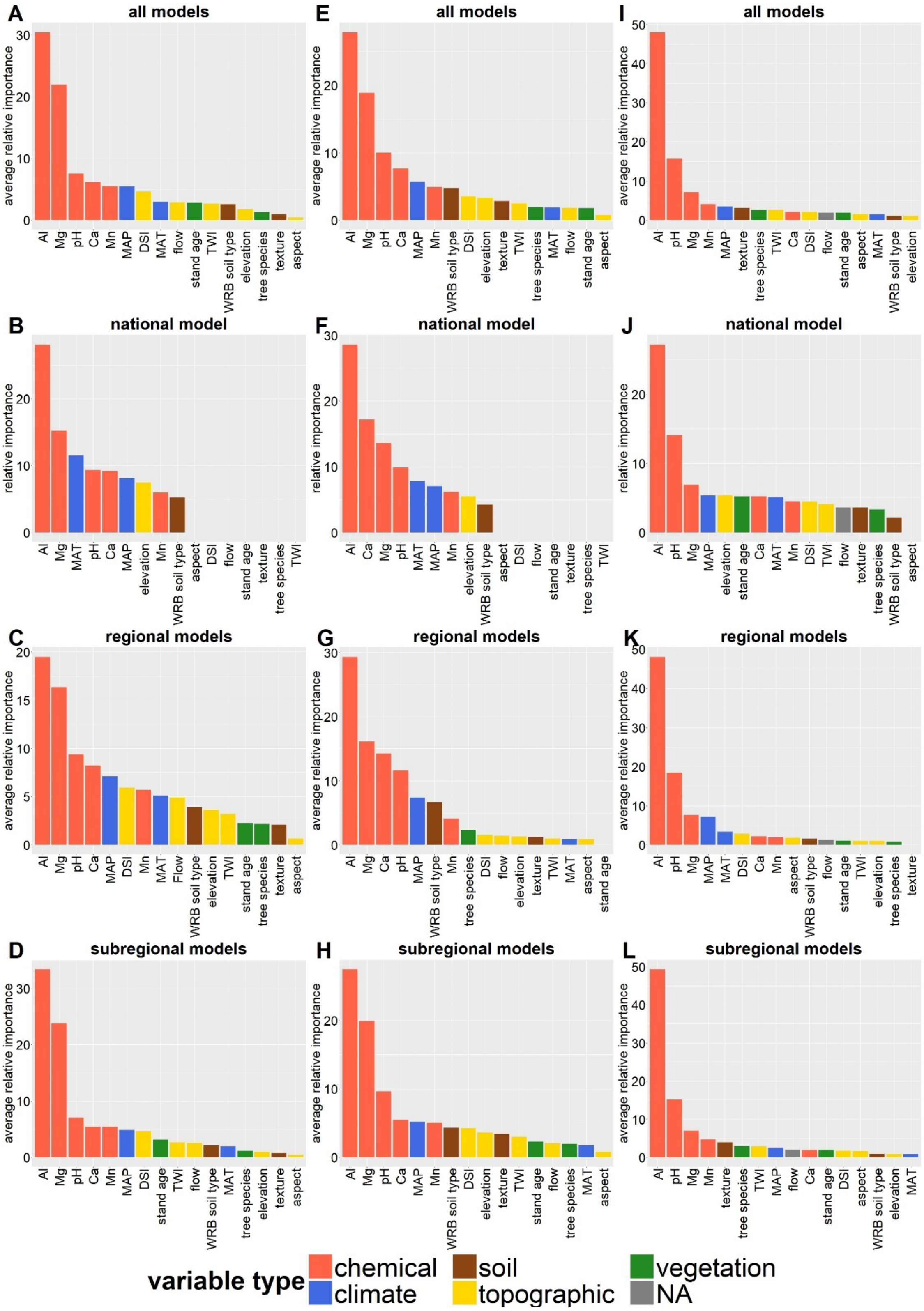

Figure 6: Relative variable importances of final random forest models for the mineral layer. Panels A to D refer to the depth increment 0-10cm, panels E to H to 10-20cm and panels I to L to 55-65cm depth. The panels in the first row show the average variable importance across all 21 models, those in the second row only the variable importances in the respective national models, those in the third row averages across the four regional models and those in the last row the averages across the 16 subregional models.

3.2 Linear regressions

In the organic layer, there were negative relationships between pH and the carbon concentration (R^2 from 0.10 to 0.41, Table 12) in all areas and positive relationships between the log-transformed magnesium concentration (R^2 from 0.04 to 0.44, Table 12) and the carbon concentration in nearly all areas. In regressions with calcium, the relationship between the log-transformed calcium concentration was also positive, with one exception in subregion 3D. The coefficients of determination were lower than for Mg and pH, with values between 0.03 and 0.15 (Table 12). Stand age was also found to be positively related to the carbon concentration in three areas each (Table 12). For none of the regressions regarding the carbon concentration were the residuals normally distributed.

In comparison, slightly more different variables were used in the regressions for the carbon stocks in the organic layer, since the corresponding RF models used more predictors. The most frequently used variables for the linear regressions of organic layer stocks were the log transformed Mn and Al concentrations with only one and two areas for which they were not used. Mn was negatively related to carbon stocks (R^2 from 0.04 to 0.40, Table 13) and Al was positively related to carbon stocks (R^2 from 0.05 to 0.26, Table 13). Although Ca was among the top 3 predictors in 9 RF models, only two of its regressions featured p-values below 0.05. Similarly, pH was used for four regressions, but none of those were statistically significant. Lastly, carbon stocks were negatively related to the log transformed DSI (9 cases) and positively related to the TWI (2 cases). The residuals of about two thirds of all significant regressions were normally distributed (Table 13).

Table 12. Results of the linear regressions with carbon concentration in the organic layer as dependent variable. Variables were transformed as described in 2.2.4.The 1st, 2nd and 3rd predictor refer to the ranking of the predictors in the final RF model for the respective area. The bracketed signs (+) and (-) refer to the sign of the slope.

Area	**1st predictor**	**p-value**	**adj. R^2**	**2nd predictor**	**p-value**	**adj. R^2**	**3rd predictor**	**p-value**	**adj. R^2**
National	pH	< 0.01	0.22 (-)	Mg	< 0.01	0.15 (+)	Ca	< 0.01	0,05 (+)
Region 1	pH	< 0.01	0.30 (-)	Mg	< 0.01	0.36 (+)	Ca	< 0.01	0.10 (+)
Region 2	pH	< 0.01	0.34 (-)	Mg	< 0.01	0.17 (+)	Ca	< 0.01	0.04 (+)
Region 3	pH	< 0.01	0.27 (-)	Mg	< 0.01	0.04 (+)	stand age	< 0.01	0.10 (+)
Region 4	pH	< 0.01	0.13 (-)	Mg	< 0.01	0.09 (+)	Ca	< 0.01	0.03 (+)
Subregion 1A	pH	< 0.01	0.36 (-)	Mg	< 0.01	0.28 (+)	Ca	< 0.01	0.11 (+)
Subregion 1B	pH	< 0.01	0.29 (-)	Mg	< 0.01	0.31 (+)	Ca	< 0.01	0.09 (+)
Subregion 1C	Mg	< 0.01	0.43 (+)	pH	< 0.01	0.28 (-)	Ca	< 0.01	0.15 (+)
Subregion 1D	Mg	< 0.01	0.44 (+)	pH	< 0.01	0.27 (-)	Ca	< 0.01	0.13 (+)
Subregion 2A	pH	< 0.01	0.41 (-)	Ca	> 0.05		Mg	< 0.01	0.10 (+)
Subregion 2B	pH	< 0.01	0.39 (-)	Mg	< 0.01	0.13 (+)	Ca	> 0.05	
Subregion 2C	Mg	< 0.01	0.36 (+)	pH	< 0.01	0.24 (-)	Al	> 0.05	
Subregion 2D	pH	< 0.01	0.33 (-)	Mg	< 0.01	0.17 (+)	-		
Subregion 3A	pH	< 0.01	0.20 (-)	Mg	< 0.01	0.15 (+)	Ca	0.01	0.04 (+)
Subregion 3B	pH	< 0.01	0.31 (-)	Mg	< 0.01	0.09 (+)	Ca	< 0.01	0.07 (+)
Subregion 3C	Mg	< 0.01	0.07 (+)	pH	< 0.01	0.14 (-)	Ca	> 0.05	
Subregion 3D	pH	< 0.01	0.31 (-)	Ca	> 0.05				
Subregion 4A	Mg	< 0.01	0.15 (+)	pH	< 0.01	0.12 (-)	Ca	< 0.01	0.04 (+)
Subregion 4B	pH	< 0.01	0.11 (-)	Mg	< 0.01	0.08 (+)	Ca	> 0.05	
Subregion 4C	pH	< 0.01	0.18 (-)	Ca	> 0.05		stand age	< 0.01	0.05 (+)
Subregion 4D	Mg	< 0.01	0.19 (+)	pH	< 0.01	0.10 (-)	Ca	< 0.01	0.09 (+)

Table 13. Results of the linear regressions with carbon stock in the organic layer as dependent variable. Variables were transformed as described in 2.2.4.The 1st, 2nd and 3rd predictor refer to the ranking of the predictors in the final RF model for the respective area. Variables in bold font indicate normality of residuals. The bracketed signs (+) and (-) refer to the sign of the slope.

Area	**1st predictor**	**p-value**	**adj. R^2**	**2nd predictor**	**p-value**	**adj. R^2**	**3rd predictor**	**p-value**	**adj. R^2**
National	Mn	< 0.01	0.29 (-)	Al	<0.01	0.12 (+)	DSI	< 0.01	0.09 (-)
Region 1	Mn	< 0.01	0.34 (-)	Al	<0.01	0.14 (+)	DSI	< 0.01	0.08 (-)
Region 2	Mn	< 0.01	0.25 (-)	**Al**	**<0.01**	**0.10 (+)**	Ca	>0.05	
Region 3	Ca	>0.05		**Mn**	**< 0.01**	**0.18 (-)**	Al	<0.01	0.05 (+)
Region 4	Mn	< 0.01	0.19 (-)	Al	<0.01	0.09 (+)	DSI	< 0.01	0.11 (-)
Subregion 1A	**Mn**	**< 0.01**	**0.23 (-)**	**Al**	**<0.01**	**0.08 (+)**	**WRB soil type**	**< 0.01**	**0.18**
Subregion 1B	**Mn**	**< 0.01**	**0.34 (-)**	**WRB soil type**	**< 0.01**	**0.14**	DSI	< 0.01	0.15 (-)
Subregion 1C	Mn	< 0.01	0.33 (-)	Al	<0.01	0.17 (+)	Mg	< 0.01	0.05 (-)
Subregion 1D	**Mn**	**< 0.01**	**0.35 (-)**	**Al**	**<0.01**	**0.20 (+)**	**TWI**	**0.01**	**0.03 (+)**
Subregion 2A	**Mn**	**< 0.01**	**0.28 (-)**	Ca	>0.05		**Al**	**<0.01**	**0.06 (+)**
Subregion 2B	**Mn**	**< 0.01**	**0.40 (-)**	**Al**	**<0.01**	**0.10 (+)**	**DSI**	**< 0.01**	**0.12 (-)**
Subregion 2C	**Mn**	**< 0.01**	**0.23 (-)**	**Al**	**<0.01**	**0.17 (+)**	**DSI**	**< 0.01**	**0.07 (-)**
Subregion 2D	**Mn**	**< 0.01**	**0.07 (-)**	**DSI**	**< 0.01**	**0.08 (-)**	Ca	>0.05	
Subregion 3A	Ca	>0.05		**Al**	**<0.01**	**0.26 (+)**	**Mn**	**<0.01**	**0.26 (+)**
Subregion 3B	Mn	< 0.01	0.11 (-)	pH	>0.05		**Al**	**<0.01**	**0.10 (+)**
Subregion 3C	Ca	>0.05		**TWI**	**0.02**	**0.02 (+)**	pH	>0.05	
Subregion 3D	**Ca**	**<0.01**	**0.05 (+)**	**Mn**	**< 0.01**	**0.30 (-)**	pH	>0.05	
Subregion 4A	**Mn**	**< 0.01**	**0.24 (-)**	**DSI**	**< 0.01**	**0.15 (-)**	Al	<0.01	0.08 (+)
Subregion 4B	**Mn**	**< 0.01**	**0.31 (-)**	Ca	>0.05		**DSI**	**< 0.01**	**0.14 (-)**
Subregion 4C	Mg	>0.05		**Mn**	**< 0.01**	**0.16 (-)**	**Al**	**<0.01**	**0.12 (+)**
Subregion 4D	**Mn**	**< 0.01**	**0.04 (-)**	**Ca**	**0.01**	**0.03 (+)**	pH	>0.05	

4. Discussion

In this thesis I compared relative variable importances from RF models that were trained to predict soil organic carbon concentrations at different depth increments and soil organic carbon stocks in the organic layer of Swedish forest soils. The clearest finding regarding the selection of variables was the high importance of chemical variables in these models, at all scales and for all depths. For most relationships though, the hypotheses about the way they behave on different scales were not confirmed.

Model performances in this thesis were slightly better compared to those in a similar study that also used RF models to predict SOC stocks in Sweden at two scales, although with different environmental variables (Hounkpatin et al. 2021). The authors divided Sweden into three sub-areas based on climatic conditions and reported R^2 values from cross validation of RF models with different sets of predictors ranging from 0.15 to 0.28 for the national models and from 0.08 to 0.32 for models for the sub-areas. The final carbon stock model in my case achieved an R^2 of 0.44±0.07 at the national scale and the final models at regional scale 0.40±0.14. I am not aware of other studies with a comparable scope and geographic focus against which to compare the performance of the models in this thesis.

Fewer studies seem to model carbon concentration as compared to carbon stocks. Given this, two studies using data from the Tibetan plateau for training RF models of carbon concentration have to be used for comparison (Yang et al. 2016; Dai et al. 2022). Using somewhat similar environmental variables (topography, climate, land cover/use) the authors reported R^2 of 0.32 and 0.68 for RF models of the carbon concentration in the top 20cm. This range roughly corresponds to the range of performances of the RF models trained for this thesis and underlines how much variation in model variation can be observed even with similar data inputs between areas. Especially for the organic layer models it is also important to note that the carbon concentration in these samples is by definition high as the organic layer is defined by its organic matter content which is closely related to soil organic carbon concentration (Blume et al. 2016; Institutionen för skoglig resurshushållning och Institutionen för mark och miljö (SLU) 2022). Variation of carbon concentration can thus probably be best explained by mixing in of mineral material due to biological activity, an aspect not directly covered by the environmental factors chosen for this study which mostly focus on decomposition of SOM (Institutionen för skoglig resurshushållning och Institutionen för mark och miljö (SLU) 2022). Carbon stocks on the other hand should arguably be less affected by this caveat.

Mn was found to be the most important predictor of SOC stocks in the organic layer in this study and this relationship is practically independent of scale. This prominent role of exchangeable Mn is in line with findings by Stendahl et al. (2017) who also studied Swedish forest soils using SFSI data and found Mn to be the most important predictor of SOC, using another statistical approach than this thesis. The authors propose a mechanism centred around Mn-peroxidases produced by fungi to explain the negative relationship between Mn and SOC stocks (Stendahl et al. 2017). Given the nature of this thesis this mechanism could neither be confirmed not rejected, but their findings about the relevance of Mn for SOC storage held.
For this specific relationship between Mn and SOC stocks, a connection to scale can be seen in the variable importance. The relative importance of Mn in the models for the organic layer stock increases from 17.0% to 20.1% at the regional and 22.4% at the subregional scale. Neither for Mn in other model types nor for the chemical variables taken together a comparable gradient of importance with regards to scale was observed, so that the hypothesis regarding geochemistry could not be confirmed. Depending on the model type, i.e. sample depth and target variable, the highest total importance of chemical variables was reached at different scales or increased.

The importance of climatic variables in relation to scale were in agreement with my hypothesis and they also changed with depth in line with what the literature suggests, although there have also been studies in which models did not exhibit interaction between depth and the climate to SOC relationship (Wiesmeier et al. 2019; Heitkamp et al. 2021). The combined importance of the climatic variable MAT and MAP mostly decreases with depth and at smaller spatial scales. In contrast to this trend, subregional models for the organic layer carbon concentration do not use climatic variables at all, while they are to some degree used at all other scales and for all model types, which does not necessarily mean that all (sub)regional models included them. This can be explained by the fact that the carbon concentration models for the organic layer generally utilize the fewest number of variables with an average of 4.2. Decreasing importance of climatic variables as predictors can plausibly be explained by their decreasing variability. The IQR of MAT for example is 4.3°C at the national scale, decreasing to an average of 1.3°C at the regional and 1.0°C at the subregional scale. The trend of decreasing importance with depth is also unsurprising, given that the deeper parts of soils are somewhat insulated from influence of e.g. temperature fluctuations at the surface.

Elevation can be seen as a variable standing somewhat in between other topographic and climatic variables. It obviously is a topographic one by definition but also used as a proxy for climatic conditions (Wiesmeier et al. 2019). In nearly all datasets for this thesis, elevation was highly correlated to climate mostly to MAT. This impacts how the total importance of topographic variables behaves. Including elevation, the total importance of topographic variables for the organic layer carbon stock models steadily decreases from the national to the regional and subregional scale, from 27.2% to 19.0%. Leaving out elevation and only adding up the remaining four topographic variables, accumulated flow, aspect, TWI and DSI, removes this clear pattern, resulting in values of 21.77% at the national, 17.8% at the regional and 18.6% at the subregional scale. Similarly, the total importance of topographic variables excluding elevation doesn't vary strongly between scales for the carbon concentration models too. With both approaches, the fourth hypothesis does not hold, since the total importance of topographic variables is either highest at the national scale or fairly similar across scales, but certainly not the highest at the subregional scale.

The texture class was rarely used in my models and evaluated as being relatively unimportant, which was surprising since previous research using a smaller part of the data from the SFSI found higher carbon concentrations in the upper 10cm of the mineral layer of fine-textured soils (Spohn & Stendahl 2024). For other regions of the world, soil texture has been successfully used as a predictor in regression models of the SOC storage capacity; an overview can be found in Wiesmeier et al. (2019). The low importance of the texture class as a predictor of SOC in this thesis could be due to the way in which this variable is recorded. In the SFSI texture is recorded as one of eight classes, while Wiesmeier et al. (2019) compiled studies utilizing the proportion of the fine mineral fraction. However, as previously stated, other authors still reported significant influence of soil texture on the SOC content in the organic layer at the national scale working with the same type of texture data from the SFSI, so this explanation cannot fully account for the low relevance of texture in my models (Spohn & Stendahl 2024). Secondly, most samples have either the texture class fine or medium sand, meaning that for a sizeable proportion of samples in all datasets, regardless of scale, differences between their SOC cannot be explained by looking at their texture.

Keeping the problem of overall low importance of the texture class in mind, it can still be said that there is no clear trend for the importance of texture with scale and that there are only small differences between the importance of texture at different scales. Therefore, the second hypothesis, i.e. that the importance of texture will be similar across scales, can be tentatively confirmed but given the low overall importance of texture, I advise for caution regarding the validity of this finding.

5. Conclusion

In this thesis the carbon content in Swedish forest soils and the carbon stocks in the organic layer were modelled on the basis of up to 16 environmental using random forest with recursive feature elimination variables at different spatial scales. The performance of the final RF models and the number of predictors that were actually used varied strongly between the organic and mineral layer and between the target variable: carbon concentration or carbon stock.

Across all model types and scales, chemical variables as a group were of the highest importance for the RF models, while the most important variable depends on scale and model type. Mn was found to be of high importance for SOC stocks in the organic layer and negatively related to them, in line with prior research on Swedish forest soils.

The hypothesis of decreasing importance of climatic factors at smaller scales was confirmed for all model types and can be explained by a decreasing variability of MAT at smaller scales. Texture, which was hypothesised to be of similar importance across scales, was in total one of the least important predictors on my models while many authors were able to produce quite well performing models of SOC with texture as a predictor. This difference might arise from the fact that texture in the SFSI is recorded as a categorical variable with 8 classes most of which were rare. In any case the hypothesis of similar importance across scales for texture is not refuted by my findings. For none of the other variable types could the hypothesis regarding how their relationships to SOC change with scale be confirmed. Elevation presents itself as a somewhat special case, being categorised as a topographic variable, but its role in the models often closely resembles that of MAT to which it is highly related.

In summary, the statistical approach was able to uncover the influence of the factor scale onto relationships between SOC and environment variables, but less often than hypothesised. For the cases that were found, the influence of scale was expected to show itself in that way and can be explained.

References

Adhikari, K., Mishra, U., Owens, P.R., Libohova, Z., Wills, S.A., Riley, W.J., Hoffman, F.M. & Smith, D.R. (2020). Importance and strength of environmental controllers of soil organic carbon changes with scale. *Geoderma*, 375, 114472. https://doi.org/10.1016/j.geoderma.2020.114472

Arachchige, C.N.P.G., Prendergast, L.A. & Staudte, R.G. (2022). Robust analogs to the coefficient of variation. *Journal of Applied Statistics*, 49 (2), 268–290. https://doi.org/10.1080/02664763.2020.1808599

Beillouin, D., Corbeels, M., Demenois, J., Berre, D., Boyer, A., Fallot, A., Feder, F. & Cardinael, R. (2023). A global meta-analysis of soil organic carbon in the Anthropocene. *Nature Communications*, 14 (1), 3700. https://doi.org/10.1038/s41467-023-39338-z

Blume, H.-P., Brümmer, G.W., Fleige, H., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K. & Wilke, B.-M. (2016). *Scheffer/Schachtschabel Soil Science*. 1st ed. 2016. Springer Berlin Heidelberg : Imprint: Springer. https://doi.org/10.1007/978-3-642-30942-7

Chapin, F.S., Matson, P.A. & Vitousek, P.M. (2011). *Principles of Terrestrial Ecosystem Ecology*. Springer New York. https://doi.org/10.1007/978-1-4419-9504-9

Dai, L., Ge, J., Wang, L., Zhang, Q., Liang, T., Bolan, N., Lischeid, G. & Rinklebe, J. (2022). Influence of soil properties, topography, and land cover on soil organic carbon and total nitrogen concentration: A case study in Qinghai-Tibet plateau based on random forest regression and structural equation modeling. *Science of The Total Environment*, 821, 153440. https://doi.org/10.1016/j.scitotenv.2022.153440

De Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L. & Van Wesemael, B. (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. https://doi.org/10.1111/ejss.12193

De Vos, B., Cools, N., Ilvesniemi, H., Vesterdal, L., Vanguelova, E. & Carnicelli, S. (2015). Benchmark values for forest soil carbon stocks in Europe: Results from a large scale forest soil survey. *Geoderma*, 251–252, 33–46. https://doi.org/10.1016/j.geoderma.2015.03.008

Deluca, T.H. & Boisvenue, C. (2012). Boreal forest soil carbon: distribution, function and modelling. *Forestry*, 85 (2), 161–184. https://doi.org/10.1093/forestry/cps003

Doetterl, S., Berhe, A.A., Nadeu, E., Wang, Z., Sommer, M. & Fiener, P. (2016). Erosion, deposition and soil carbon: A review of process-level controls, experimental tools and models to address C cycling in dynamic landscapes. *Earth-Science Reviews*, 154, 102–122. https://doi.org/10.1016/j.earscirev.2015.12.005

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., Casanova-Katny, A., Muñoz, C., Boudin, M., Zagal Venegas, E. & Boeckx, P. (2015). Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience*, 8 (10), 780–783. https://doi.org/10.1038/ngeo2516

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37 (12), 4302–4315. https://doi.org/10.1002/joc.5086

González-Domínguez, B., Niklaus, P.A., Studer, M.S., Hagedorn, F., Wacker, L., Haghipour, N., Zimmermann, S., Walthert, L., McIntyre, C. & Abiven, S. (2019). Temperature and moisture are minor drivers of regional-scale soil organic carbon dynamics. *Scientific Reports*, 9 (1), 6422. https://doi.org/10.1038/s41598-019-42629-5

Guo, Z., Adhikari, K., Chellasamy, M., Greve, M.B., Owens, P.R. & Greve, M.H. (2019). Selection of terrain attributes and its scale dependency on soil organic carbon prediction. *Geoderma*, 340, 303–312. https://doi.org/10.1016/j.geoderma.2019.01.023

Heitkamp, F., Ahrends, B., Evers, J. & Meesenburg, H. (2021). Spatial 3D mapping of forest soil carbon stocks in Hesse, Germany. *Journal of Plant Nutrition and Soil Science*, 184 (6), 635–656. https://doi.org/10.1002/jpln.202100138

Hjerdt, K.N., McDonnell, J.J., Seibert, J. & Rodhe, A. (2004). A new topographic index to quantify downslope controls on local drainage. *Water Resources Research*, 40 (5), 2004WR003130. https://doi.org/10.1029/2004WR003130

Hounkpatin, K.O.L., Stendahl, J., Lundblad, M. & Karltun, E. (2021). Predicting the spatial distribution of soil organic carbon stock in Swedish forests using a group of covariates and site-specific data. *SOIL*, 7 (2), 377–398. https://doi.org/10.5194/soil-7-377-2021

Institutionen för skoglig resurshushållning och Institutionen för mark och miljö (SLU) (2022). Fältinstruktion Riksinventeringen av skog.

Intergovernmental Panel On Climate Change (2023). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1. ed Cambridge University Press. https://doi.org/10.1017/9781009157896

Jenny, H. (1941). Factors of Soil Formation: *Soil Science*, 52 (5), 415. https://doi.org/10.1097/00010694-194111000-00009

Kuhn & Max (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28 (5), 1–26. https://doi.org/10.18637/jss.v028.i05

Manning, P., De Vries, F.T., Tallowin, J.R.B., Smith, R., Mortimer, S.R., Pilgrim, E.S., Harrison, K.A., Wright, D.G., Quirk, H., Benson, J., Shipley, B., Cornelissen, J.H.C., Kattge, J., Bönisch, G., Wirth, C. & Bardgett, R.D. (2015). Simple measures of climate, soil properties and plant traits predict national-scale grassland soil carbon stocks. Wilsey, B. (ed.) (Wilsey, B., ed.) *Journal of Applied Ecology*, 52 (5), 1188–1196. https://doi.org/10.1111/1365-2664.12478

McBratney, A.B., Mendonça Santos, M.L. & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117 (1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Miller, B.A., Koszinski, S., Wehrhan, M. & Sommer, M. (2015). Impact of multi-scale predictor selection for modeling soil properties. *Geoderma*, 239–240, 97–106. https://doi.org/10.1016/j.geoderma.2014.09.018

Moni, C., Chabbi, A., Nunan, N., Rumpel, C. & Chenu, C. (2010). Spatial dependance of organic carbon–metal relationships. *Geoderma*, 158 (3–4), 120–127. https://doi.org/10.1016/j.geoderma.2010.04.014

Oostra, S., Majdi, H. & Olsson, M. (2006). Impact of tree species on soil carbon stocks and soil acidity in southern Sweden. *Scandinavian Journal of Forest Research*, 21 (5), 364–371. https://doi.org/10.1080/02827580600950172

Osman, K.T. (2013). *Forest Soils: Properties and Management*. Springer International Publishing. https://doi.org/10.1007/978-3-319-02541-4

Ranneby, B., Cruse, T., Hagglund, B., Jonasson, H. & Sward, J. (1987). Designing a new national forest survey for Sweden.

Rasmussen, C., Heckman, K., Wieder, W.R., Keiluweit, M., Lawrence, C.R., Berhe, A.A., Blankinship, J.C., Crow, S.E., Druhan, J.L., Hicks Pries, C.E., Marin-Spiotta, E., Plante, A.F., Schädel, C., Schimel, J.P., Sierra, C.A., Thompson, A. & Wagai, R. (2018). Beyond clay: towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry*, 137 (3), 297–306. https://doi.org/10.1007/s10533-018-0424-3

Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S. & Trumbore, S.E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, 478 (7367), 49–56. https://doi.org/10.1038/nature10386

Seibert, J., Stendahl, J. & Sørensen, R. (2007). Topographical influences on soil properties in boreal forests. *Geoderma*, 141 (1–2), 139–148. https://doi.org/10.1016/j.geoderma.2007.05.013

Shapiro, H.M. (2003). *Practical flow cytometry*. 4th ed. Wiley-Liss.

Spohn, M. & Stendahl, J. (2022). Carbon, nitrogen, and phosphorus stoichiometry of organic matter in Swedish forest soils and its relationship with climate, tree species, and soil texture. *Biogeosciences*, 19 (8), 2171–2186. https://doi.org/10.5194/bg-19-2171-2022

Spohn, M. & Stendahl, J. (2024). Soil carbon and nitrogen contents in forest soils are related to soil texture in interaction with pH and metal cations. *Geoderma*, 441, 116746. https://doi.org/10.1016/j.geoderma.2023.116746

Stendahl, J., Berg, B. & Lindahl, B.D. (2017). Manganese availability is negatively associated with carbon storage in northern coniferous forest humus layers. *Scientific Reports*, 7 (1), 15487. https://doi.org/10.1038/s41598-017-15801-y

Vesterdal, L., Clarke, N., Sigurdsson, B.D. & Gundersen, P. (2013). Do tree species influence soil carbon stocks in temperate and boreal forests? *Forest Ecology and Management*, 309, 4–18. https://doi.org/10.1016/j.foreco.2013.01.017

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., Von Lützow, M., Marin-Spiotta, E., Van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J. & Kögel-Knabner, I. (2019). Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. *Geoderma*, 333, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G. & Li, D.-C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological Indicators*, 60, 870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

Zhou, Y., Biswas, A., Ma, Z., Lu, Y., Chen, Q. & Shi, Z. (2016). Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain. *Geoderma*, 271, 71–79. https://doi.org/10.1016/j.geoderma.2016.02.006

Popular science summary

Soils around the world contain vast amounts of carbon, more than the atmosphere and vegetation combined. Most of this carbon stored in soils, especially in Sweden is soil organic carbon (SOC). The storage of SOC is influenced by environmental conditions, like the climate, vegetation, topography and other soil properties. Relationships between these environmental variables can be used in statistical models to predict SOC. It is recognised that these relationships change with scale, meaning that depending on what area you look at, different sets of variables are well suited to set up models for SOC prediction. In this thesis I use Swedish soil data to test the following hypotheses: Climatic variables are more important when predicting SOC for larger areas, texture is similarly important for different areas and topography and chemistry are more important for smaller areas.

Data on soils and vegetation from the Swedish forest soil inventory was combined with topographic and climatic data to serve as the input for statistical models. The soil samples in this inventory are taken from different depths in the soil, so that five model are created for every area: One for the carbon concentration and one for the stocks in the organic layer and three for different depths in the mineral layer. Additionally, I divided Sweden into four regions and each of these regions into 4 subregions, so that models were trained for 21 areas each. The type of model is called random forest (RF) with recursive feature elimination. It works well with all types of data and choses the variables that are best suited itself, but its output cannot be nicely expressed in a simple function, because it basically is the combined output of many smaller models. Although the importance of variables can be assessed by it, I additionally used linear regressions which gave information about the types of relationships between the most important predictors in RF models and SOC.

Across the three scales, all models gave very high importances to chemical variables, with Mn being most important for organic layer carbon stocks, pH for organic laser concentrations and Al for mineral layer concentrations. The climatic variables were more important at the national scale, as I hypothesized. Texture was mostly unimportant in all models, making it hard to satisfyingly draw conclusions about the second hypothesis. Topographic and chemical variables didn't exhibit a clear pattern like climate. Their highest importances occurred at different scales in different models.

Acknowledgements

First and foremost, I want to thank my supervisor Marie Spohn and my co-supervisor Klaus Katzensteiner who helped me with their inspiration and feedback throughout the course of this whole thesis.

I also want to thank my examiner Björn Lindahl and my student opponent Ronja Stubbe, your feedback and questions during the oral presentation were a valuable addition to my work.

Lastly, I can’t forget to mention Christofer Engberg Hydén and Johan Stendahl. Your explanations of and help with the SFSI data made it possible to actually realise the analysis that Marie and Klaus helped me to conceptualise.

Appendix

Table A 1. Definition of texture classes

Texture class	Grain size
Cobble	20-200 mm
Gravel	2-20 mm
Coarse sand	0.6-2 mm
Medium sand	0.2-0.6 mm
Fine sand	0.06-0.2 mm
Coarse silt	0.02-0.06 mm
Fine silt	0.002-0.02 mm
Clay	<0.002 mm

Table A 2. Definition of tree species classes

Tree species class	Definition	corresponding values of “tradblandning” in the SFSI
Pine	> 70% Pine	1
Spruce	> 70% Spruce	2
Mixed conifers	> 70% Pine + Spruce	3
Deciduous	> 70% deciduous	4, 5, 10, 11
Mixed	< 70% conifers **and** < 70% deciduous	6, 7, 8, 9

Table A 3. Cross-validation results of the RF models for carbon concentration in the organic layer

		10- fold cross- validation		
Area	number of predictors used	R^2	RMSE (g C*kg^{-1} d.w.)	MAE (g C*kg^{-1} d.w.)
National	6	0.71 ± 0.03	49.3 ± 2.26	38.9 ± 1.66
Region 1	6	0.74 ± 0.05	50.4 ± 4.13	40.6 ± 3.67
Region 2	6	0.69 ± 0.08	53.2 ± 4.68	42.7 ± 2.90
Region 3	6	0.64 ± 0.06	54.4 ± 5.21	42.4 ± 4.32
Region 3	6	0.52 ± 0.11	53.6 ± 6.26	40.6 ± 4.22
Subregion 1A	4	0.70 ± 0.14	56.5 ± 10.50	45.7 ± 6.97
Subregion 1B	3	0.64 ± 0.21	54.0 ± 9.67	45.4 ± 8.27
Subregion 1C	6	0.57 ± 0.10	57.0 ± 7.19	47.2 ± 6.33
Subregion 1D	4	0.70 ± 0.12	53.2 ± 8.38	42.5 ± 6.48
Subregion 2A	3	0.63 ± 0.15	53.3 ± 8.89	42.8 ± 7.27
Subregion 2B	4	0.65 ± 0.14	59.0 ± 10.62	47.0 ± 7.79
Subregion 2C	4	0.60 ± 0.19	57.3 ± 10.22	46.8 ± 7.61
Subregion 2D	2	0.57 ± 0.19	58.0 ± 12.04	45.4 ± 9.32
Subregion 3A	3	0.61 ± 0.13	57.7 ± 12.24	46.0 ± 8.07
Subregion 3B	3	0.68 ± 0.17	57.4 ± 16.30	46.2 ± 8.60
Subregion 3C	3	0.55 ± 0.13	45.6 ± 6.96	35.7 ± 5.47
Subregion 3D	2	0.55 ± 0.17	58.3 ± 13.92	46.6 ± 11.49
Subregion 4A	4	0.50 ± 0.11	58.7 ± 12.51	47.2 ± 10.15
Subregion 4B	4	0.41 ± 0.28	56.2 ± 13.48	43.4 ± 11.79
Subregion 4C	6	0.41 ± 0.19	60.2 ± 13.33	47.2 ± 9.79
Subregion 4D	3	0.60 ± 0.20	49.7 ± 10.12	38.4 ± 7.37

Table A 4. Cross-validation results of the RF models for carbon stock in the organic layer

		10- fold cross- validation		
Area	number of predictors used	R^2	RMSE (t C*ha^{-1})	MAE (t C*ha)
National	16	0.44 ± 0.07	22.4 ± 2.82	14.4 ± 0.78
Region 1	10	0.41 ± 0.08	29.2 ± 5.62	19.3 ± 2.28
Region 2	14	0.39 ± 0.15	22.8 ± 3.38	15.7 ± 1.49
Region 3	6	0.43 ± 0.19	17.1 ± 5.10	11.4 ± 1.88
Region 3	16	0.36 ± 0.14	17.1 ± 3.71	11.0 ± 1.43
Subregion 1A	6	0.53 ± 0.15	21.4 ± 6.80	14.9 ± 3.77
Subregion 1B	6	0.45 ± 0.24	33.1 ± 16.29	21.7 ± 7.70
Subregion 1C	5	0.36 ± 0.17	29.1 ± 8.92	21.1 ± 4.70
Subregion 1D	7	0.40 ± 0.27	27.9 ± 7.45	20.5 ± 4.03
Subregion 2A	3	0.34 ± 0.27	22.7 ± 7.85	15.8 ± 4.06
Subregion 2B	14	0.42 ± 0.26	27.8 ± 8.68	19.7 ± 4.54
Subregion 2C	14	0.26 ± 0.18	21.6 ± 6.77	16.2 ± 3.46
Subregion 2D	11	0.28 ± 0.20	18.7 ± 6.94	13.9 ± 3.95
Subregion 3A	4	0.45 ± 0.23	19.6 ± 7.90	13.6 ± 3.89
Subregion 3B	7	0.37 ± 0.24	15.5 ± 5.45	11.4 ± 3.32
Subregion 3C	15	0.23 ± 0.19	15.6 ± 5.83	11.2 ± 2.88
Subregion 3D	3	0.61 ± 0.24	15.5 ± 4.18	10.7 ± 2.25
Subregion 4A	15	0.50 ± 0.27	16.5 ± 6.39	10.8 ± 2.07
Subregion 4B	15	0.39 ± 0.19	21.0 ± 11.65	14.7 ± 5.35
Subregion 4C	11	0.18 ± 0.12	12.5 ± 3.88	9.1 ± 2.21
Subregion 4D	10	0.24 ± 0.14	14.5 ± 4.99	10.3 ± 2.29

Table A 5. Cross-validation results of RF models for carbon concentration in the mineral layer (0-10cm)

		10- fold cross- validation		
Area	number of predictors used	R^2	RMSE (g C*kg^{-1} d.w.)	MAE (g C*kg^{-1} d.w.)
National	9	0.51 ± 0.14	18.4 ± 6.65	9.3 ± 0.58
Region 1	15	0.50 ± 0.13	23.0 ± 12.22	12.6 ± 2.52
Region 2	9	0.51 ± 0.19	21.6 ± 12.17	11.2 ± 1.33
Region 3	9	0.38 ± 0.15	11.7 ± 2.56	8.3 ± 1.24
Region 3	13	0.43 ± 0.15	12.3 ± 4.30	7.8 ± 1.62
Subregion 1A	9	0.41 ± 0.19	25.2 ± 27.83	15.0 ± 8.54
Subregion 1B	2	0.52 ± 0.23	21.4 ± 14.92	13.9 ± 5.96
Subregion 1C	2	0.67 ± 0.23	20.5 ± 15.47	11.3 ± 4.99
Subregion 1D	2	0.60 ± 0.19	13.8 ± 2.47	9.7 ± 1.53
Subregion 2A	10	0.35 ± 0.27	31.1 ± 30.99	17.1 ± 9.82
Subregion 2B	11	0.55 ± 0.19	23.3 ± 9.99	15.7 ± 4.26
Subregion 2C	3	0.68 ± 0.18	14.2 ± 6.19	9.3 ± 2.80
Subregion 2D	9	0.63 ± 0.18	9.7 ± 2.94	6.6 ± 0.92
Subregion 3A	2	0.38 ± 0.2	15.3 ± 5.36	10.5 ± 2.84
Subregion 3B	5	0.29 ± 0.24	10.5 ± 5.39	7.9 ± 3.04
Subregion 3C	9	0.42 ± 0.21	11 ± 3.29	8.1 ± 2.30
Subregion 3D	5	0.32 ± 0.28	9.8 ± 2.26	7.6 ± 1.34
Subregion 4A	13	0.40 ± 0.20	13.7 ± 8.98	8.3 ± 2.77
Subregion 4B	5	0.45 ± 0.18	9.4 ± 1.97	6.9 ± 1.05
Subregion 4C	15	0.40 ± 0.33	15.4 ± 5.94	10.1 ± 1.91
Subregion 4D	9	0.41 ± 0.23	9.3 ± 3.28	6.3 ± 1.56

Table A 6. Cross-validation results of RF models for carbon concentration in the mineral layer (10-20cm)

		10- fold cross- validation		
Area	number of predictors used	R^2	RMSE (g C*kg^{-1} d.w.)	MAE (g C*kg^{-1} d.w.)
National	9	0.48 ± 0.10	15.2 ± 5.63	8.3 ± 0.64
Region 1	7	0.46 ± 0.19	20.2 ± 14.07	10.3 ± 2.86
Region 2	6	0.50 ± 0.08	14.7 ± 2.07	9.5 ± 0.99
Region 3	6	0.36 ± 0.11	11.4 ± 1.69	8.2 ± 0.85
Region 3	14	0.38 ± 0.11	10.5 ± 4.07	6.3 ± 0.78
Subregion 1A	13	0.40 ± 0.26	22.4 ± 31.19	13.5 ± 11.79
Subregion 1B	14	0.44 ± 0.22	24.9 ± 27.09	14.3 ± 8.18
Subregion 1C	9	0.55 ± 0.12	12.4 ± 4.70	8.8 ± 1.99
Subregion 1D	7	0.60 ± 0.21	13.8 ± 4.05	9.7 ± 2.30
Subregion 2A	3	0.48 ± 0.25	15.2 ± 9.02	10.5 ± 4.44
Subregion 2B	11	0.57 ± 0.24	13.1 ± 6.60	9.9 ± 4.18
Subregion 2C	10	0.51 ± 0.17	14.3 ± 5.24	10.8 ± 2.02
Subregion 2D	4	0.63 ± 0.24	13.2 ± 7.03	8.3 ± 3.23
Subregion 3A	7	0.53 ± 0.16	12.0 ± 3.13	8.7 ± 1.94
Subregion 3B	4	0.44 ± 0.16	8.6 ± 1.66	6.7 ± 1.13
Subregion 3C	5	0.41 ± 0.25	10.3 ± 3.10	7.7 ± 1.95
Subregion 3D	12	0.25 ± 0.25	12.5 ± 5.01	9.0 ± 2.40
Subregion 4A	9	0.51 ± 0.24	6.8 ± 2.77	4.7 ± 1.36
Subregion 4B	10	0.51 ± 0.18	7.5 ± 2.28	5.9 ± 1.46
Subregion 4C	2	0.42 ± 0.29	12.8 ± 6.73	8.5 ± 3.07
Subregion 4D	8	0.42 ± 0.26	7.3 ± 3.05	5.2 ± 1.45

Table A 7. Cross-validation results of RF models for carbon concentration in the mineral layer (55-65cm)

		10- fold cross- validation		
Area	number of predictors used	R^2	RMSE (g C*kg^{-1} d.w.)	MAE (g C*kg^{-1} d.w.)
National	15	0.51 ± 0.07	5.9 ± 2.38	3.1 ± 0.51
Region 1	6	0.58 ± 0.16	7.0 ± 3.30	3.9 ± 1.11
Region 2	6	0.50 ± 0.26	5.4 ± 2.08	3.2 ± 0.54
Region 3	2	0.42 ± 0.25	4.3 ± 2.05	2.6 ± 0.88
Region 3	12	0.55 ± 0.10	5.1 ± 4.29	2.6 ± 0.85
Subregion 1A	3	0.66 ± 0.30	6.8 ± 5.40	4.5 ± 2.75
Subregion 1B	2	0.42 ± 0.31	8.9 ± 6.47	5.3 ± 2.91
Subregion 1C	1	0.79 ± 0.20	5.4 ± 1.70	4.0 ± 1.16
Subregion 1D	13	0.59 ± 0.30	6.5 ± 4.21	4.2 ± 1.72
Subregion 2A	6	0.49 ± 0.33	4.5 ± 1.54	3.3 ± 0.84
Subregion 2B	7	0.41 ± 0.33	5.3 ± 2.80	3.6 ± 1.44
Subregion 2C	2	0.33 ± 0.31	4.1 ± 0.96	3.0 ± 0.59
Subregion 2D	3	0.81 ± 0.09	5.5 ± 3.25	3.3 ± 1.77
Subregion 3A	1	0.46 ± 0.35	4.3 ± 2.06	2.9 ± 0.87
Subregion 3B	14	0.30 ± 0.22	4.3 ± 2.87	2.9 ± 1.10
Subregion 3C	13	0.40 ± 0.22	3.4 ± 2.62	2.3 ± 1.09
Subregion 3D	7	0.51 ± 0.31	3.8 ± 2.31	2.4 ± 0.95
Subregion 4A	10	0.56 ± 0.34	2.7 ± 1.47	1.7 ± 0.57
Subregion 4B	2	0.41 ± 0.31	3.8 ± 2.00	2.4 ± 0.88
Subregion 4C	15	0.48 ± 0.31	8.9 ± 9.12	5.6 ± 3.18
Subregion 4D	3	0.58 ± 0.17	2.0 ± 1.06	1.3 ± 0.40

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file. If you are more than one author, the checked box will be applied to all authors. You will find a link to SLU's publishing agreement here:

- https://libanswers.slu.se/en/faq/228318.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.