Assessing top- and subsoil organic carbon stocks of Low-Input High-Diversity systems using soil and vegetation characteristics

Sam Ottoy[a*], Koenraad Van Meerbeek[a], Anicet Sindayihebura[a,b], Martin Hermy[a] and Jos Van Orshoven[a]

[a] Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E box 2411, 3001 Leuven, Belgium

[b] Department of Earth Sciences, Burundi University, P.O. Box 1550 Bujumbura, Burundi

[*]Corresponding author: e-mail: sam.ottoy@kuleuven.be , tel. +321632974

Abstract

The soil organic carbon (SOC) stock is an important indicator in ecosystem service assessments. Even though a considerable fraction of the total stock is stored in the subsoil, current assessments often consider the topsoil only. Furthermore, mapping efforts are hampered by the limited spatial density of these topsoil measurements. The aim of this study was to assess the SOC stock in the upper 100 cm of soil in 30,556 ha of Low-Input High-Diversity systems, such as nature reserves, in Flanders (Belgium) and compare this estimate with the stock found in the topsoil (upper 15 cm). To this end, we combined depth extrapolation of 139 measurements limited to the topsoil with four digital soil mapping techniques: multiple linear regression, boosted regression trees, artificial neural networks and least-squares support vector machines. Particular attention was given to vegetation characteristics as predictors. For both the stock in the upper 15 cm and 100 cm, a boosted regression trees approach was most informative as it resulted in the lowest cross-validation errors and provided insights in the relative importance of predictors. The predictors of the stock in the upper 100 cm were soil type, groundwater level, clay fraction and community weighted mean (CWM) and variance (CWV) of plant height. These predictors, together with the CWM of specific leaf area, aboveground biomass production, CWV and CWM of rooting depth, terrain slope, CWM of mycorrhizal associations and species diversity also explained the topsoil stock. Our total stock estimates show that focusing on the topsoil (1.63 Tg OC) only considers 36% of the stock in the upper 100 cm (4.53 Tg OC). Given the magnitude of subsoil OC and its dependency on typical ecosystem characteristics, it should not be neglected in regional ecosystem service assessments.

Keywords: ecosystem services, depth extrapolation, digital soil mapping, regional assessment

1. Introduction

The effective and potential level of services that ecosystems provide is increasingly inspiring land use planning (Goldstein et al., 2012; Broekx et al., 2013; Galati et al., 2016). For such mapping and assessments, the soil organic carbon (SOC) stock is an important indicator (Maes et al., 2016). Whereas a considerable fraction of the total SOC stock is known to be stored in the subsoil (Batjes, 1996; Jobbágy and Jackson, 2000) and should not be neglected in an ecosystem service context (Jandl et al., 2014), routinely available measurement data and hence stock estimates are often limited to the topsoil, e.g. see Minasny et al. (2013). To include the subsoil stock, vertical extrapolation of the topsoil measurement is often necessary. However, the commonly used exponential decline function is not capable to accurately model the stock in soil types characterised by SOC-rich subsurface horizons, such as spodic and peat horizons (Sleutel et al., 2003; Aldana Jague et al., 2016). To take these 'anomalies' into account, we have developed an exponential change decline function in earlier research, assuming that not the OC content but rather the difference between the target (2009-2011) and the historical (1947-1974) reference topsoil measurement value declines exponentially with depth (Ottoy et al., 2016).

Another shortcoming of routine soil sampling is its limited and heterogeneous spatial density which is a weak basis for regional SOC stock assessments (Carré et al., 2007; Ottoy et al., 2015). In many cases, soil profiles have been sampled for the major land units (LUs), but are lacking for the many minor LUs. To cope with this lack of data, digital soil mapping or 'SCORPAN' approaches have been proposed which exploit the covariance of a soil variable (s) with predictors representing Jenny's (1941) soil forming factors (climate (c), organisms (o), topography (r), parent material (p), age (a)) extended with geographic position (n) (McBratney et al., 2003). For the case of SOC stock modelling, numerous techniques have been proposed, ranging from multiple linear regression (Meersmans et al., 2008) to more

58 recently developed machine-learning methods like Boosted Regression Trees, Artificial
59 Neural Networks and Support Vector Machines (Martin et al., 2014; Were et al., 2015;
60 Taghizadeh-Mehrjardi et al., 2016).

61 These SCORPAN-methods do not only contribute to more reliable SOC stock assessments,
62 but also provide insights in the relative importance of the candidate predictors of the SOC
63 stock and hence in the functioning of the soil system. At the biome level, climate variables
64 such as mean annual precipitation and temperature and their interaction with vegetation are
65 important controls of the SOC storage capacity of soils (Jobbágy and Jackson, 2000;
66 O'Rourke et al., 2015). At the regional scale, physical and chemical soil variables like the
67 texture fraction, moisture content, pH and soil profile development are typically identified as
68 variables explaining SOC storage (Meersmans et al., 2008; Wiesmeier et al., 2011; Were et
69 al., 2015). In addition, land use intensity including manure application was found to explain
70 regional variations in the SOC stock (van Wesemael et al., 2010; Parras-Alcántara et al.,
71 2015a; Manning et al., 2015). Another important representative of SCORPAN's 'organism'
72 factor is the vegetation, which can contribute to controlling both soil carbon input and loss
73 (Chapin, 2003; De Deyn et al., 2008) and hence the resulting SOC stock (Grigulis et al., 2013;
74 Manning et al., 2015). Similarly, diversity of plant species (Tilman et al., 2006) and
75 functional groups (Steinbeiss et al., 2008) were found to affect SOC storage.

76 The aim of this study was to assess the SOC stock in the upper 100 cm of soil of Low-Input
77 High-Diversity (LIHD) systems in Flanders (Belgium) using available topsoil (upper 15 cm)
78 measurements. Managed nature reserves are typical LIHD systems characterised by low
79 levels of inputs (e.g. manure application) and high species diversity. Recently, these systems
80 have come into the picture due to their high potential to mitigate climate change through the
81 production of bioenergy (Tilman et al., 2006; Van Meerbeek et al., 2016), but their SOC
82 storage capacity remained relatively underexplored. To include the subsoil in our regional

assessment and spatially densify the available measurements, we combined depth extrapolation of topsoil measurements with digital soil mapping. Additionally, this estimate was compared with the stock found in the topsoil only. Through this process, we aimed at identifying the main predictors of top- and subsoil stocks, considering various soil properties, plant functional traits and trait diversity measures.

2. Material and Methods

2.1 Study area

We assessed the SOC stocks in the upper 15 and 100 cm of mineral soil of LIHD systems in the region of Flanders, N. Belgium. This region of 13,522 km² is characterised by a maritime temperate climate, with a mean annual temperature of 9.8 – 10.5°C (mean minimum of 6.7°C and maximum of 13.8°C) and a mean annual precipitation of 733 – 832 mm (Peel et al., 2007). A pronounced gradient of decreasing sand and increasing silt fractions is present from north to south.

2.2 Soil and environmental data

2.2.1 Soil and vegetation sampling

From 2009 to 2011, 139 sites in nature reserves across different ecoregions were visited and sampled following the procedure described in Van Meerbeek et al. (2014). At each site, a plot of 10 x 10 m was positioned in a homogeneous vegetation patch. Therein three subplots of 0.5 x 0.5 m were randomly selected, forming a composite sample. In each subplot, the topsoil was sampled to a depth of 15 cm. The SOC content (%) was determined using a modified version of the Walkley and Black (1934) method. A correction factor of 1.14 was applied to account for incomplete oxidation (Lettens et al., 2005). Also the aboveground biomass was harvested in each subplot. The SOC content and the dry weight of the harvested biomass were averaged over the three subplots to obtain one value per plot. Furthermore, the cover (%) of each plant species was visually estimated for the subplots.

2.2.2 Plant functional traits and trait diversity

Trait-based diversity indices were chosen to represent the two main classes of effects of biodiversity on ecosystem processes, namely the complementarity effect and the selection effect (Loreau and Hector, 2001). First, the community weighted mean value (CWM) was calculated for each trait in each plot. Weighting was done by the relative abundance (cover,

%) of the plant species. CWM corresponds to the selection effect in which dominance by species with particular traits affects ecosystem processes (Loreau and Hector, 2001). Next, the functional dispersion (FDis) index is the weighted mean distance of the species to their centroid in a multivariate trait space (Laliberté and Legendre, 2010), and is an indicator of variability of the trait values in a community. The third trait-based index considered was the community weighted variance (CWV) (Sonnier et al., 2010). It is the weighted variance of trait values with respect to the CWM. Both FDis and CWV are used as proxies for the complementarity effects, in which niche complementarity leads to higher resource use and ecosystem functioning (Loreau and Hector, 2001).

To compute the three selected trait-based diversity indices, we selected twelve functional traits based on their assumed relevance for belowground carbon sequestration (De Deyn et al., 2008; Pérez-Harguindeguy et al., 2013) and extracted corresponding trait values from the TRY database (Kattge et al., 2011). Because of the high percentage of missing values in the trait matrix (41%), we estimated the missing values using a phylogeny with the 'Rphylopars' package of R-software (Goolsby et al., 2016). This package can perform missing data imputation on an estimated evolutionary model, in our case a brownian motion model. The phylogenetic tree used in this analysis was constructed from the dated phylogeny for higher plants of Western Europe (Durka and Michalski, 2012) with the 'Picante' package of R-software (Kembel et al., 2010). From the pool of twelve traits, we selected five traits that optimally represented the trait space by means of a PCA. CWM and CWV were calculated for each of the traits, whereas FDis was derived based on all five traits. FDis and CWM were calculated using the *dbFD* function available in the 'FD' package of R-software (Laliberté et al., 2014) and CWV according to Sonnier et al. (2010).

2.2.3 Additional environmental data

The soil map unit and corresponding Reference Soil Group (RSG) according to the third edition of the World Reference Base (WRB) classification system (IUSS Working Group WRB, 2014) of each of the 139 plots were derived from the soil map of the Flemish region (source scale 1:20,000) converted to WRB classification (Dondeyne et al., 2014). Apart from this soil map, Belgium's national soil survey (1947 – 1974) resulted in an extensive soil profile dataset (Van Orshoven et al., 1993). The data for the region of Flanders are gathered in the Aardewerk-Vlaanderen-2010 database comprising the location and descriptive data of 7020 soil profiles and descriptive and analytical data of 42,529 associated horizons. Using this database, statistical soil profiles were computed at a reference level and at four hierarchical levels of generalisation (Ottoy et al., 2015). For each soil map unit in which the 139 plots were positioned, the most detailed statistical profile was selected in the database. For each horizon in the statistical profile, SOC contents and soil granulometrical fractions were retrieved. By proportional weighting according to horizon thickness, mean values of the clay, silt and sand fraction for the upper 100 cm were calculated. These texture fractions were used for digital soil mapping, while the historical SOC contents of each horizon were used for depth extrapolation of the topsoil measurements (Figure 1). In line with Sleutel et al. (2003), we assumed that the soil texture fractions have not changed between 1947-1974 and 2009-2011.

Data of the highest (reduction horizon) and lowest (oxidation horizon) groundwater level were retrieved from the ECOPLAN database, with a spatial resolution of 5 x 5 m (Staes, 2016). The slope of the terrain was derived from the Digital Elevation Model of Flanders, resolution 5 x 5 m (AGIV, 2006).

163 2.3 Depth extrapolation and SOC stock calculation

164 As the measured 2009-2011 SOC contents were limited to the upper 15 cm of soil (C_0), depth
165 extrapolation was necessary to derive estimates for the upper 100 cm (Figure 1). Therefore,
166 the exponential change decline function presented by Ottoy et al. (2016) was implemented to
167 estimate the SOC content (%) at depth z (cm):

168 $$z \leq 15\ cm : C(z) = C_0 \qquad \text{Eq. (1)}$$

169 $$z > 15\ cm : C(z) = C_{hist,z} + \left(C_0 - C_{hist,0}\right)e^{-k_{ECD}z} \qquad \text{Eq. (2)}$$

170 where $C_{hist,z}$ and $C_{hist,0}$ are the SOC contents of the statistical (historical) soil profile at depth
171 z and at the surface, respectively. In Eq. (2), z represents the mean depth below the topsoil.
172 Parameter k_{ECD} describes the shape of the exponential curve; a larger parameter value
173 corresponds to a more pronounced decrease with depth of the difference in SOC content
174 recorded for the topsoil. The value of k_{ECD} was estimated based on the measurements for
175 grassland soil used by Ottoy et al. (2016) with the 'nlstools' package of R-software (Baty et
176 al., 2015). For the sandy agricultural regions, k_{ECD} was found to be 0.10, whereas a value of
177 0.35 was implemented for the regions characterised by finer soil textures.

178 After obtaining the SOC content of each horizon (C_i), stocks were calculated using Eq. (3).
179 Bulk density was not determined during soil sampling, but was estimated for each horizon by
180 the pedotransfer function of Rawls (1983), Eq. (4).

181 $$\text{SOCS} = \sum_{i=1}^{n}(C_i \times \text{BD}_i \times d_i), \qquad \text{Eq. (3)}$$

182 $$\text{BD}_i = \frac{100}{\frac{\text{SOM}_i}{\text{BD}_{\text{SOM}}} + \frac{100-\text{SOM}_i}{\text{BD}_{\text{MF}}}}, \qquad \text{Eq. (4)}$$

183 where SOCS (kg OC m^{-2}) is the SOC stock in the upper 15 cm resp. 100 cm, n is the total
184 number of horizons, BD_i (kg m^{-3}) is the bulk density of horizon i, d_i (m) the thickness of
185 horizon i, SOM_i (%) the soil organic matter content of horizon i (SOC content × 2, analogous
186 to Lettens et al. (2004)), BD_{SOM} is the bulk density of soil organic matter (0.224×10^3 kg m^{-3})

187 and BD_{MF} is the bulk density of the mineral fraction, reported by Lettens et al. (2004). The
188 bulk density of peat soil was set to 0.31×10^3 kg m^{-3} (Batjes, 1996).

189 2.4 Digital soil mapping: model training

190 2.4.1 Exploratory data analysis

191 To detect collinearity between the predictors, correlations (r) were calculated using R 3.0.2
192 software (R Core Team, 2013). For two continuous variables, Spearman's rank correlation (r_s)
193 was calculated, whereas for a continuous and a binary variable, the point-biserial correlation
194 (r_p) was determined. One of each pair of highly correlated variables ($|r| > 0.70$) was omitted
195 from further modelling.

196 2.4.2 Digital soil mapping

197 Four digital soil mapping techniques, one linear and three non-linear, have been applied to
198 model the SOC stock in the upper 15 cm and 100 cm. To assess the performance of each
199 modelling technique, the coefficient of determination (R^2), adjusted R^2 (R^2_{adj}), root mean
200 squared error (RMSE) and relative RMSE (rRMSE) were calculated using 10-fold cross-
201 validation.

202 *Multiple Linear Regression (MLR)*

203 In ecological modelling, MLR models are often selected for their user friendliness and
204 straightforward interpretability (Aertsen et al., 2010; Van Meerbeek et al., 2014). Model
205 selection was based on multimodel inference using the 'MuMIn' package of R-software
206 (Barton, 2016). The global model contained all predictor variables listed in Table 1, except
207 those omitted after the collinearity analysis. The *dredge* function generates a set of
208 submodels, nested in the global model, and ranks these according to the Second-order Akaike
209 Information Criterion (AICc). The so-called best model is the one with the lowest AICc value.
210 Next, model averaged coefficients were calculated using the submodels with $\Delta AICc < 2$
211 compared to the best model, as these models are considered to explain variation in the data

substantially (Burnham and Anderson, 2002). The relative variable importance is reflected by the sum of the Akaike weights over all submodels including the explanatory variable.

Boosted Regression Trees (BRT)

The technique of BRT aims to improve (boost) the performance of a single model (tree) by fitting many models (trees) and combining them for prediction (Elith et al., 2008). This method has several advantages, including the capacity of combining predictor variables of different data types and accommodating missing data. BRT were developed using the 'dismo' package of R-software (Hijmans et al., 2016). Fitting a BRT requires specification of three meta-parameters: (i) the learning rate which determines the contribution of each tree to the growing model, (ii) the tree complexity which controls the order of interactions that can be fitted, and (iii) the number of trees required for optimal prediction, which is determined by (i) and (ii). The optimal values of the learning rate and tree complexity were derived by performing a grid search using the *train* function of R's 'caret' package (Kuhn, 2008). Models were fitted with the *gbm.step* function and simplified by reducing the number of predictors using the *gbm.simplify* function. The effect of each variable on the SOC stock can be visualised by partial dependence plots. As these plots account for the average effects of all other model variables, caution is required when interpreting them in the case of strongly correlated variables. Furthermore, the relative importance of each variable can be estimated, taking into account the number of times a variable is selected for splitting and the improvement of the model's performance as a result of each split. The relative importance of each variable is scaled so that the sum adds to 100%.

Artificial Neural Networks: Multi-Layer Perceptron (ANN)

Inspired by the architecture of the human brain, Artificial Neural Networks (ANN) are interconnected structures of simple, non-linear processing elements called nodes or neurons (Haykin, 1998). A very popular class of ANN are the multi-layer perceptron networks, in

which the neurons are arranged in layers: an input layer ingesting the value of the predictor variables, one or more hidden layers and an output layer producing the value of the response variable. Each neuron is connected to the neurons in the next layer whereby the strength of this connection is represented by an interconnection weight. By training the ANN with reference samples, the initial interconnection weights of the network are iteratively adjusted, hereby minimising the prediction error. During this process, three meta-parameters need to be specified: (i) the number of neurons in the hidden layer, (ii) the learning algorithm, and (iii) the number of training iterations. The number of neurons and the learning algorithm were determined in a grid search by testing multiple combinations and selecting the one with the lowest prediction error. The tested learning algorithms were gradient descent backpropagation, Quasi-Newton learning, Levenberg-Marquardt learning and conjugate gradient learning. The number of training iterations was determined using an early stopping procedure. This avoids overfitting as the training procedure is stopped as soon as the error on a separate validation set starts to increase. Overfitting was also avoided by specifying a regularisation parameter, which describes the trade-off between model complexity and flexibility. ANN were developed using the 'Neural Network' toolbox of Matlab-software (Beale et al., 2012).

Least-Squares Support Vector Machines (LS-SVM)

Support vector machines use kernel functions to map the input variables in a high-dimensional space, in which a linear regression model is constructed. For Least-squares Support Vector Machines, inequality constraints are replaced by equality constraints and a squared loss function is minimised (Suykens et al., 2002). During the training phase, two meta-parameters need to be specified: (i) the kernel type and associated kernel parameter, describing the non-linear mapping function to the feature space, and (ii) the regularisation parameter. The optimal values for the meta-parameters were determined in a cross-validation

approach within each training dataset using the *tunelssvm* function in the Statistical library for least squares support vector machines (STATLSSVM) toolbox of Matlab-software (De Brabanter et al., 2013).

2.5 SOC stock upscaling

The location and spatial extent of the LIHD systems in Flanders were retrieved from the ECOPLAN land cover and land use geodatasets with a resolution of 5 x 5 m (Vrebos, 2015). The land cover dataset is based on the Biological valuation map (De Saeger et al., 2010) and the agricultural parcels map (ALV, 2013) and distinguishes the following land cover classes: herb-rich vegetation, nutrient-poor grassland, nutrient-rich grassland, dry heath, wet heath, marsh, reed marsh and salt marsh. The ECOPLAN land use dataset describes the type and intensity of land management and was used to differentiate low-input from high-input grasslands. Only grassland with an extensive management regime was retained.

Land units were defined by topological overlay of these land cover and land use datasets with the soil map converted to WRB classification (Figure 1). This resulted in 129,794 polygons belonging to 6518 distinct land units. For each polygon, the mean lowest groundwater level was retrieved from the ECOPLAN dataset. Soil texture fractions of each LU were derived from the most detailed statistical (historical) soil profile as described above. As there was no region-wide information available about functional traits, species diversity and biomass production, the corresponding training data were generalised per land cover class. Finally, the SOC stock of the land units was estimated by applying the best performing digital soil mapping method.

3. Results

3.1 Depth extrapolation of topsoil OC

The SOC stock in the upper 15 cm was on average 5.55 kg OC m^{-2}, with a minimum of 0.95 kg OC m^{-2} and a maximum of 15.32 kg OC m^{-2} (Table 1). These topsoil OC measurements were vertically extrapolated using the exponential change decline function (Eqs. 1 & 2), which was able to model SOC-rich subsoil horizons such as spodic horizons (Figure 2a) and peat layers (Figure 2b). The resulting SOC stock in the upper 100 cm of the training dataset was on average 15.75 kg OC m^{-2}, with a minimum of 4.65 kg OC m^{-2}, a median of 13.30 kg OC m^{-2} and a maximum of 72.84 kg OC m^{-2} (Table 1).

3.2 Exploratory data analysis

According to the results of the PCA, the following five traits optimally represented the trait space: specific leaf area (SLA, mm^2 mg^{-1}), leaf nitrogen (LN, mg g^{-1}), vegetation height (H, m), mycorrhizal associations (MA, %) and rooting depth (RD, m). The collinearity analysis among the candidate predictors (Figure 3 and SI1) resulted in four variables omitted from further modelling: the silt and sand fraction (r_s=0.71 and r_s=-0.85, $p \leq 0.01$ with the clay fraction), the highest groundwater level (r_s=0.90, $p \leq 0.01$ with the lowest groundwater level) and the CWM of LN (r_s=0.70, $p \leq 0.01$ with the CWM of SLA). Significant correlation was observed between SOC stock in the upper 15 cm and the following candidate predictors: clay fraction (r_s=0.21, $p \leq 0.05$), lowest groundwater level (r_s=-0.25, $p \leq 0.01$), biomass production (r_s=0.31, $p \leq 0.01$), CWM of H (r_s=0.28, $p \leq 0.01$) and MA (r_s=-0.23, $p \leq 0.01$), CWV of H (r_s=0.17, $p \leq 0.01$) and soil classification as Histosol (r_p=-0.19, $p \leq 0.05$), Phaeozem (r_p=0.19, $p \leq 0.05$), Retisol (r_p=-0.21, $p \leq 0.05$) or another WRB reference soil group (r_p=-0.19, $p \leq 0.05$). The SOC stock in the upper 100 cm, on the other hand, was significantly correlated with clay fraction (r_s=0.19, $p \leq 0.05$), lowest groundwater level (r_s=-0.34, $p \leq 0.01$), slope (r_s=-0.25 $p \leq 0.01$), biomass production (r_s=0.21, $p \leq 0.05$), CWM of SLA (r_s=-0.23, $p \leq 0.01$) and H

(r_s=0.20, p≤0.05) and soil classification as Arenosol (r_p=-0.20, p≤0.05), Gleysol (r_p=-0.35, p≤0.01), Histosol (r_p=0.20, p≤0.05) or another WRB reference soil group (r_p=-0.21, p≤0.05).

3.3 Model training and performance

3.3.1 SOC stock in the upper 15 cm

MLR highlighted the importance of seven variables in predicting the SOC stock in the upper 15 cm: lowest groundwater level, biomass production, CWV of MA and soil classification as Arenosol, Histosol, Retisol or other WRB RSG (Table 2). When considering all 97 submodels with a difference in AICc smaller than two compared to the so-called best model, additional variables affecting the SOC stock were identified: clay fraction, slope, CWM of SLA, H and MA, CWV of SLA, LN, H and RD and the soil classes Cambisol, Gleysol, Phaeozem and Podzol. Overall, MLR was outperformed by the other – non-linear, machine-learning – modelling techniques (Table 3). Of the latter, the BRT model resulted in the best fit of the training data ($R^2_{adj,train}$ = 0.69, $RMSE_{train}$ = 1.45 kg m^{-2}) and after cross-validation ($R^2_{adj,CV}$ = 0.19, $RMSE_{CV}$ = 2.22 kg m^{-2}). The optimal meta-parameters of the BRT model consisted of a tree complexity of 4, a learning rate of 0.002 and a bag fraction of 0.5 (SI2). This resulted in 1850 trees in the final BRT model. After simplification, WRB reference soil groups (relative importance of 25.1 %), CWM of SLA (12.9%), biomass production (8.2%), CWV of H (7.5%), CWV of RD (7.3%), CWM of H (7.1%), slope (6.6%), clay fraction (5%), CWM of RD (5.2%), CWM of MA (5.1%), species diversity (4.9%) and lowest groundwater level (4.6%) were identified as key predictors of the SOC stock in the upper 15 cm (Figure 4). The corresponding partial dependence plots show the effect of each variable on the SOC stock when accounting for the average effects of all other variables in the model. Larger SOC stocks in the upper 15 cm were found in Gleysols, Umbrisols, Phaeozems and Cambisols, whereas Retisols, other WRB groups and Histosols stored smaller stocks. These plots further indicate a negative effect of increasing CWM of SLA and MA, slope and depth of the

groundwater level and a positive effect of increasing biomass production, CWV and CWM of H and RD, clay fraction, and species diversity.

The optimal ANN consisted of one hidden layer with seven neurons trained using the Levenberg-Marquardt algorithm. An early stopping procedure and a regularisation parameter of 0.01 prevented overfitting. The optimal kernel function of the LS-SVM was the linear kernel.

3.3.2 SOC stock in the upper 100 cm

For the SOC stock in the upper 100 cm, the MLR best model consisted of ten variables: clay fraction, lowest groundwater level, slope, CWM and CWV of SLA, functional dispersion and soil classification as Gleysol, Histosol, Podzol or Umbrisol (Table 2). When considering all 55 submodels in the multimodel average, additional variables affecting the SOC stock were aboveground biomass production, CWM of H and MA, CWV of LN, H and RD and soil classification as 'other'. Again, MLR was outperformed by the other modelling techniques (Table 4), of which the BRT model resulted in the best fit after cross-validation ($R^2_{adj,CV}$ = 0.44, $RMSE_{CV}$ = 6.99 kg m^{-2}). An optimal tree complexity of 4, learning rate of 0.05 and bag fraction of 0.5 (SI3) resulted in a total number of 3500 trees. After simplification, soil type (relative importance of 39.6 %), lowest groundwater level (21%), clay fraction (15.7%) and CWV (12.1%) and CWM of H (11.6%) were identified as the main predictors of the SOC stock in the upper 100 cm (Figure 5). According to the partial dependence plots, larger stocks were found in Gleysols, Umbrisols and Histosols, whereas Arenosols and other WRB groups stored smaller stocks. Similar to the results of the topsoil OC, a negative effect of an increasing depth position of the groundwater level and positive effects of increasing clay fractions and CWM of H were observed.

The optimal ANN consisted of one hidden layer with five neurons trained using the Levenberg-Marquardt algorithm in combination with an early stopping procedure and a

358 regularisation parameter of 0.01. The optimal kernel function of the LS-SVM was the
359 Gaussian additive kernel.

360 3.4 Model application and upscaling

361 Given the best performance of the BRT model, it was applied to the 129,794 polygons
362 covering 30,556 ha. This resulted in a total SOC stock in the upper 15 cm of 1.63 Tg OC, with
363 a median and average value of 5.17 kg m^{-2} and 5.34 kg m^{-2}, respectively. Considering the
364 upper 100 cm led to larger estimates: a total of 4.53 Tg OC, with a median and average value
365 of 13.72 kg m^{-2} and 14.83 kg m^{-2}, respectively. Figure 6 shows the regional distribution of the
366 difference in SOC stock between the 100 cm and 15 cm estimate. LUs which showed the
367 largest differences, i.e. LUs in which the depth interval [15,100] is relatively carbon-rich,
368 corresponded to Gleysols and Histosols.

4. Discussion

4.1 Modelling regional subsoil OC stocks

Ecosystem service assessments rely on ecological indicators such as the SOC stock. However, routine soil sampling is often limited in depth and spatial density, which hampers accurate assessments of the SOC stock. To increase the vertical extent, several methods have been proposed including exponential decline (Hilinski, 2001), logarithmic (Jobbágy and Jackson, 2000) and power (Veronesi et al., 2014) functions. Taking full advantage of the presence of a detailed legacy soil dataset, an exponential change decline function was developed (Ottoy et al., 2016). Subsoil reference profiles are needed to enable detailed calibration and validation of this function. As subsoil measurements for LIHD systems are currently lacking, in future inventories soil should be sampled by pedogenetic horizon down to the parent material (Wiesmeier et al., 2012; Parras-Alcántara et al., 2015b). Our results (Figure 2) show the added value of this function to model SOC-rich subsoil horizons such as spodic (in Podzols) and peat horizons (in Gleysols and Histosols) in LIHD systems. Given the considerable spatial extent of these LUs (Figure 6), they should not be neglected in regional ecosystem service assessments. In our case, limiting the assessment to the topsoil (1,63 Tg OC) would result in a substantially smaller regional stock compared to the estimate of the upper 100 cm (4,53 Tg OC). Besides incomplete stock estimates, limitation to the topsoil can result in incomplete understanding of stock changes over time (Chapman et al., 2013; Jandl et al., 2014). Our average predicted stock in the upper 100 cm of soil under LIHD systems (14.83 kg OC m^{-2}) is larger than those observed in agricultural soils in Flanders: 10.4 kg OC m^{-2} under arable land and 12.5 kg OC m^{-2} under grassland (Ottoy et al., 2016). Forests, on the other hand, were found to contain a slightly larger stock (14.8 – 15.5 kg OC m^{-2}) in the complete territory of Belgium (Lettens et al., 2005). These findings show that LIHD systems – apart from their

potential biomass production for bioenergy (Tilman et al., 2006; Van Meerbeek et al., 2016) – also considerably contribute to climate change mitigation through SOC storage.

Both the regional stock of the upper 15 cm and 100 cm were estimated using BRT models, as this modelling technique outperformed the other considered techniques (MLR, ANN and LS-SVM). The observed goodness-of-fit indicator values are in line with those reported in earlier studies (Meersmans et al., 2008; Martin et al., 2014). In contrast to our results, Viscarra Rossel and Behrens (2010) found that SVM outperformed by BRT in predicting SOC content. An important advantage of BRT compared to ANN and LS-SVM is its ability to identify important predictors and visualise their effects through partial dependence plots (Elith et al., 2008).

Upscaling these predictors requires regional coverage of the predictor variables. Such information is not always available, e.g. plant functional traits are typically sampled at the (local) plot level. To this end, we aggregated the training data per land cover class. This approach is, however, limited because it does not account for the within-class variation, which can be as large as 75% of the total overall variation for some functional traits (Kattge et al., 2011). To take this local variation into account remote sensing techniques can offer an appropriate solution, e.g. the use of high-resolution LiDAR (Light Detection And Ranging) information to estimate CWM and CWV of vegetation height (Abelleira Martínez et al., 2016).

4.2 SOC stock predictors

Interesting insights in the soil system can be gained from the partial dependence plots of the BRT models. These results show that the model explaining the topsoil stock contained all variables of the model explaining the stock in the upper 100 cm (WRB RSG, LGWL, Clay, CWV and CWM of height) supplemented by information on CWM of specific leaf area, biomass production, CWV of rooting depth, slope, CWM of rooting depth, CWM of

mycorrhizal associations, and species diversity. The predictor with the highest relative importance in both models (25,1% resp. 39,6%) is the WRB reference soil group. For mineral SOC stocks in European forests, De Vos et al. (2015) also found that WRB RSG was given the highest relative importance. The effects of each soil group can be explained by their specific diagnostic characteristics (IUSS Working Group WRB, 2015): the presence of permanently high groundwater levels (Gleysols), thick organic layers (Histosols), spodic horizons (Podzols) or the accumulation of OC in the topsoil (Phaeozems and Umbrisols). On the other hand, sandy soils with limited or no profile development (Arenosols) and acid, nutrient-poor soils with an interfingered clay illuviation horizon (Retisols) store smaller SOC stocks (e.g. De Vos et al. (2015)). Remarkably, our results indicate that smaller stocks are stored in the topsoil of Histosols compared to the other soil groups. As the soil classification was retrieved from the soil map dating from 1947 – 1974, part of the information might be outdated. For soils classified as peat on the Dutch national soil map (1962 – 1995), Kempen et al. (2009) pointed to large-scale changes in land and water management after the national soil survey, resulting in (i) decreased peat thickness, or (ii) changed (from organic to mineral) soil type. For (extremely) wet grassland soils in Belgium, Meersmans et al. (2009) explained carbon losses between 1960 and 2006 by more intensive drainage during the last decades. These findings indicate that caution is required when using potentially outdated information on soil classification, especially for Histosols. Additional soil variables such as the clay fraction and groundwater level have but a minor effect on the topsoil stock, but are key predictors of the subsoil stock with a relative importance of 15,7% and 21%, respectively. The positive effect of an increasing clay fraction can be explained by the formation of stable, clay-protected organo-mineral complexes (Six et al., 2002), while the positive effect of a decreasing depth of the groundwater level is due to the hampering effect of oxygen deficiency on the decomposition rate of organic matter (Callesen et al., 2003).

443 The negative relationship between the CWM of specific leaf area and the topsoil stock has
444 also been observed by Manning et al. (2015) and Grigulis et al. (2013), who found larger
445 stocks under vegetation with thick and/or dense leaves. The positive effect of increasing
446 values of biomass production, CWV and CWM of height can be explained by larger carbon
447 inputs into the soil. Chapin (2003) already stressed the importance of plant traits related to
448 size and growth rate in the carbon cycle. Additionally, larger CWV and CWM values of
449 rooting depth can increase belowground carbon inputs and can promote interaction of SOC
450 with soil minerals to form stable organo-mineral complexes (Lorenz and Lal, 2005; De Deyn
451 et al., 2008). Furthermore, these variables were found to be positively correlated with the clay
452 fraction and negatively correlated with the depth of the groundwater table. The effect of
453 increasing mycorrhizal associations is rather ambiguous. Hodge et al. (2001) found that
454 arbuscular mycorrhizal symbiosis enhanced decomposition of grass leaves, which can have a
455 negative effect on the resulting SOC stock. However, mycorrhizal fungi are also found to
456 reduce carbon losses by slowing down root decomposition (Langley et al., 2006), producing
457 relatively stable glomalin carbon and promoting the formation of soil aggregates (Zhu and
458 Miller, 2003). The observation of a negative effect of increasing mycorrhizal associations in
459 our study is likely to be a result of correlated soil characteristics: larger CWM values are
460 found on dry (r_s=-0.43, $p \leq 0.01$), sandy (r_s= 0.25, $p \leq 0.01$) soils. Similarly, a higher species
461 diversity was found on nutrient-richer clayey (r_s=0.17, $p \leq 0.05$) or silty (r_s=0.21, $p \leq 0.05$) soils,
462 and is accompanied by larger CWV values of e.g. height (r_s=0.34, $p \leq 0.01$) and rooting depth
463 (r_s=0.57, $p \leq 0.01$).

5. Conclusions

Our study has shown that the topsoil (0-15 cm) of Low-Input High-Diversity systems in Flanders stored but 36% of the SOC stock found in the upper 100 cm. Apart from soil variables, vegetation characteristics such as specific leaf area, aboveground biomass production, plant height, rooting depth, mycorrhizal associations and species diversity influence the SOC stock. These findings indicate that the subsoil should not be neglected in ecosystem service assessments. We could obtain these results by combining depth extrapolation of topsoil OC measurements with digital soil mapping using a boosted regression trees approach. Legacy soil profile data are key inputs for the vertical exponential change decline function as they provide reference data about the depth dependence of the SOC content. This information is especially valuable for soil types in which the subsoil contributes substantially to the total SOC stock, i.e. by the presence of SOC-rich subsoil horizons which are typically overlooked by alternative depth extrapolation approaches. Secondly, legacy data provide soil predictors, such as clay fraction, for spatial upscaling. Further improvements of the training of the boosted regression trees and of the spatial upscaling might be gained by updating the soil map to provide actual information on the distribution of SOC-rich soil types. To capture the local variation in vegetation functional traits as predictors for upscaling, novel remote sensing techniques like high-resolution Light Detection And Ranging (LiDAR) are promising.

Acknowledgements

We acknowledge the anonymous reviewers for their constructive comments and suggestions. The research was supported by the IWT SBO Project 120014 “EcoPlan” funded by the Agency for Innovation by Science and Technology (IWT); and by PDM grant 3E150537 by the KU Leuven.

References

Abelleira Martínez, O.J., Fremier, A.K., Günter, S., Ramos Bendaña, Z., Vierling, L., Galbraith, S.M., Bosque-Pérez, N.A., Ordoñez, J.C., 2016. Scaling up functional traits for ecosystem services with remote sensing: concepts and methods. Ecol. Evol. 6, 4359–4371.

Aertsen, W., Kint, V., Van Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecol. Modell. 221, 1119–1130.

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

Aldana Jague, E., Sommer, M., Saby, N.P.A., Cornelis, J.-T., van Wesemael, B., Van Oost, K., 2016. High resolution characterization of the soil organic carbon depth profile in a soil landscape affected by erosion. Soil Tillage Res. 156, 185–193.

ALV, 2013. Landbouwgebruikspercelen (Agricultural land use parcels map).

Barton, K., 2016. Package " MuMIn ". R Package v1.15.6.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. Eur. J. Soil Sci. 47, 151–163.

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P., Delignette-Muller, M.-L., 2015. A Toolbox for Nonlinear Regression in R : The Package nlstools. J. Stat. Softw. 66, 1–21.

Beale, M., Hagan, M.T., Demuth, H.B., 2012. Neural Network Toolbox, User's Guide, MATLAB.

Broekx, S., Liekens, I., Peelaerts, W., De Nocker, L., Landuyt, D., Staes, J., Meire, P., Schaafsma, M., Van Reeth, W., Van den Kerckhove, O., Cerulus, T., 2013. A web application to support the quantification and valuation of ecosystem services. Environ. Impact Assess. Rev. 40, 65–74.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information - theoretic approach, 2nd editio. ed. Springer, New York.

Callesen, I., Liski, J., Raulund-Rasmussen, K., Olsson, M.T., Tau-Strand, L., Vesterdal, L., Westman, C.J., 2003. Soil carbon stores in Nordic well-drained forest soils-relationships with climate and texture class. Glob. Chang. Biol. 9, 358–370.

Carré, F., McBratney, A.B., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141, 1–14.

Chapin, F.S., 2003. Effects of plant traits on ecosystem and regional processes: a conceptual framework for predicting the consequences of global change. Ann. Bot. 91, 455–463.

Chapman, S.J., Bell, J.S., Campbell, C.D., Hudson, G., Lilly, A., Nolan, A.J., Robertson, A.H.J., Potts, J.M., Towers, W., 2013. Comparison of soil carbon stocks in Scottish soils between 1978 and 2009. Eur. J. Soil Sci. 64, 455–465.

De Brabanter, K., Suykens, J.A.K., De Moor, B., 2013. Nonparametric regression via StatLSSVM. J. Stat. Softw. 55, 1 – 21.

De Deyn, G.B., Cornelissen, J.H.C., Bardgett, R.D., 2008. Plant functional traits and soil carbon sequestration in contrasting biomes. Ecol. Lett. 11, 516–531.

De Saeger, S., Ameeuw, G., Berten, B., Bosch, H., Brichau, I., De Knijf, G., Demolder, H., Erens, G., Guelinckx, R., Oosterlynck, P., Rombouts, K., Scheldeman, K., T'jollyn, F., Van Hove, M., Van Ormelingen, J., Vriens, L., Zwaenepoel, A., Van Dam, G., Verheirstraeten, M., Wils, C., Paelinckx, D., 2010. Biologische Waarderingskaart versie 2.2. Rapporten van het Instituut voor Natuur- en Bosonderzoek 2010 (36). Instituut voor Natuur- en Bosonderzoek, Brussel

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.

Dondeyne, S., Vanierschot, L., Langohr, R., Van Ranst, E., Deckers, J., 2014. The soil map of the Flemish region converted to the 3rd edition of the World Reference Base for soil resources.

Durka, W., Michalski, S.G., 2012. Daphne: a dated phylogeny of a large European flora for phylogenetically informed ecological analyses. Ecology 93, 2297–2297.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.

Galati, A., Crescimanno, M., Gristina, L., Keesstra, S., Novara, A., 2016. Actual provision as an alternative criterion to improve the efficiency of payments for ecosystem services for C sequestration in semiarid vineyards. Agric. Syst. 144, 58–64.

Goldstein, J.H., Caldarone, G., Duarte, T.K., Ennaanay, D., Hannahs, N., Mendoza, G., Polasky, S., Wolny, S., Daily, G.C., 2012. Integrating ecosystem-service tradeoffs into land-use decisions. Proc. Natl. Acad. Sci. U. S. A. 109, 7565–7570.

Goolsby, E.W., Bruggeman, J., Ane, C., 2016. Package "Rphylopars."

Grigulis, K., Lavorel, S., Krainer, U., Legay, N., Baxendale, C., Dumont, M., Kastl, E., Arnoldi, C., Bardgett, R.D., Poly, F., Pommier, T., Schloter, M., Tappeiner, U., Bahn, M., Clément, J.-C., 2013. Relative contributions of plant traits and soil microbial properties to mountain grassland ecosystem services. J. Ecol. 101, 47–57.

Haykin, S., 1998. Neural networks: A comprehensive foundation. 2nd edition. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2016. Package "dismo."

Hilinski, T., 2001. Implementation of exponential depth distribution of organic carbon in the CENTURY model [WWW Document]. URL http://www.nrel.colostate.edu/projects/century5/documents/ExponentialCDistribution.pdf (accessed 9.3.16).

Hodge, A., Campbell, C.D., Fitter, A.H., 2001. An arbuscular mycorrhizal fungus accelerates decomposition and acquires nitrogen directly from organic material. Nature 413, 297–299.

IUSS Working Group WRB, 2015. World reference base for soil resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps, World Soil Resources Reports No. 106. FAO, Rome.

IUSS Working Group WRB, 2014. World Reference Base for Soil Resources 2014: International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D.D., Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status , uncertainty and future needs in soil organic carbon monitoring. Sci. Total Environ. 468-469, 376–383.

Jenny, H., 1941. Factors of soil formation. A system of quantitative pedology. McGraw-Hill, New York, USA.

Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10, 423–436.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I.J., Cornelissen, J.H.C., Violle, C., Harrison, S.P., Van Bodegom, P.M., Reichstein, M., Enquist, B.J., Soudzilovskaia, N.A., Ackerly, D.D., Anand, M., Atkin, O., Bahn, M., Baker, T.R., Baldocchi, D., Bekker, R., Blanco, C.C., Blonder, B., Bond, W.J., Bradstock, R., Bunker, D.E., Casanoves, F., Cavender-Bares, J., Chambers, J.Q., Chapin, F.S., Chave, J., Coomes, D., Cornwell, W.K., Craine, J.M., Dobrin, B.H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W.F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G.T., Fyllas, N.M., Gallagher, R. V., Green, W. a., Gutierrez, a. G., Hickler, T., Higgins, S.I., Hodgson, J.G., Jalili, A., Jansen, S., Joly, C. a., Kerkhoff, a. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J.M.H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T.D., Leishman, M., Lens, F., Lenz, T., Lewis, S.L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M.D., Manning, P., Massad, T., Medlyn, B.E., Messier, J., Moles, a. T., Müller, S.C., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V.G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W. a., Patiño, S., Paula, S., Pausas, J.G., Peñuelas, J., Phillips, O.L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S.J., Yguel, B., Zaehle, S., Zanne, a. E., Wirth, C., 2011. TRY - a global database of plant traits. Glob. Chang. Biol. 17, 2905–2935.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P., Webb, C.O., 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26, 1463–1464.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311–326.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Laliberté, E., Legendre, P., 2010. A distance-based framework for measuring functional diversity from multiple traits. Ecology 91, 299–305.

Laliberté, E., Legendre, P., Shipley, B., 2014. Package "FD."

Langley, J.A., Chapman, S.K., Hungate, B.A., 2006. Ectomycorrhizal colonization slows root decomposition: The post-mortem fungal legacy. Ecol. Lett. 9, 955–959.

Lettens, S., Van Orshoven, J., van Wesemael, B., De Vos, B., Muys, B., 2005. Stocks and fluxes of soil organic carbon for landscape units in Belgium derived from heterogeneous data sets for 1990 and 2000. Geoderma 127, 11–23.

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

Loreau, M., Hector, A., 2001. Partitioning selection and complementarity in biodiversity experiments. Nature 412, 72–76.

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

Maes, J., Liquete, C., Teller, A., Erhard, M., Paracchini, M.L., Barredo, J.I., Grizzetti, B., Cardoso, A., Somma, F., Petersen, J.-E., Meiner, A., Gelabert, E.R., Zal, N., Kristensen, P., Bastrup-Birk, A., Biala, K., Piroddi, C., Egoh, B., Degeorges, P., Fiorina, C., Santos-Martín, F., Naruševičius, V., Verboven, J., Pereira, H.M., Bengtsson, J., Gocheva, K., Marta-Pedroso, C., Snäll, T., Estreguil, C., San-Miguel-Ayanz, J., Pérez-Soba, M., Grêt-Regamey, A., Lillebø, A.I., Malak, D.A., Condé, S., Moen, J., Czúcz, B., Drakou, E.G., Zulian, G., Lavalle, C., 2016. An indicator framework for assessing ecosystem services in support of the EU Biodiversity Strategy to 2020. Ecosyst. Serv. 17, 14–23.

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. J. Appl. Ecol. 52, 1188–1196.

640 Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P.A., Paroissien, J.B., Jolivet,
641 C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting
642 the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223-
643 225, 97–107.

644 McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping.
645 Geoderma 117, 3–52.

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

649 Meersmans, J., van Wesemael, B., De Ridder, F., Dotti, M.F., De Baets, S., Van Molle, M.,
650 2009. Changes in organic carbon distribution with depth in agricultural soils in northern
651 Belgium, 1960-2006. Glob. Chang. Biol. 15, 2739–2750.

652 Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital Mapping of Soil
653 Carbon. Adv. Agron. 118, 1–47.

654 O'Rourke, S.M., Angers, D.A., Holden, N.M., Mcbratney, A.B., 2015. Soil organic carbon
655 across scales. Glob. Chang. Biol. 21, 3561–3574.

656 Ottoy, S., Beckers, V., Jacxsens, P., Hermy, M., Van Orshoven, J., 2015. Multi-level
657 statistical soil profiles for assessing regional soil organic carbon stocks. Geoderma 253-
658 254, 12–20.

659 Ottoy, S., Elsen, A., Van De Vreken, P., Gobin, A., Merckx, R., Hermy, M., Van Orshoven,
660 J., 2016. An exponential change decline function to estimate soil organic carbon stocks
661 and their changes from topsoil measurements. Eur. J. Soil Sci. 67, 816–826.

662 Parras-Alcántara, L., Díaz-Jaimes, L., Lozano-García, B., 2015a. Management effects on soil
663 organic carbon stock in Mediterranean open rangelands-treeless grasslands. L. Degrad.
664 Dev. 26, 22–34.

665 Parras-Alcántara, L., Lozano-García, B., Brevik, E.C., Cerdá, A., 2015b. Soil organic carbon
666 stocks assessment in Mediterranean natural areas: A comparison of entire soil profiles
667 and soil control sections. J. Environ. Manage. 155, 219–228.

668 Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-
669 Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644.

670 Pérez-Harguindeguy, N., Díaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P.,
671 Bret-Harte, M.S., Cornwell, W.K., Craine, J.M., Gurvich, D.E., Urcelay, C., Veneklaas,
672 E.J., Reich, P.B., Poorter, L., Wright, I.J., Ray, P., Enrico, L., Pausas, J.G., de Vos, A.C.,
673 Buchmann, N., Funes, G., Quétier, F., Hodgson, J.G., Thompson, K., Morgan, H.D., ter
674 Steege, H., van der Heijden, M.G.A., Sack, L., Blonder, B., Poschlod, P., Vaieretti, M.
675 V., Conti, G., Staver, A.C., Aquino, S., Cornelissen, J.H.C., 2013. New Handbook for
676 standardized measurment of plant functional traits worldwide. Aust. J. Bot. 61, 167–234.

677 R Core Team, 2013. R: A language and environment for statistical computing. R Foundation
678 for Statistical Computing, Vienna, Austria.

679 Rawls, W.J., 1983. Estimating soil bulk density from particle size analysis and organic matter
680 content. Soil Sci. 135, 123–125.

681 Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms of soil organic
682 matter: Implications for C-saturation of soils. Plant Soil 241, 155–176.

683 Sleutel, S., de Neve, S., Hofman, G., 2003. Estimates of carbon stock changes in Belgian
684 cropland. Soil Use Manag. 19, 166–171.

685 Sonnier, G., Shipley, B., Navas, M.-L., 2010. Quantifying relationships between traits and
686 explicitly measured gradients of stress and disturbance in early successional plant
687 communities. J. Veg. Sci. 21, 1014–1024.

688 Staes, J., 2016. Historisch gemiddelde grondwaterstand [WWW Document]. URL
689 http://ecosysteemdiensten.be/cms/nl/indicator/bl_ghg_glg (accessed 24.10.2016).

690 Steinbeiss, S., Beßler, H., Engels, C., Temperton, V.M., Buchmann, N., Roscher, C.,
691 Kreutziger, Y., Baade, J., Habekost, M., Gleixner, G., 2008. Plant diversity positively
692 affects short-term soil carbon storage in experimental grasslands. Glob. Chang. Biol. 14,
693 2937–2949.

694 Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least
695 Squares Support Vector Machines. World Scientific, Singapore.

696 Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic
697 carbon at multiple depths using different data mining techniques in Baneh region, Iran.
698 Geoderma 266, 98–110.

699 Tilman, D., Hill, J., Lehman, C., 2006. Carbon-Negative Biofuels from Low-Input High-
700 Diversity Grassland Biomass. Science (80-.). 314, 1598–1600.

701 Van Meerbeek, K., Ottoy, S., de Andrés García, M., Muys, B., Hermy, M., 2016. The
702 bioenergy potential of Natura 2000 - a synergy between climate change mitigation and
703 biodiversity protection. Front. Ecol. Environ. 14, 473–478.

704 Van Meerbeek, K., Van Beek, J., Bellings, L., Aertsen, W., Muys, B., Hermy, M., 2014.
705 Quantification and Prediction of Biomass Yield of Temperate Low-Input High-Diversity
706 Ecosystems. Bioenergy Res. 7, 1120–1130.

707 Van Orshoven, J., Deckers, J.A., Vandenbroucke, D., Feyen, J., 1993. The completed
708 database of Belgian soil profile data and its applicability in the planning and
709 management of rural land. Bull. des Rech. Agron. Gembloux 28, 197–222.

710 van Wesemael, B., Paustian, K., Meersmans, J., Goidts, E., Barancikova, G., Easter, M.,
711 2010. Agricultural management explains historic changes in regional soil carbon stocks.
712 Proc. Natl. Acad. Sci. U. S. A. 107, 14926–14930.

713 Veronesi, F., Corstanje, R., Mayr, T., 2014. Landscape scale estimation of soil carbon stock
714 using 3D modelling. Sci. Total Environ. 487, 578–586.

715 Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil
716 diffuse reflectance spectra. Geoderma 158, 46–54.

717 Vrebos, D., 2015. Bodembedekking [WWW Document]. URL
718 http://www.ecosysteemdiensten.be/cms/indicator/bl_lc (accessed 24.10.2016).

719 Walkley, A., Black, I., 1934. An examination of the Degtjareff method for determining soil
720 organic matter, and a proposed modification of the chromic acid titration method. Soil
721 Sci. 37, 29–38.

722 Wei, T., Simko, V., 2016. Package "corrplot."

723 Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support
724 vector regression, artificial neural networks, and random forests for predicting and
725 mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52,
726 394–403.

727 Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil
728 organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem.
729 Plant Soil 340, 7–24.

730 Wiesmeier, M., Spörlein, P., Geuß, U., Hangen, E., Haug, S., Reischl, A., Schilling, B., von
731 Lützow, M., Kögel-Knabner, I., 2012. Soil organic carbon stocks in southeast Germany
732 (Bavaria) as affected by land use, soil type and sampling depth. Glob. Chang. Biol. 18,
733 2233–2245.

734 Zhu, Y.-G., Miller, R.M., 2003. Carbon cycling by arbuscular mycorrhizal fungi in soil-plant
735 systems. Trends Plant Sci. 8, 407–409.

736 Table 1: Descriptive statistics of the different model variables in the training dataset.

Variables	Data type [unit]	Model training				
		n	min	max	mean	med
SOC stock 15 cm	Cont. [kg m^{-2}]	139	0.95	15.32	5.55	5.30
SOC stock 100 cm	Cont. [kg m^{-2}]	139	4.65	72.84	15.75	13.30
Soil texture fraction	Cont. [%]					
Clay		139	0	44.00	12.04	10.32
Silt		139	1.67	81.09	27.20	17.99
Sand		139	1.91	96.55	60.26	72.25
Groundwater level	Cont. [cm]					
Highest		139	1.50	738.29	66.56	27.85
Lowest		139	2.50	760.44	129.89	97.65
Slope	Cont. [%]	139	0.18	41.71	2.56	1.48
Biomass production	Cont. [ton DM ha^{-1}]	139	0.19	25.32	5.30	4.21
Species diversity	Cont. []	139	1.00	33.00	12.58	13.00
CWM						
Specific leaf area	Cont. [mm^2 mg^{-1}]	139	8.08	34.16	23.02	24.17
Leaf Nitrogen	Cont. [mg g^{-1}]	139	11.37	38.71	23.97	24.44
Height	Cont. [m]	139	0.26	1.68	0.68	0.60
Mycorrhizal associations	Cont. [%]	139	12.87	78.65	55.26	63.39
Rooting depth	Cont. [m]	139	0.23	1.26	0.49	0.46
CWV						
Specific leaf area	Cont. [mm^2 mg^{-1}]	139	0.00	121.30	31.51	23.45
Leaf Nitrogen	Cont. [mg g^{-1}]	139	0.00	378.87	40.63	30.17
Height	Cont. [m]	139	0.00	0.40	0.05	0.03
Mycorrhizal associations	Cont. [%]	139	0.00	1138.28	265.05	182.90
Rooting depth	Cont. [m]	139	0.00	0.29	0.05	0.03
Functional dispersion	Cont. []	139	0.00	2.17	1.20	1.23
Soil type	Bin. [Presence]					
Arenosols		14				
Cambisols		26				
Gleysols		6				
Histosols		15				
Other		10				
Phaeozems		23				
Podzols		32				
Retisols		7				
Umbrisols		6				

737 cont., continuous; bin., binary; CWM, community weighted mean; CWV, community weighted variance.

738 Table 2: Results of the multiple linear regression (MLR) models to estimate the SOC stock in
739 the upper 15 and 100 cm. The relative importance (RI), coefficients (Coef.) and their 95%
740 confidence interval (95% CI) of the multimodel average (MMA) together with the
741 coefficients of the best model (BM).

Predictor	MLR SOC_{15cm}				MLR SOC_{100cm}			
	MMA			BM	MMA			BM
	RI	Coef.	95% CI	Coef.	RI	Coef.	95% CI	Coef.
Intercept		4.46	[2.90, 6.02]	5.06		14.15	[5.70, 22.60]	16.32
Clay	0.41	0.03	[-0.01, 0.08]		1.00	0.42	[0.24, 0.61]	0.45
LGWL	0.55	-0.003	[-0.006, 0.001]	-0.003	1.00	-0.02	[-0.03, -0.002]	-0.02
Slope	0.02	0.04	[-0.06, 0.13]		0.81	0.32	[-0.05, 0.68]	0.32
BP	1.00	0.16	[0.04, 0.28]	0.19	0.21	0.25	[-0.12, 0.62]	
SD					0.01	-0.14	[-0.38, 0.11]	
CWM								
SLA	0.09	-0.03	[-0.1, 0.04]		0.43	-0.19	[-0.44, 0.05]	-0.21
LN								
H	0.39	1.44	[-0.38, 3.26]		0.03	2.77	[-3.11, 8.65]	
MA	<0.01	-0.01	[-0.04, 0.02]		0.15	-0.06	[-0.16, 0.04]	
RD					0.01	-4.29	[-12.91, 4.33]	
CWV								
SLA	0.03	0.01	[-0.01, 0.02]		0.52	0.05	[-0.01, 0.11]	0.05
LN	0.19	0.01	[-0.004, 0.02]		0.16	-0.02	[-0.06, 0.01]	
H	0.18	4.59	[-2.11, 11.28]		0.17	14.54	[-9.94, 39.02]	
MA	0.93	0.002	[0.0001, 0.003]	0.002				
RD	<0.01	3.01	[-4.49, 10.52]		0.15	-17.45	[-47.11, 12.20]	
FDis					1.00	-3.41	[-6.32, -0.51]	-4.02
WRB RSG								
AR	0.48	-1.00	[-2.29, 0.28]	-1.10	0.03	1.95	[-3.41, 7.31]	
CM	0.03	0.45	[-0.54, 1.44]					
GL	0.48	1.39	[-0.42, 3.20]		1.00	19.92	[13.07, 26.78]	19.98
HS	1.00	-2.59	[-3.86, -1.32]	-2.66	1.00	5.92	[1.10, 10.74]	5.54
PH	0.02	-0.54	[-1.72, 0.63]					
PZ	0.03	-0.52	[-1.53, 0.48]		1.00	6.98	[2.78, 11.17]	7.23
RT	1.00	-2.59	[-4.31, -0.87]	-2.78				
UM					1.00	10.92	[3.97, 17.86]	10.75
Other	1.00	-1.99	[-3.45, -0.53]	-2.08	0.04	-2.49	[-8.12, 3.14]	

742 LGWL, lowest groundwater level; BP, biomass production; SD, species diversity; CWM, community weighted
743 mean; CWV, community weighted variance; SLA, specific leaf area; LN, leaf nitrogen; H, height; MA,
744 mycorrhizal associations; RD, rooting depth; FDis, functional dispersion; WRB RSG, world reference base
745 reference soil group; AR, Arenosol; CM, Cambisol; GL, Gleysol; HS, Histosol; PH, Phaeozem; PZ, Podzol; RT,
746 Retisol; UM; Umbrisol.

747 Table 3: Goodness-of-fit measurements of the four digital soil mapping techniques (MLR,
748 multiple linear regression; BRT, boosted regression trees; ANN, artificial neural networks;
749 LS-SVM, least-squares support vector machines) in modelling the SOC stock of the upper 15
750 cm. These indicators were calculated on the complete training dataset (train) and after 10-fold
751 cross-validation (CV).

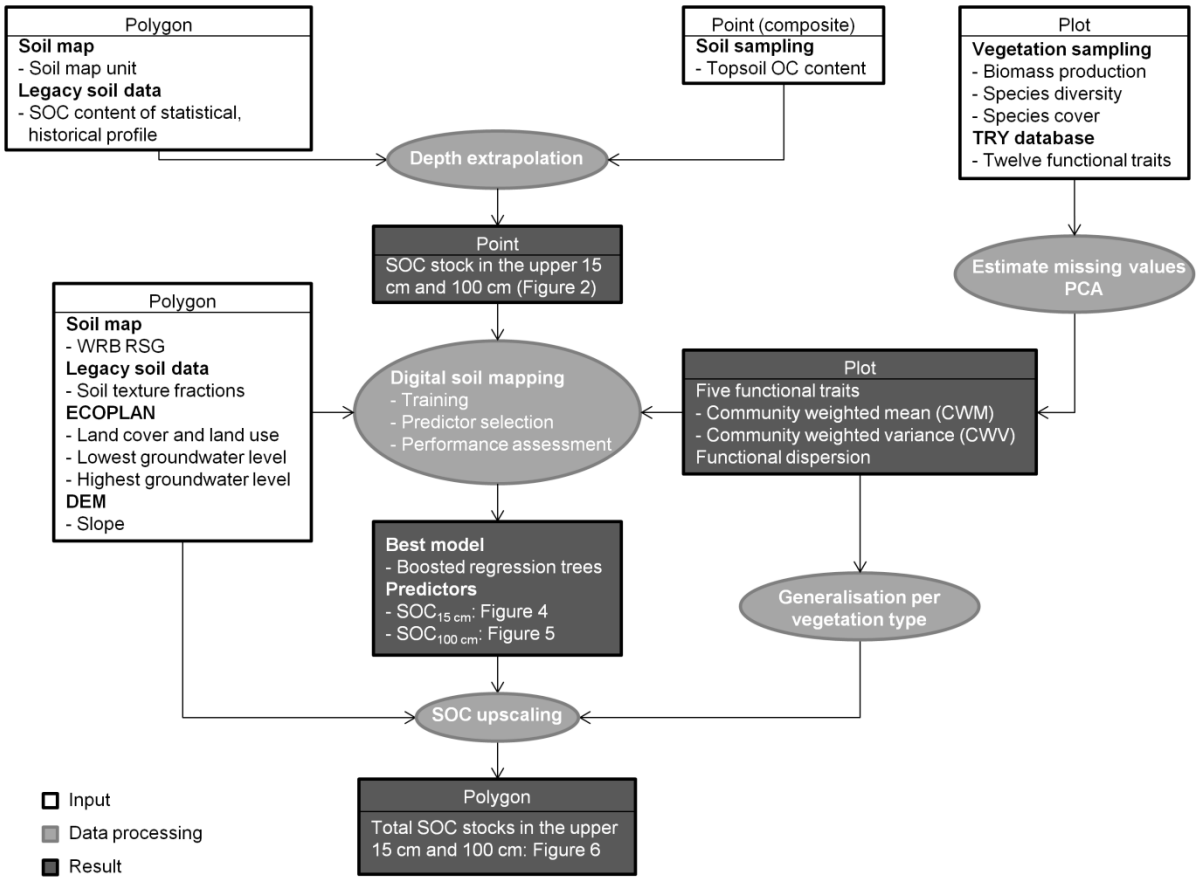
Goodness-of-fit indicator	MLR		BRT	ANN	LS-SVM
	MMA	BM			
R^2_{train}	0.29	0.29	0.72	0.37	0.33
$R^2_{adj,train}$	0.17	0.26	0.69	0.28	0.23
$RMSE_{train}$ (kg m^{-2})	2.16	2.07	1.45	1.96	2.03
$rRMSE_{train}$ (%)	38.98	37.34	26.18	35.30	36.57
R^2_{CV}	-	0.18	0.26	0.19	0.27
$R^2_{adj,CV}$	-	0.14	0.19	0.07	0.17
$RMSE_{CV}$ (kg m^{-2})	-	2.26	2.22	2.58	2.24
$rRMSE_{CV}$ (%)	-	40.62	40.03	45.62	40.56

752 R^2, coefficient of determination; R^2_{adj}, adjusted R^2; RMSE, root mean squared error; rRMSE, relative RMSE.

753 Table 4: Goodness-of-fit measurements of the four digital soil mapping techniques (MLR,
754 multiple linear regression; BRT, boosted regression trees; ANN, artificial neural networks;
755 LS-SVM, least-squares support vector machines) in modelling the SOC stock of the upper
756 100 cm. These indicators were calculated on the complete training dataset (train) and after 10-
757 fold cross-validation (CV).

Goodness-of-fit indicator	MLR		BRT	ANN	LS-SVM
	MMA	BM			
R^2_{train}	0.38	0.40	0.99	0.63	0.99
$R^2_{adj,train}$	0.28	0.36	0.99	0.61	0.99
$RMSE_{train}$ (kg m^{-2})	10.14	7.68	0.16	5.88	0.38
$rRMSE_{train}$ (%)	64.38	48.79	1.04	37.39	2.48
R^2_{CV}	-	0.17	0.45	0.37	0.43
$R^2_{adj,CV}$	-	0.10	0.44	0.32	0.38
$RMSE_{CV}$ (kg m^{-2})	-	9.54	6.99	8.59	7.67
$rRMSE_{CV}$ (%)	-	60.58	44.38	53.65	47.82

758 R^2, coefficient of determination; R^2_{adj}, adjusted R^2; RMSE, root mean squared error; rRMSE, relative RMSE.



759

760 Figure 1: A schematic overview of the applied procedure. A distinction is made between input
761 data, data processing and results.

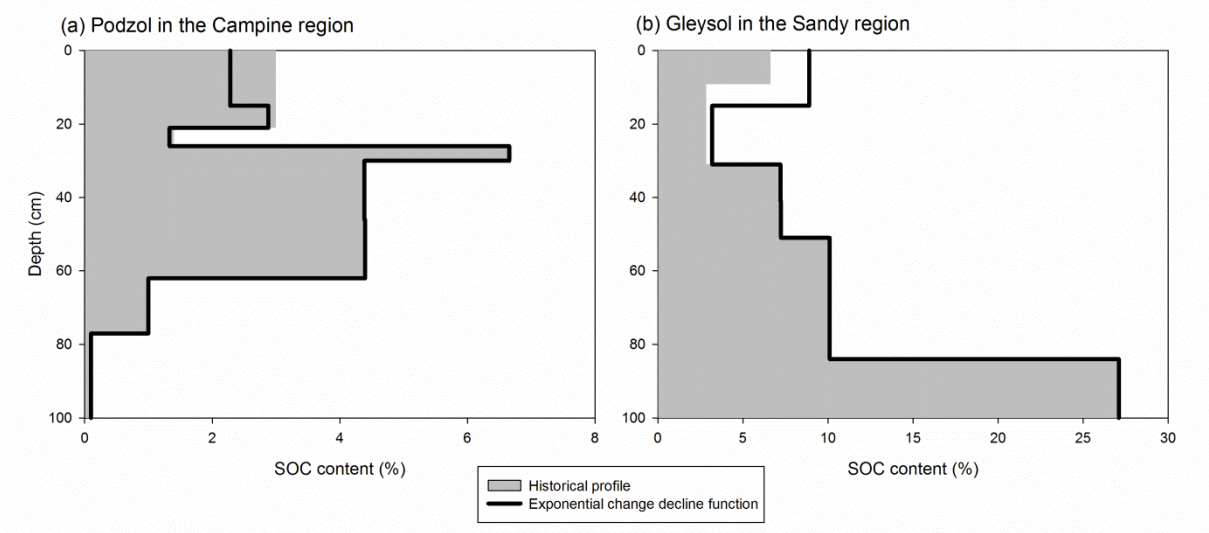


Figure 2: The modelled vertical distribution of soil organic carbon (SOC) content (%) using the exponential change decline function (black line) together with the historical SOC profile (grey bars) of two land units: (a) Podzol in the Campine region and (b) Gleysol in the Sandy region.

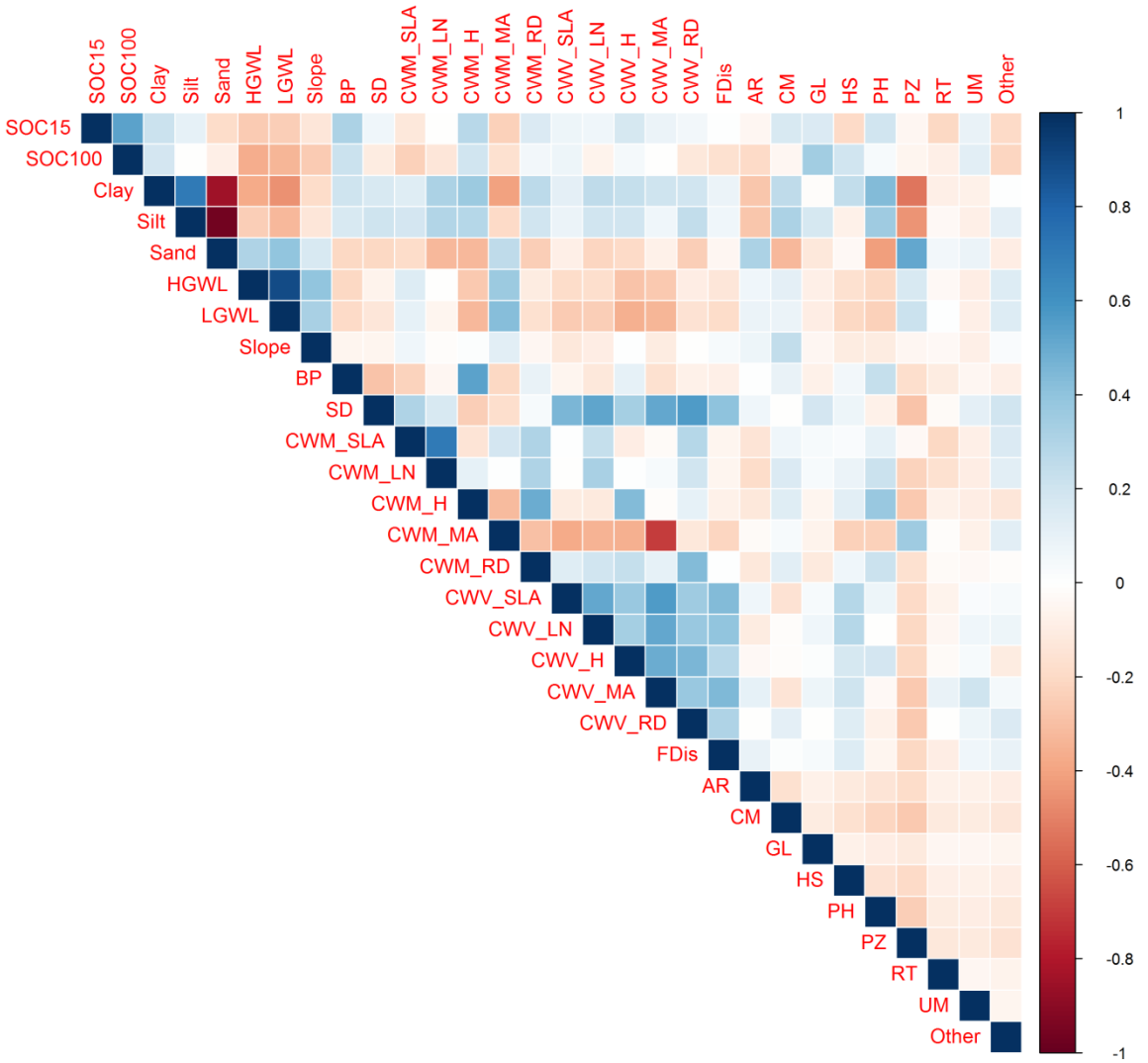


Figure 3: Graphical display of the correlation matrix, using the 'corrplot' package of R-software (Wei and Simko, 2016). It comprises Spearman's rank correlations for two continuous variables and point-biserial correlations for one continuous and one binary variable. The corresponding correlation coefficients and p-values can be found in SI1. HGWL, highest groundwater level; LGWL, lowest groundwater level; BP, biomass production; SD, species diversity; CWM, community weighted mean; CWV, community weighted variance; SLA, specific leaf area; LN, leaf nitrogen; H, height; MA, mycorrhizal associations; RD, rooting depth; FDis, functional dispersion; AR, Arenosol; CM, Cambisol; GL, Gleysol; HS, Histosol; PH, Phaeozem; PZ, Podzol; RT, Retisol; UM; Umbrisol.

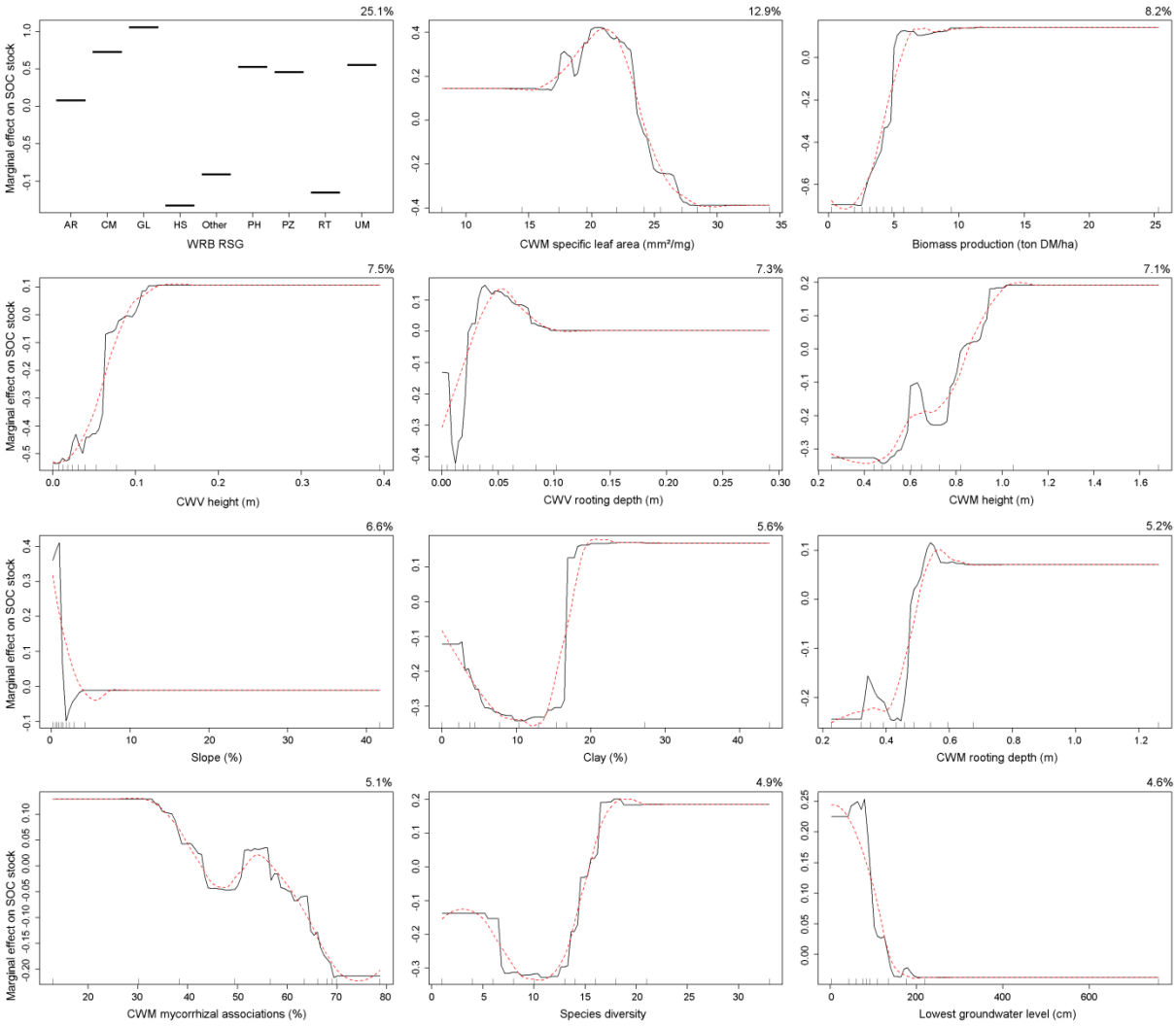


Figure 4: Partial dependence plots of the soil organic carbon stock in the upper 15 cm to the predictors in the boosted regression trees. The plots indicate the effect of each predictor, given the average effects of all other predictors in the model. The relative importance (%) of each predictor is reported above the upper right corner of each plot.
CWM, community weighted mean; CWV, community weighted variance; WRB RSG, world reference base reference soil group; AR, Arenosol; CM, Cambisol; GL, Gleysol; HS, Histosol; PH, Phaeozem; PZ, Podzol; RT, Retisol; UM; Umbrisol.

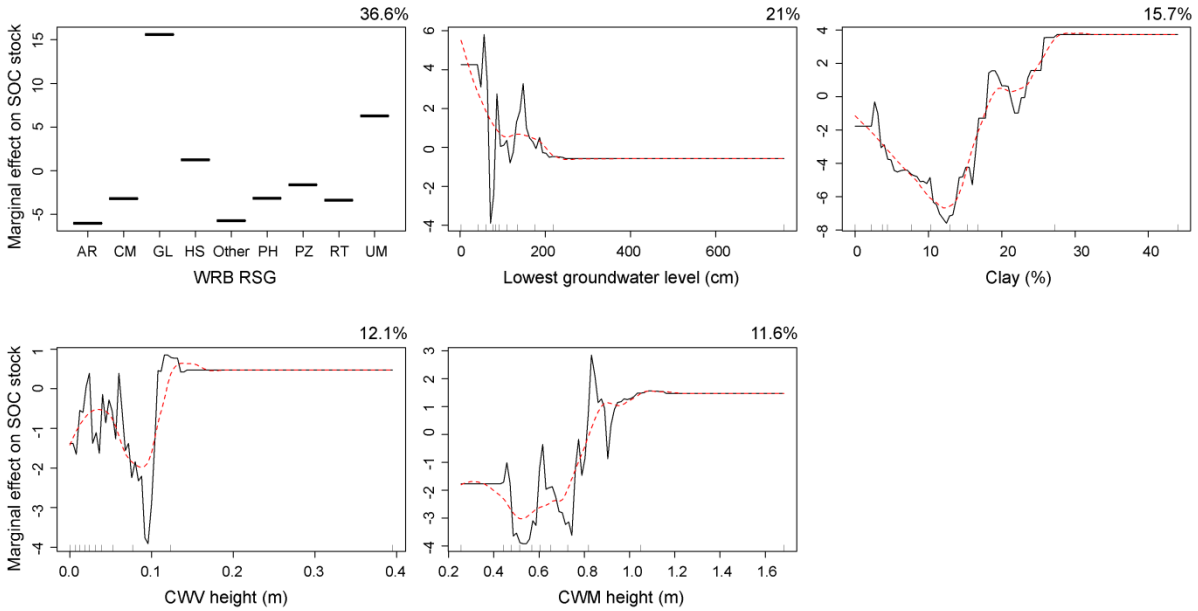
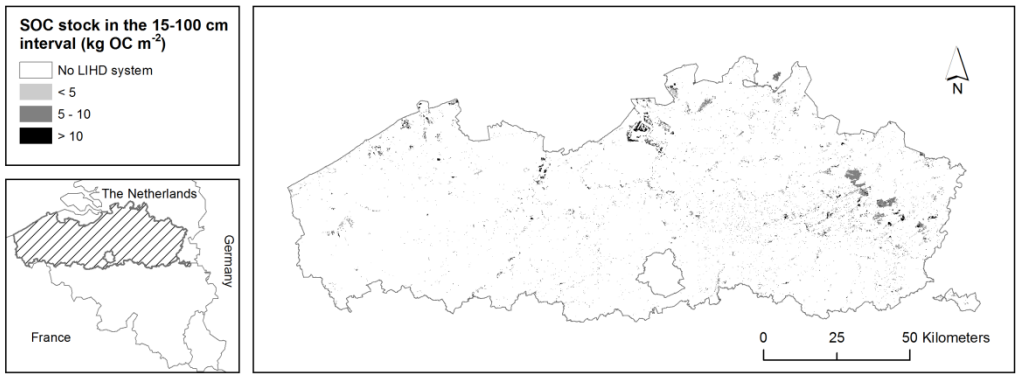


Figure 5: Partial dependence plots of the soil organic carbon stock in the upper 100 cm to the predictors in the boosted regression trees. The plots indicate the effect of each predictor, given the average effects of all other predictors in the model. The relative importance (%) of each predictor is reported above the upper right corner of each plot. CWM, community weighted mean; CWV, community weighted variance; WRB RSG, world reference base reference soil group; AR, Arenosol; CM, Cambisol; GL, Gleysol; HS, Histosol; PH, Phaeozem; PZ, Podzol; RT, Retisol; UM; Umbrisol.



794

795 Figure 6: Geographical distribution of the difference in soil organic carbon stock between the
796 15 cm and 100 cm estimate by the resulting boosted regression trees for Low-Input
797 High-Diversity (LIHD) systems in Flanders, Belgium.