Supplementary information 1:

Selection of the controlling factors of OC_{S+C} with a general additive mixed effect model (GAMM)

A GAMM includes the feature of both generalized additive models (GAM) and linear mixed effect models (LMM) (Zuur et al., 2009). The GAM features of the GAMM allow to model non-linear data, because the response variable can depend on a smoother function of the selected predictor variable. In this study, it allows to include a smoother for the predictor variable "depth", which has a non-linear relationship with OC_{S+C}. The smoother corrects the prediction of the dependent variable OC_{S+C} accordingly to the non-linear nature of this relationship. The LMM features of the GAMM allow to separate predictor variables into fixed effect and random effect. This is particularly useful when more than one measurement is made on a given statistical unit. It is therefore adopted in this study, as at each location multiple soil samples were collected at different depths. The variable "site" was included as a random effect, on the one hand because there are multiple observations for each site (for different depths). On the other hand, because there are 40 sites and using "site" as a fixed effect would be very expensive in terms of degrees of freedom. Finally, our study aims at highlighting a general relationship for any site, not specifically for these 40 sites.

To determine the best approach to model the data, in a first step a linear model of OC_{S+C} was fitted to assess its performance (*LM* function in R). This approach resulted in a clear violation of homogeneity of variances, which can be explained by the non-linear relationship between the predictor variable "depth" and the dependent variable OC_{S+C} (Figure S6). A solution for dealing with a non-linear relationship is to use a general additive model (GAM). This allows to include a smoother for the predictor variable "depth". This smoother corrects the prediction of the dependent variable OC_{S+C} accordingly to the non-linear nature of its relationship with depth. Another issue with applying a statistical model to this dataset is the dependence between the samples of each individual soil profile, as these were collected at the same location. The location (i.e. the variable "site") was therefore included as a random effect. Because it includes all the required features, a GAMM was considered the best approach for this study.

As OC_{S+C} concentration depth profiles are different for the different land uses (Figure S19), using different depth smoother for the different land use to predict OC_{S+C} concentration along soil profile might be a suitable option. Two models using every selected variable and realistic interaction were fit, one with one smoother for both lands uses and one with two different smoothers. For both models, a second version with a power variance function on depth was also fitted. This type of function can help to couple with heterogeneity, which may need to be addressed because there is more variability in the topsoil than in the deeper depth. The way it was used, the model assumes homogeneity between sites but heterogeneity within sites along depth (and the strength of the heterogeneity along the depth gradient is the same for each site). It models the residual spread for the profiles in such a way that their variance is proportional to the variance covariate "depth". AIC and BIC both favored the model with only one smoother and a power variance function on depth (Table S2).

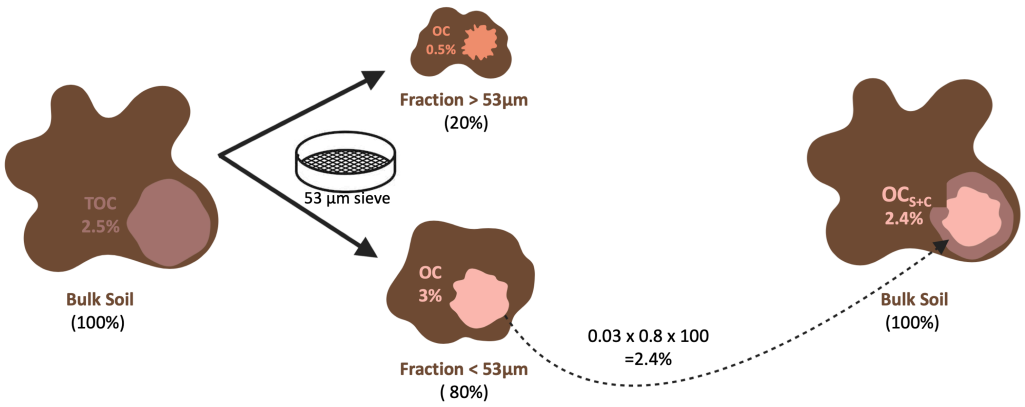


Figure S1: Scheme of the size fractionation of bulk soil into two fractions. The partition of total organic carbon (TOC) as organic carbon (OC) in the two resulting fractions is highlighted, well as the calculation of OC_{S+C}, using fictional data.

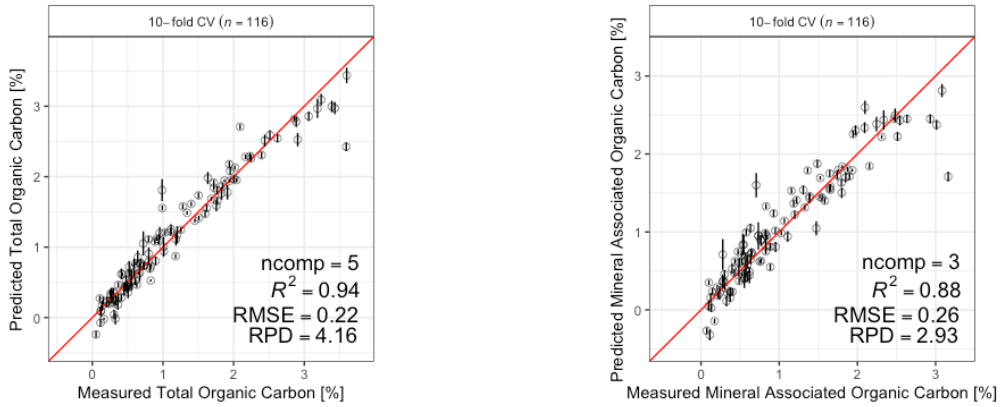


Figure S2: Measured versus predicted values of the MIR prediction models for total organic carbon and mineral associated organic carbon (i.e., OC_{S+C})

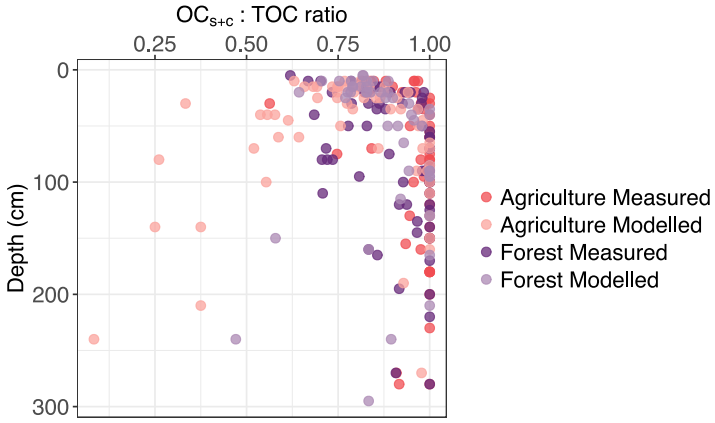


Figure S3: Measured vs Modelled OC_{s+c} to TOC ratio, before removing unrealistic data (i.e., data with OC_{s+c} to TOC ratio < 0.5). The final figure is presented in Figure 3b of the main manuscript.

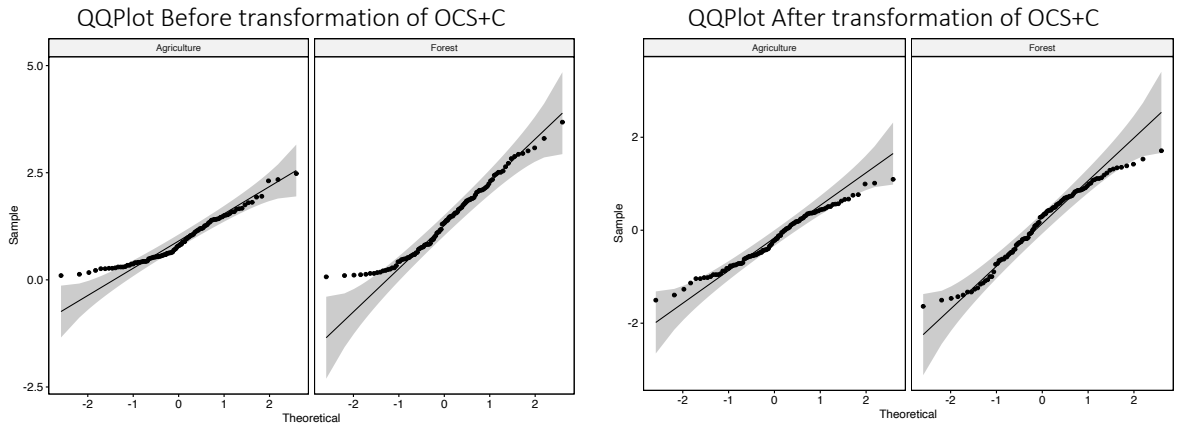


Figure S4: Q-Q plots for the ANOVA on land use effect, before (left) and after (right) box-cox transformation with λ = 0.4.

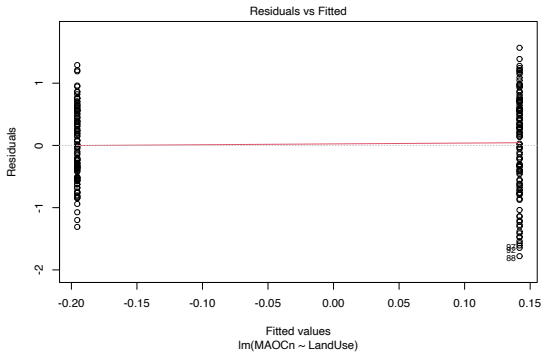


Figure S5: Residuals vs fitted values of a linear model with OC_{S+C} (i.e. MAOC) as dependent variable and land use as independent variable. There is no sign of violation of homogeneity of variance assumption, as no evident relationship between residuals and fitted value is present.

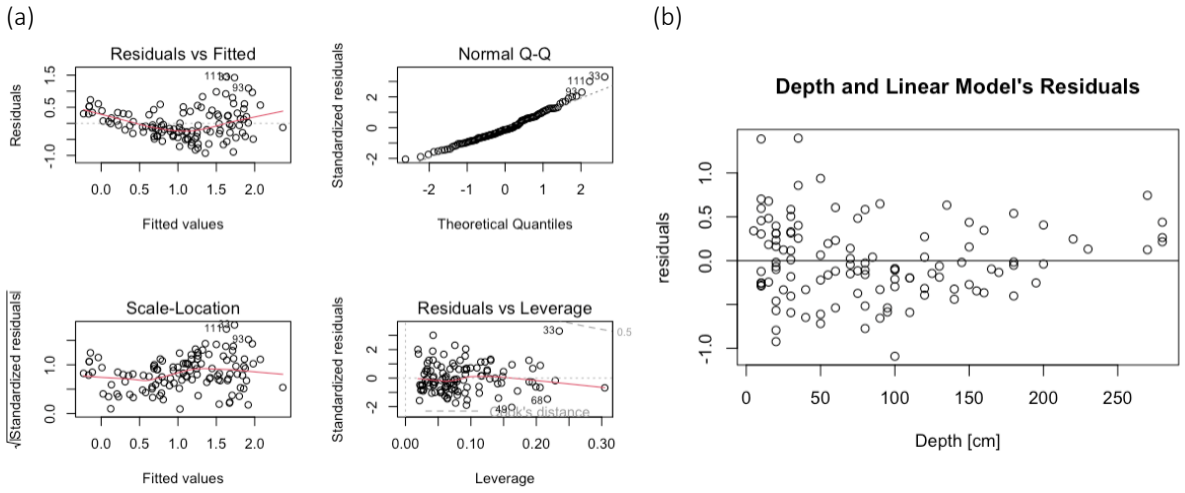


Figure S6: Regression diagnostic plots of the linear model of OC_{S+C} that was fitted before to do the GAMM. In this model, all independent variables (i.e. depth, CEC, pH, clay, silt, CIA, slope, and Al^{3+}) were computed without interaction. A clear violation of homogeneity was observed (a). This heterogeneity resulted from the non-linear relationship between the independent variable depth and the dependent variable OC_{S+C}. As the spread of the residuals is rather on the positive side (especially below 2 m depth), the linear model underestimates OC_{S+C} concentration (b).

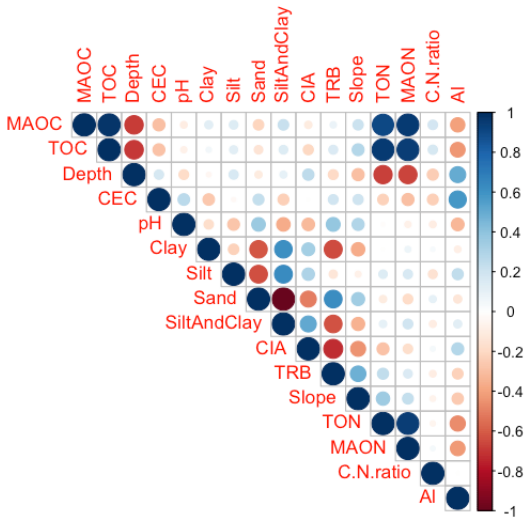


Figure S7: Correlation of all soil characteristics that were measured in the field or in the laboratory (OC_{S+C} is referred here as MAOC).

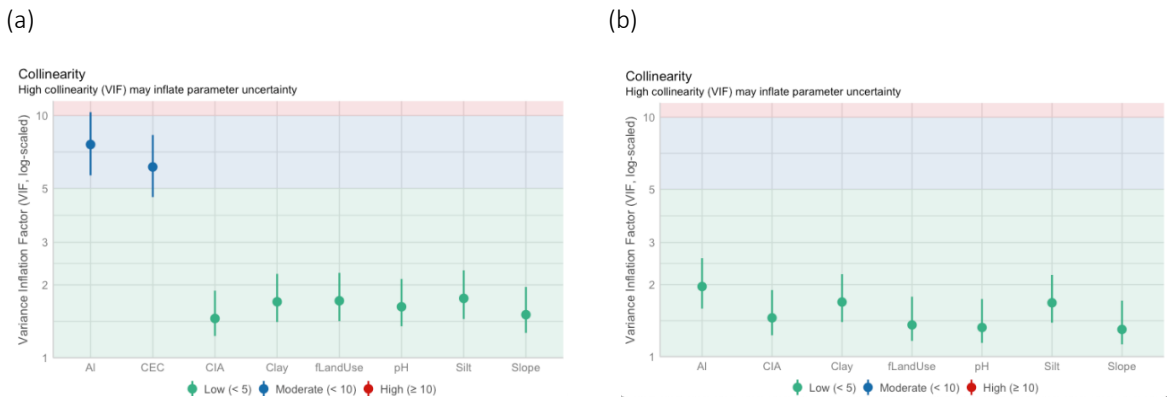


Figure S8: Variance inflation factor (VIF) of the independent variable before (a) and after (b) removing CEC from the analysis. The VIF is a measure of multicollinearity among the independent variables. If an independent variable has a VIF higher than 5, this might affect its significance.

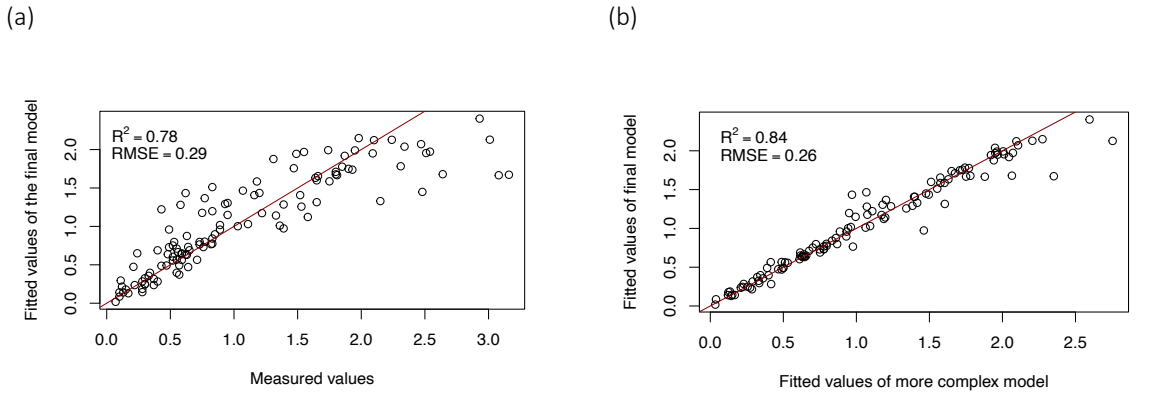


Figure S9: Validation of the general additive mixed effect model for OC_{S+C} concentration a) Fitted values of the selected model vs measured data. b) Fitted values of a more complex model vs fitted values of the selected model.

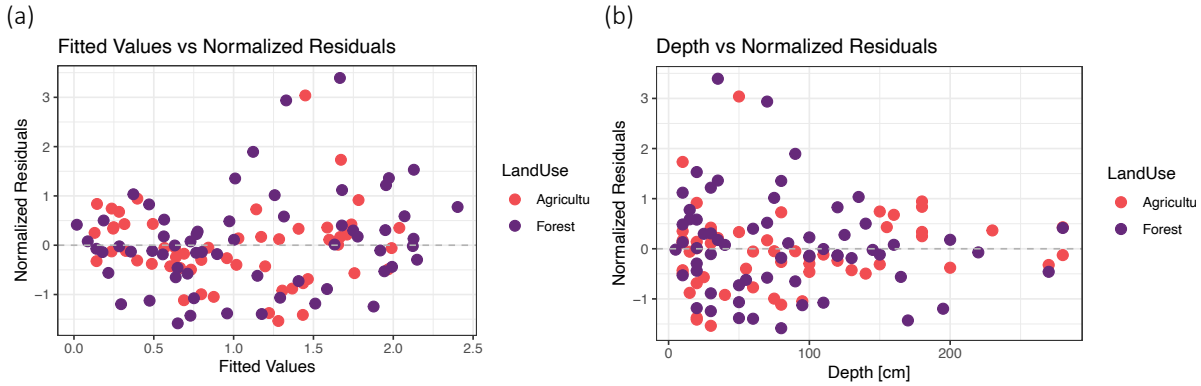


Figure S10: a) Fitted values vs residuals of the final model (i.e. the GAMM model, which includes only the selected controlling factors). There is no indication of a violation of the assumption the residuals are normally distributed. b) Independent variable depth and the model's residuals. The relationship between depth and OC_{s+c} is not linear and this is taken into account in the model. Therefore, the selected model does not underestimate the OC_{S+C} of the lower depth like the linear regression.

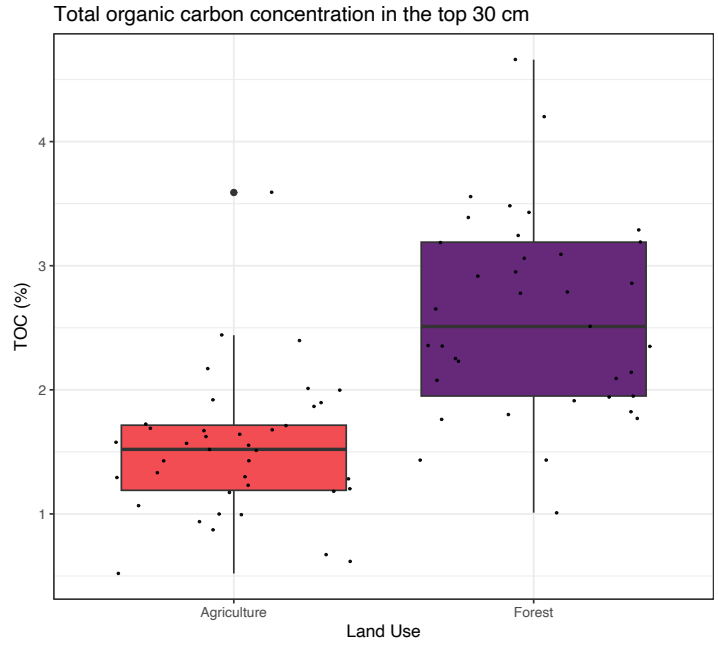
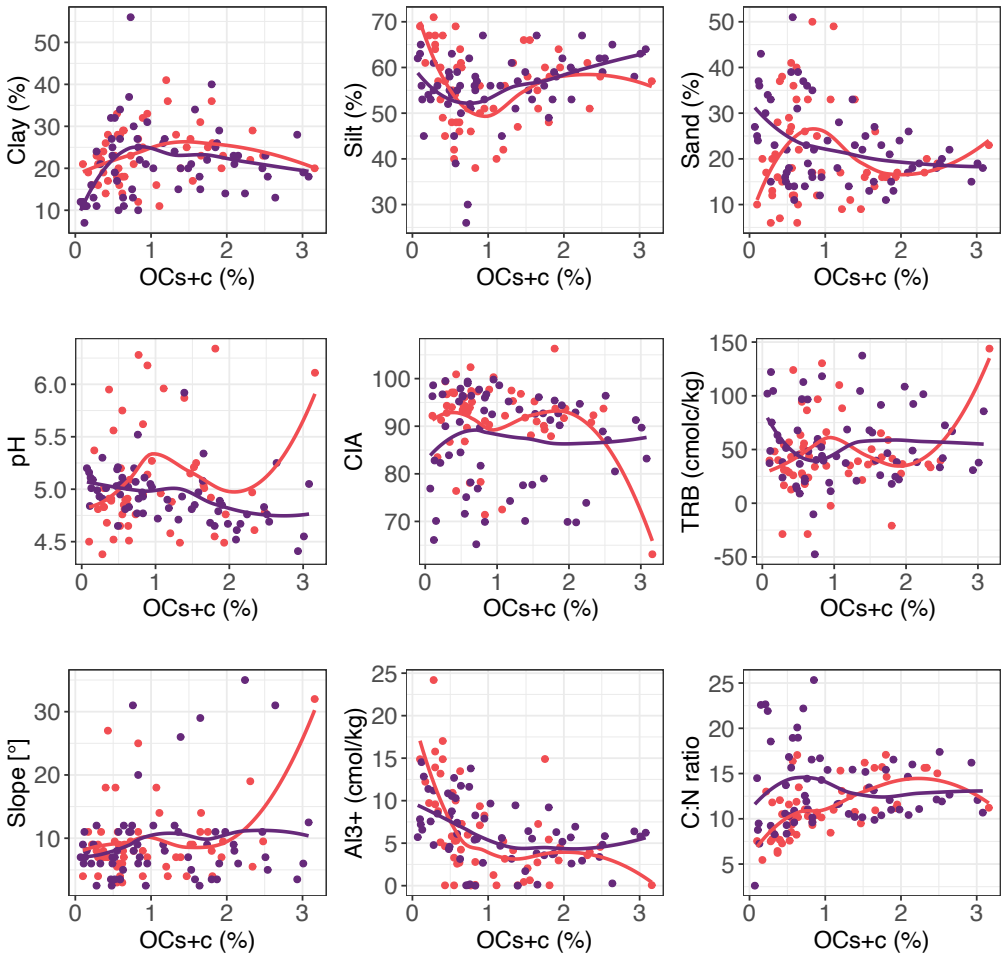


Figure S11: Topsoil (i.e. > 30 cm) total OC concentration using all data (measured and modelled)



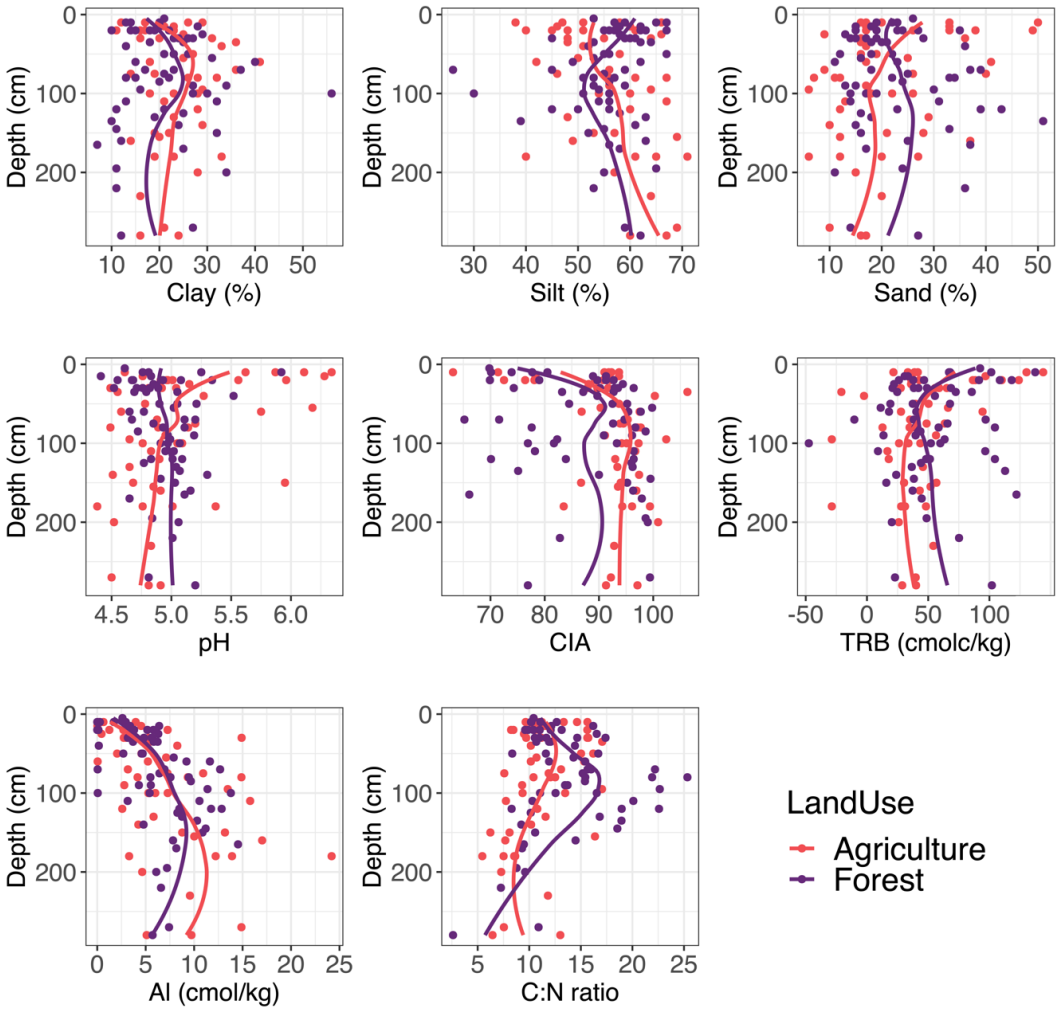


Figure S12: Soil Characteristics, OC_{s+c} concentration and Depth profile with loess smooth

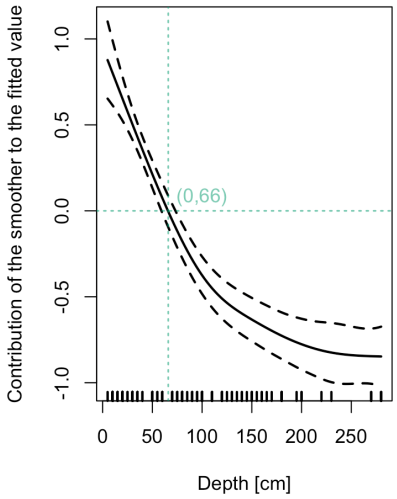


Figure S13 : Contribution of the smoother for depth to the fitted values
The unit of the contribution of the smoother is the same as the unit of the fitted value. Here it represents the percentage of carbon that is added to the regression curve that did not yet accounted for depth at each different depth.

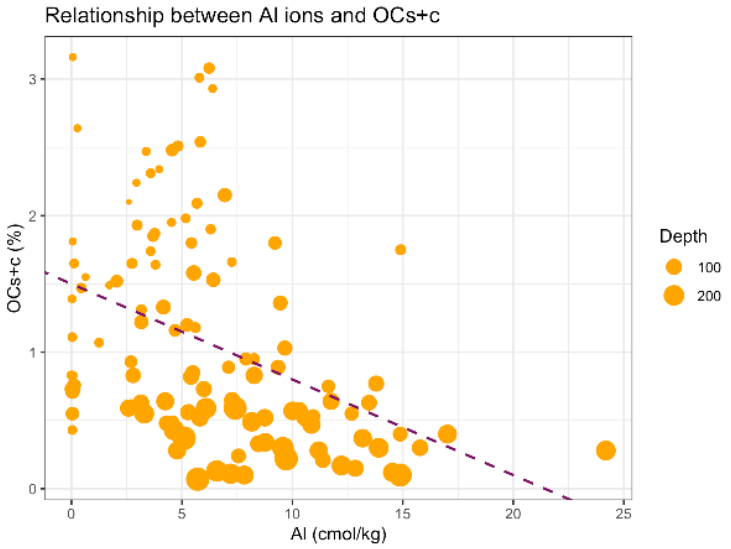


Figure S14: Relationship between Al^{3+}, OC_{S+C} concentration and soil depth.

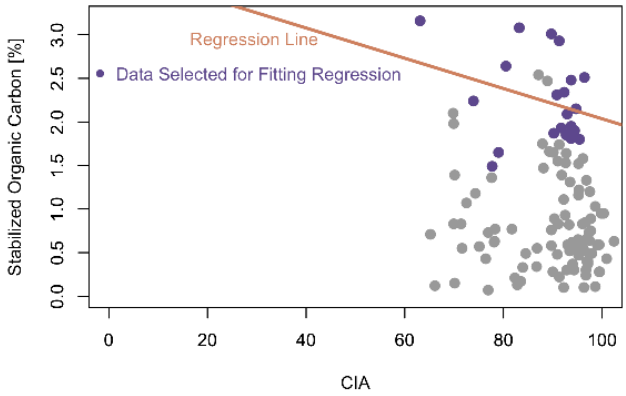


Figure S15: Relationship between CIA and OC_{S+C} concentration and regression line of the highest OC_{S+C} concentration for different ranges of CIA values.

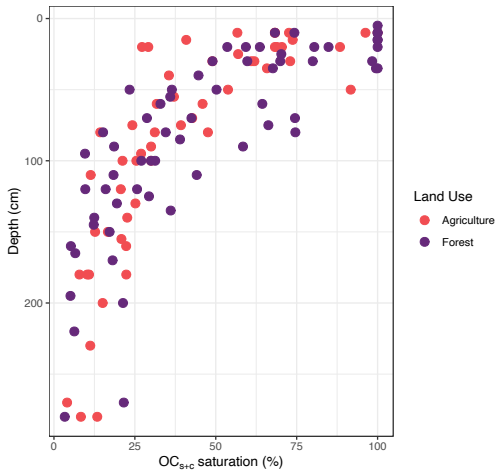


Figure S16: Soil profiles of OC_{S+C} saturation level. Below respectively 0.5 m and 1 m depth, none of the forest soils reached an OC_{S+C} saturation higher than 75% and 50%.

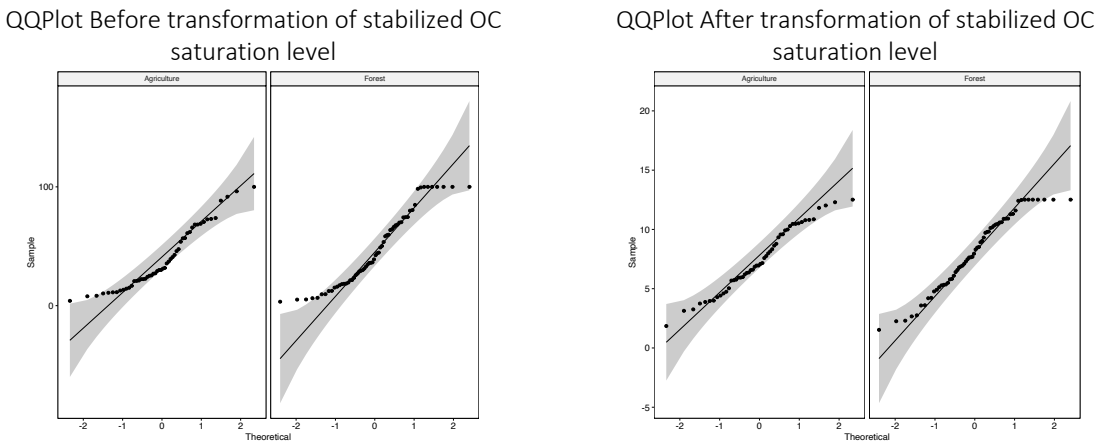


Figure S17: Q-Q plots for the ANOVA on land use effect, before (left) and after (right) box-cox transformation with λ = 0.38.

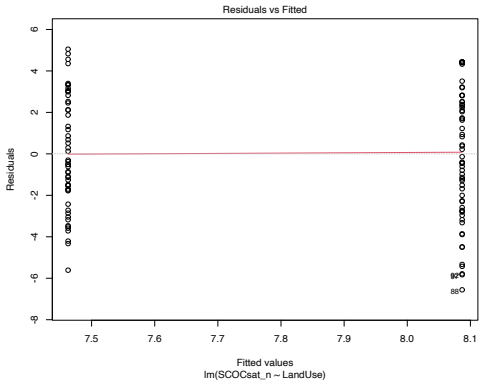


Figure S18: Residuals vs fitted values of a linear model with stabilized OC saturation level (i.e. SCOCsat_n) as dependent variable and land use as independent variable. There is no sign of violation of homogeneity of variance assumption, as no evident relationship between residuals and fitted value is present.

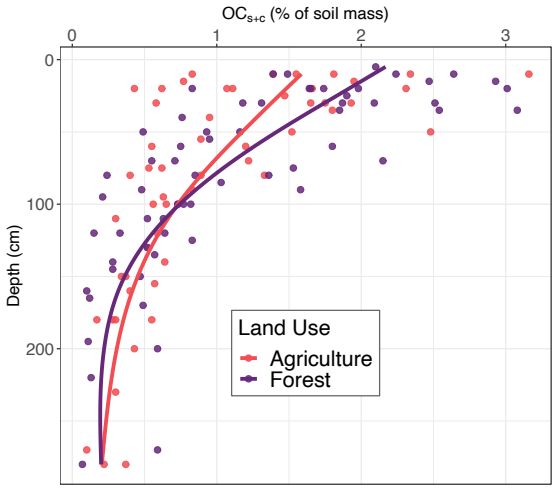


Figure S19: OC_{S+C} concentration depth profiles with measured samples only

Table S1:
Summary of the ANOVA on OC_{s+c} concentration in forest and agriculture for different depth layers

Depth layer start at [cm] (included)	Depth layer finishes at [cm] (not included)	Counts Total	p-Value	Significant	Lower Confidence Limit	Upper Confidence Limit	Counts Forest	Counts Agriculture
0	20	40	0	Yes	0.49	0.96	19	21
15	25	42	0	Yes	0.36	0.93	20	22
20	30	35	<0.05*	Yes	0.29	0.91	18	17
25	50	40	<0.05*	Yes	0.35	0.94	18	22
30	60	41	<0.05*	Yes	0.08	0.75	21	20
35	70	38	<0.05*	Yes	0.19	0.93	18	20
40	80	41	<0.05*	Yes	0.16	0.84	18	23
45	90	42	<0.05*	Yes	0.02	0.72	19	23
55	100	40	0.15	No	-0.1	0.6	20	20
65	110	40	0.13	No	-0.08	0.59	20	20
70	115	42	0.18	No	-0.1	0.53	21	21
75	130	40	0.38	No	-0.17	0.45	23	17
80	140	42	0.49	No	-0.19	0.39	25	17
85	150	39	0.54	No	-0.21	0.39	24	15
90	155	42	0.65	No	-0.22	0.34	25	17
95	180	42	0.97	No	-0.26	0.25	25	17
100	190	42	0.84	No	-0.21	0.26	22	20
110	270	42	0.73	No	-0.29	0.2	24	18
115	280	41	0.84	No	-0.29	0.23	23	18
125	300	40	0.57	No	-0.34	0.19	21	19

Table S2: AIC and BIC of the 4 models that were computed

	One depth smoother for both land uses	Two different smoother, one for each land use
Without power variance function	AIC: 333.76 BIC: 401.85	AIC: 343.02 BIC: 415.83
With power variance function	AIC: 323.71 BIC: 394.32	AIC: 333.45 BIC: 408.78

Table S3: Time since deforestation for all agricultural sites

Site ID	Time (years)	Site ID	Time (years)	Site ID	Time (years)	Site ID	Time (years)	Site ID	Time (years)
A1	>30	A11	50	A21	70	A31	70	A41	40
A2	90	A12	50	A22	50	A32	50	A42	>30
A3	90	A13	100	A23	>30	A33	>30	A43	70
A4	100	A14	>30	A24	>30	A34	60	A44	100
A5	100	A15	>30	A25	30	A35	30	A45	50
A6	100	A16	>30	A26	80	A36	50	A46	100
A7	70	A17	>30	A27	>30	A37	>30	A47	>30
A8	>30	A18	>30	A28	>30	A38	100	A48	70
A9	>30	A19	50	A29	60	A39	>30	A49	50
A10	>30	A20	70	A30	>30	A40	50	A50	>30

Table S4:
Mean of OC_{S+C} concentration in forest and agriculture for different depth layers

Depth layer start at (cm) (included)	Depth layer finishes at (cm) (not included)	Agriculture OC_{S+C} concentration and standard deviation (%)	Forest OC_{S+C} concentration and standard deviation (%)	Loss in OC_{S+C} concentration after deforestation (%)
0	20	1.29 ± 0.42	2.31 ± 0.63	44.16
15	25	1.24 ± 0.45	2.12 ± 0.75	41.51
20	30	1.16 ± 0.49	1.9 ± 0.61	38.95
25	50	1.07 ± 0.5	1.82 ± 0.54	41.21
30	60	1.15 ± 0.6	1.62 ± 0.65	29.01
35	70	0.95 ± 0.59	1.54 ± 0.67	38.31
40	80	0.83 ± 0.56	1.3 ± 0.54	36.15
45	90	0.86 ± 0.57	1.19 ± 0.55	27.73
55	100	0.75 ± 0.4	1 ± 0.54	N.S.
65	110	0.7 ± 0.4	0.94 ± 0.51	N.S.
70	115	0.68 ± 0.39	0.88 ± 0.49	N.S.
75	130	0.64 ± 0.33	0.78 ± 0.42	N.S.
80	140	0.64 ± 0.74	0.74 ± 0.38	N.S.
85	150	0.6 ± 0.7	0.7 ± 0.37	N.S.
90	155	0.58 ± 0.65	0.65 ± 0.37	N.S.
95	180	0.5 ± 0.54	0.54 ± 0.32	N.S.
100	190	0.46 ± 0.17	0.5 ± 0.27	N.S.
110	270	0.42 ± 0.16	0.42 ± 0.26	N.S.
115	280	0.41 ± 0.42	0.42 ± 0.27	N.S.
125	300	0.39 ± 0.17	0.38 ± 0.27	N.S.

Table S5:
Mean of total OC concentration in forest and agriculture for different depth layers (down to 90 cm)

Depth layer start at (cm) (included)	Depth layer finishes at (cm) (not included)	Agriculture TOC concentration and standard deviation (%)	Forest TOC concentration and standard deviation (%)	Loss in TOC concentration after deforestation (%)
0	20	1.59 ± 0.38	2.94 ± 0.75	45.92
15	25	1.52 ± 0.48	2.53 ± 0.88	39.92
20	30	1.32 ± 0.51	2.23 ± 0.68	40.81
25	50	1.23 ± 0.4	2.03 ± 0.65	39.41
30	60	1.29 ± 0.52	1.78 ± 0.73	27.53
35	70	1.07 ± 0.55	1.65 ± 0.75	35.15
40	80	0.94 ± 0.53	1.38 ± 0.54	31.88
45	90	0.91 ± 0.56	1.29 ± 0.57	29.46

Table S6:
Summary of the ANOVA on stabilized OC saturation in forest and agriculture for different depth layers

Depth layer start at (cm) (included)	Depth layer finishes at (cm) (not included)	p-Value	Significant	Lower Confidence Limit	Upper Confidence Limit	Counts Forest	Counts Agriculture
0	30	<0.05*	Yes	0.11	2.47	14	14
0	50	<0.05*	Yes	0.23	2.08	23	19
20	50	<0.05*	Yes	0.02	2.33	16	12
30	90	0.29	No	-0.63	2.03	23	16
30	100	0.42	No	-0.8	1.9	26	18
50	90	0.74	No	-1.3	1.8	14	11
50	100	0.91	No	-1.43	1.59	17	13
90	150	0.95	No	-1.35	1.27	16	9
100	300	0.69	No	-0.83	1.25	22	20