Supplement of Earth Syst. Sci. Data, 18, 2507–2548, 2026
https://doi.org/10.5194/essd-18-2507-2026-supplement

Open Access Earth System Science Data

Supplement of

Soil information and soil property maps for the Kurdistan region, Dohuk governorate (Iraq)

Mathias Bellat et al.

Correspondence to: Mathias Bellat (mathias.bellat@uni-tuebingen.de)

Table of contents

List of Figures

List of Tables

RUSLE model protocol

Introduction

Purpose

This RUSLE model is generated to produce a soil erosion map for our study area of the project (based on the EHAS survey). This soil erosion map will later be used for a **conditioned Latin hypercube sampling** (cLHS) to choose sample locations. It was based on the originals equations from Renard et al. (1991).

Procedure

We follow Thapa (2020) procedure for the RUSLE model based on 5 factors: rainfall, soil strength, slope, land use and conservation. We also used Cossart, Fressard, and Chaize (2020); Almagro et al. (2019) and Mehri et al. (2024) as reference work. The DEM has been filled with the SAGA GIS function *Fill Sinks (Wank & Liu); SAGA GIS 7.8.2* version.

Input data

All the data listed below are available freely online.

Table S1: Data used in the production of the RUSLE model.

Name/ID	Original resolution (m)	Type/Unit	Source
DEM/TE.5	25	Meters	ESA and Airbus (2022)
Prec. sum/OT.6	1000	mm	WorldClim2 Fick and Hijmans (2017)
Landuse/OT.4	10	-	Zanaga et al. (2021)
WRB Soils/OT.11	1000	-	HWSD v2.0 FAO and IIASA (2023)

Equations

R Factor

It is the precipitation factor computed from the precipitation raster based on Morgan et al. Morgan, Morgan, and Finney (1984):

$$R = 38.5 + 0.35P$$

For P = precipitation per year in mm.

K Factor

Is the soil strength factor computed from different parameters from soil map based on Kouli et al. Kouli, Soupios, and Vallianatos (2009):

- SAN = Sand in %
- SIL = Silt in %
- CLA = Clay in %
- C = Soil organic carbon in %

$$Fcsand = 0.2 + 0.3exp(-0.0256SAN(1 - \frac{SIL}{100}))$$

$$Fsi - cl = \frac{SIL}{CLA + SIL}^{0.3}$$

$$Forgc = 1 - \frac{0.25C}{C + exp(3.72 - 2.95C)}$$

$$Fhisand = 1 - \frac{0.7(\frac{SAND-1}{100})}{(\frac{SAND-1}{100}) + exp(-5.51 + 22.9(\frac{SAN-1}{100}))}$$

$$K = Fcsand * Fsi - cl * Forgc * Fhisand * 0.1317$$

LS factor

The slope length and steepness factor was computed from the DEM under *SAGA GIS 7.8.2*. We used the *LSFactor* single way with the Desmet and Govers 1996 method.

For more detail of the LS parameters:

- S_j = Slope factor for the j-th segment.
- λ_j = Distance from the lower boundary of the j-th segment to the upslope (m).
- m = length exponent of the RUSLE Ls factor.

$$LSj = \frac{Sj(\lambda_j{}^{m+1} - \lambda \frac{m+1}{j-1})}{(\lambda_j - \lambda_{j-1})22.13^m}$$

C factor

It was based on the land cover map of Copernicus data set and divided into 6 different classes. Then, based on a literature review, a ratio was established (Morgan 2005; Swarnkar et al. 2018).

Table S2: Ratio of C factor between different land uses.

Land use	Value
Water	0
Forest	0.001
Shrubland	0.01
Urban Area	0.1
Cropland	0.2
Wasteland	1

P factor

P is the conservation factor value. It is difficult to establish in the region as no previous work has been done. As no value was known for this factor, we put a 1 value default as in Mehri et al. (2024).

Results

The erosion map shows a low erosion at Mg/ha-1 year-1 in terms of quantity. Most of the values are under 1.7, and only a few are above it. However, with a quantile classification (not based on continuous classification), we were able to spot the area with more erosion. It is mainly an area on the south of the Be'khair anticline in the Swaka Tika anticline and the upper region of Zakho (under the anticline). The northern and western areas are more preserved from soil erosion as the high anticline is probably due to less land use and forest presence. When the slope factor mainly influences the wadi and river border, the precipitation factor strongly influences the upper Khabur valley. The soil factor seems to have a minor influence on the middle and upper Khabur valley and on the right bank of the Tigris. The most preserved areas are the river and Mosul Lake.

Table S3: Basic statistics for the erosion map based on RUSLE model.

Statistic	Value (Mg/ha-1 year-1)
Min. value	0
Max. value	117
Mean value	1.17561
SD	3.1574

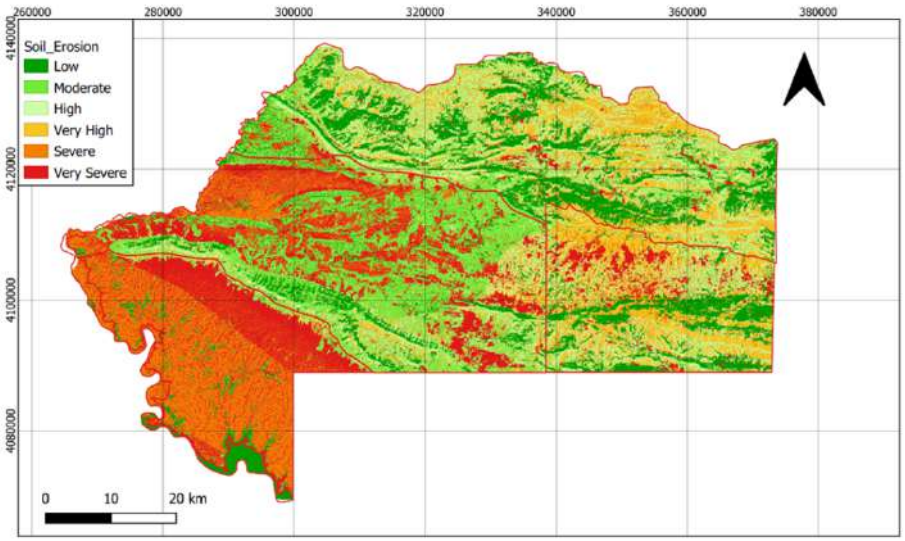

Figure S1: RUSLE map of erosion with breaks calculated by quantiles (WGS84, UTM38N).

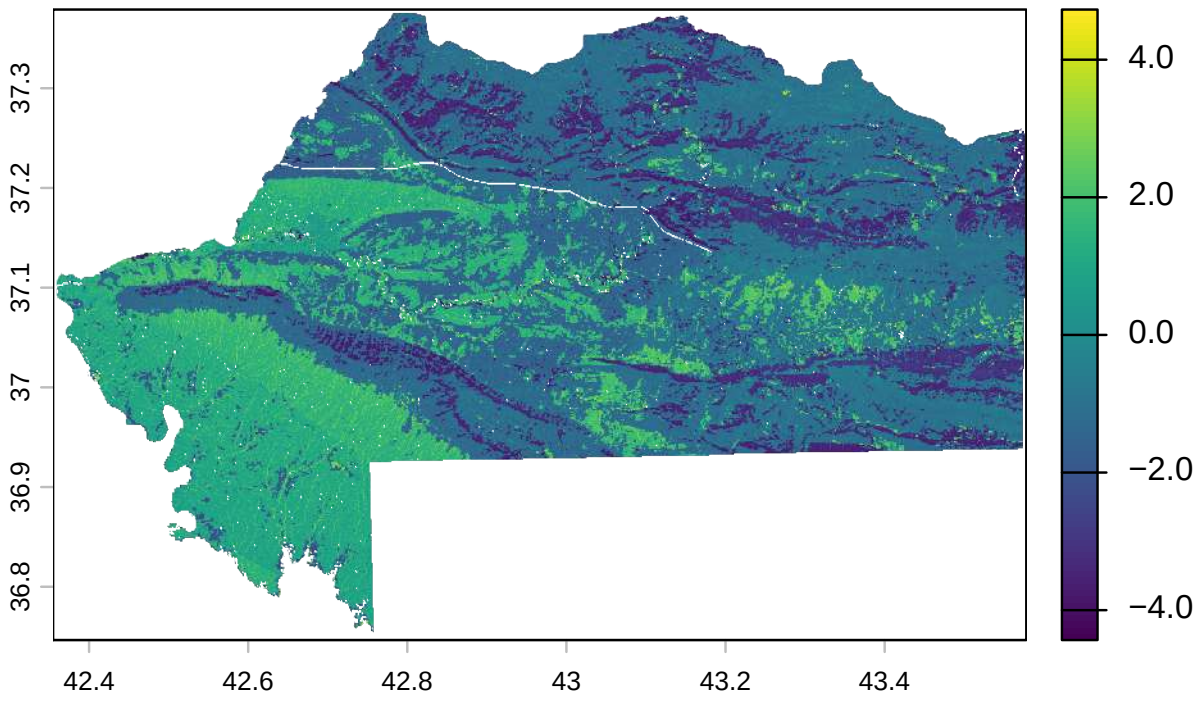

Figure S2: RUSLE map of erosion in Mg/ha-1 year-1 (WGS84, UTM38N).

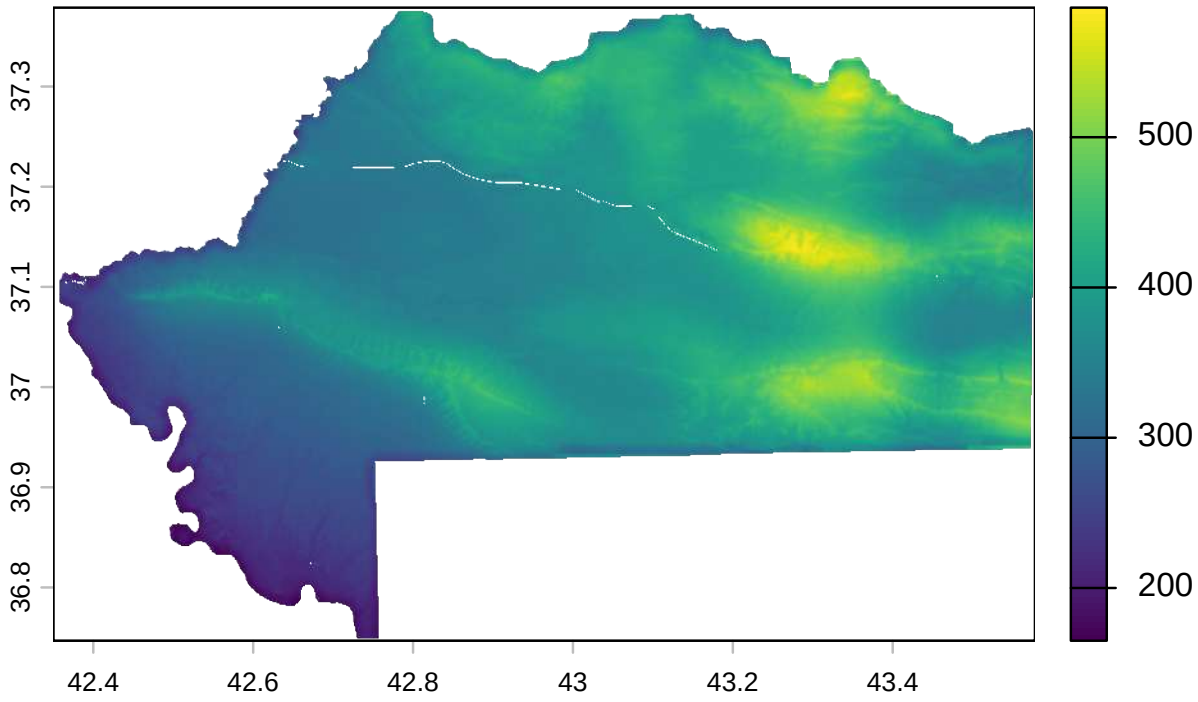

Figure S3: Rainfall (R) factor (WGS84, UTM38N).

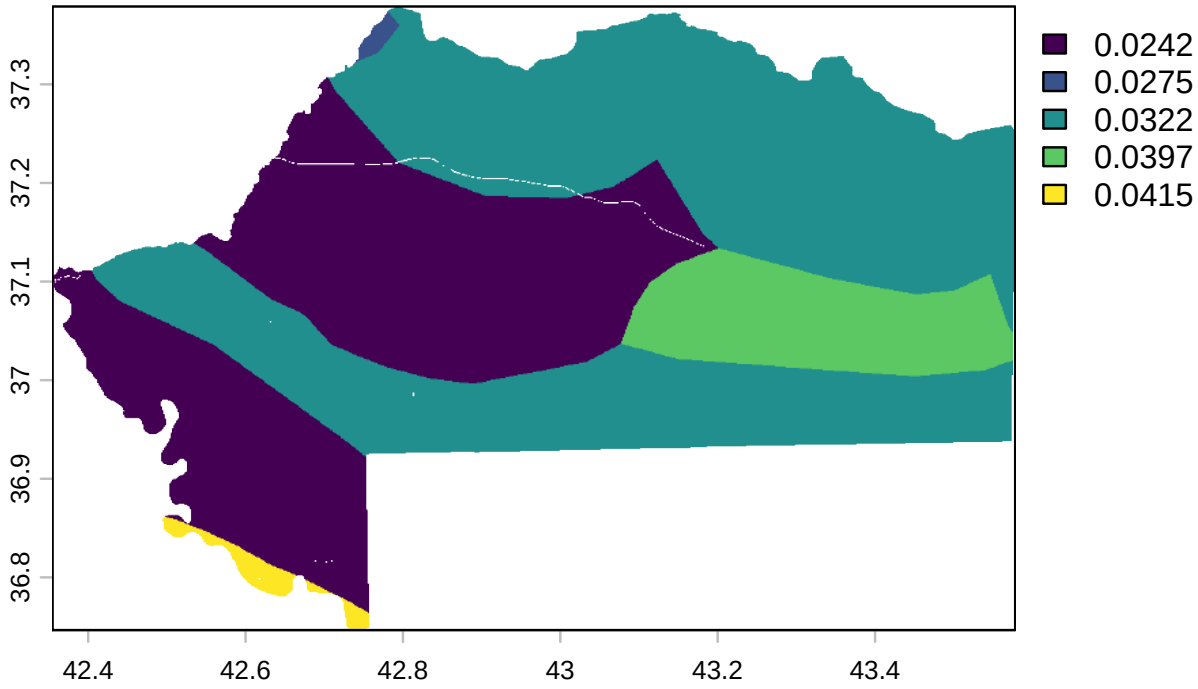

Figure S4: Soil strength (K) factor (WGS84, UTM38N).

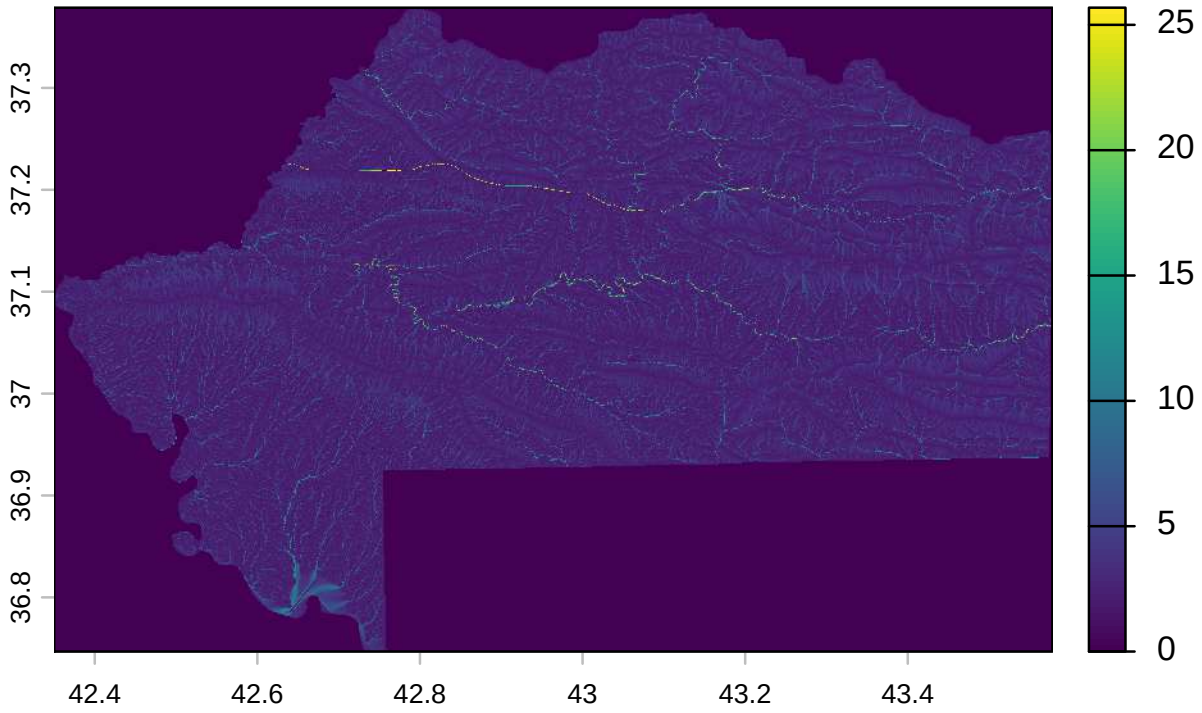

Figure S5: Slope length and steepness (LS) factor (WGS84, UTM38N).

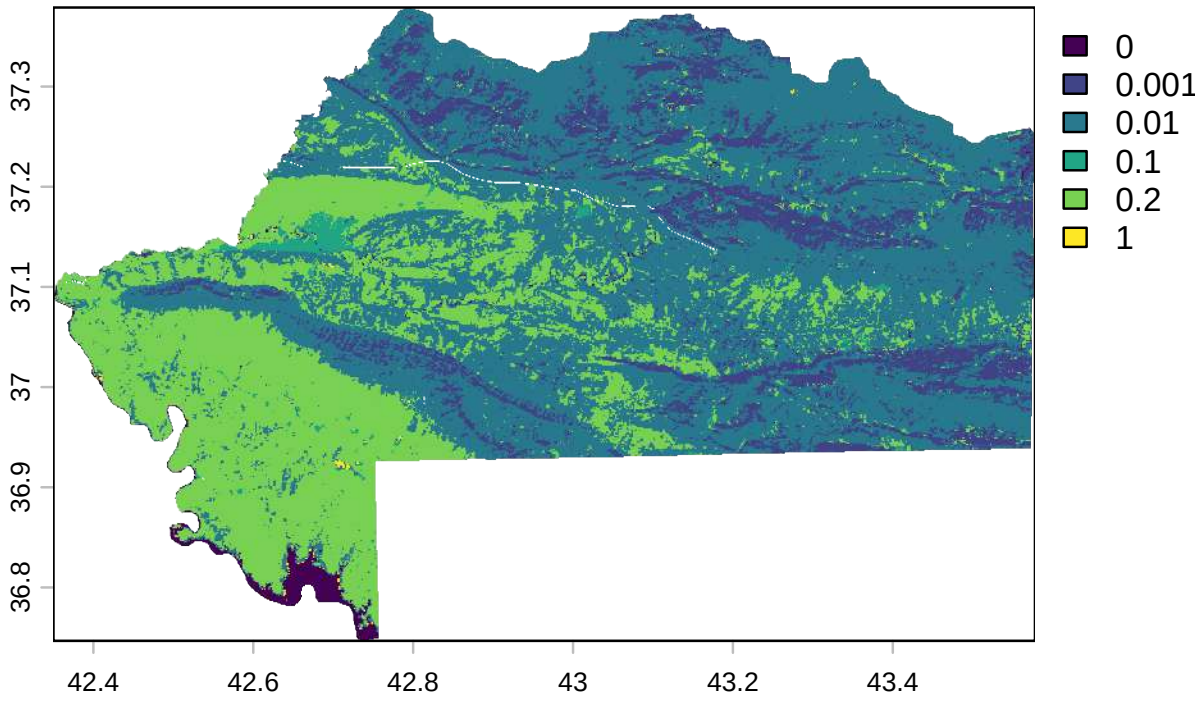

Figure S6: Land uses (C) factor (WGS84, UTM38N).

Conditioned Latin hypercube sampling

Introduction

Purpose

This document presents the results of the sampling design for the 2022 and 2023 summer field campaigns of the B07 subproject in *CRC1070*. We used a specific sampling technique called conditioned Latin hypercube sampling (cLHS). This sampling strategy based on combinations of different *strata*/layers gave a diversity of samplings according to the heterogeneity of the layers. We tried different numbers of sampling locations (100, 150, 200, 250, 300, 350 and 400) to finally agree on a set of 100 samplings (We majored the number of samples by 10% - to reach 110 locations - to avoid any issue on the file in case of non-accessible area). The study area covers over 830 km^2 in 2022 and 490 km^2 in 2023. The size of area C explored in 2023 is equal to 1450 km^2. Still, due to inaccessible places in high mountains or conflict zones between PKK (Kurdistan Workers' Party) and the Turkish army (Ertan 2022), we reduce the explored area.

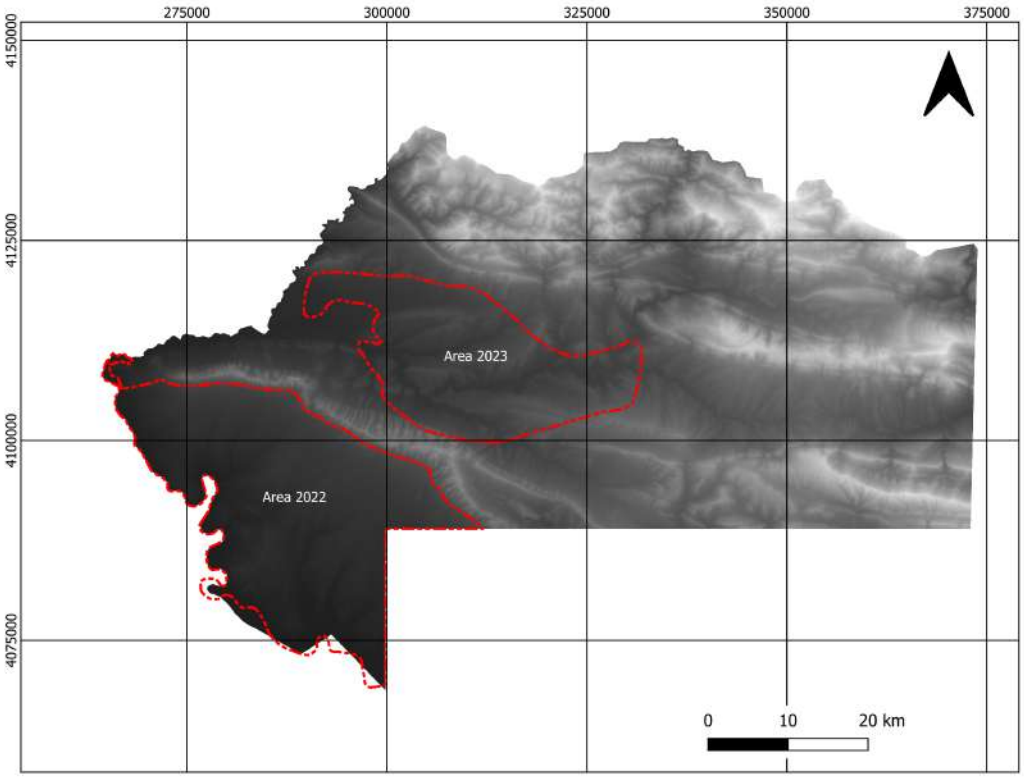

Figure S7: Sampling areas for 2022 and 2023.

Procedure

We first produced several maps composed of different factors before using them as layers for the conditioned Latin hypercube sampling. We used different software such as *GIS 3.24.1* version and *SAGA GIS 7.8.2*, *GoogleEarthEngine* (GEE) and *R 4.4* with *Rstudio*. The procedure is partially based on the Sampling package from Dick Brus and its reference book (Brus 2022).

Input data

All the data listed below are available freely online excepted for the digitized maps (geomorphology).

Table S4: Data used in the production of the conditioned Latin hypercube sampling.

Name/ID	Original resolution (m)	Type/Unit	Source
DEM/TE.5	25	Meters	ESA and Airbus (2022)
RUSLE map/OT.11	25	Mg/ha^{-1} $year^{-1}$	Mathias Bellat
Slope/TE.16	25	Radians	ESA and Airbus (2022)
TWI/TE.22	25	-	ESA and Airbus (2022)
Soil estimation/OT.12	10	-	Nafiseh Kakhani / USGS (2018)
Geomorphology/OT.3	1 : 250 000	-	Forti et al. (2021)
Landsat 8 Green/LA.3	30	0.53 - 0.59 µm	USGS (2018)
Landsat 8 Red/LA.4	30	0.64 - 0.67 µm	USGS (2018)
Landsat 8 NIR/LA.5	30	0.85 - 0.88 µm	USGS (2018)
Landsat 8 SWIR1/LA.6	30	1.57 - 1.65 µm	USGS (2018)
Landsat 8 SWIR2/LA.7	30	2.11 - 2.29 µm	USGS (2018)

Data treatement

Soil prediction map

The Landsat 8 images were collected *via* a Google Earth Engine script on a period covering 2020, the median of the composite image from *Tier 1 TOA* collection was used. To produce the soil estimation map we based our prediction on a Google Earth Engine written script. The javascript codes for scraping these images and computing the soil estimation indexes are available in the supplementary file inside the "2 - CLHS/code" folder. We computed four soil indexes:

$$Landsat\ 8\ Clay\ index = \frac{SWIR1}{SWIR2}$$

Hengl (2007), USDA/GO (former GTAC)

$$Landsat\ 8\ Ferrous\ minerals\ ratio = \frac{SWIR1}{NIR}$$

Al-Mousawi, Fouad, and Sissakian (2007), Henrich et al. (2012) (Ferric Oxides), USDA/GO (former GTAC)

$$Landsat\ 8\ Carbonate\ index = \frac{Red - Green}{Red + Green}$$

Pouget et al. (1991) (Color, Hematite/ goethite), USDA/GO (former GTAC)

$$Landsat\ 8\ Rocks\ outcrop\ index = \frac{SWIR1 - Green}{SWIR1 + Green}$$

USDA/GO (former GTAC)

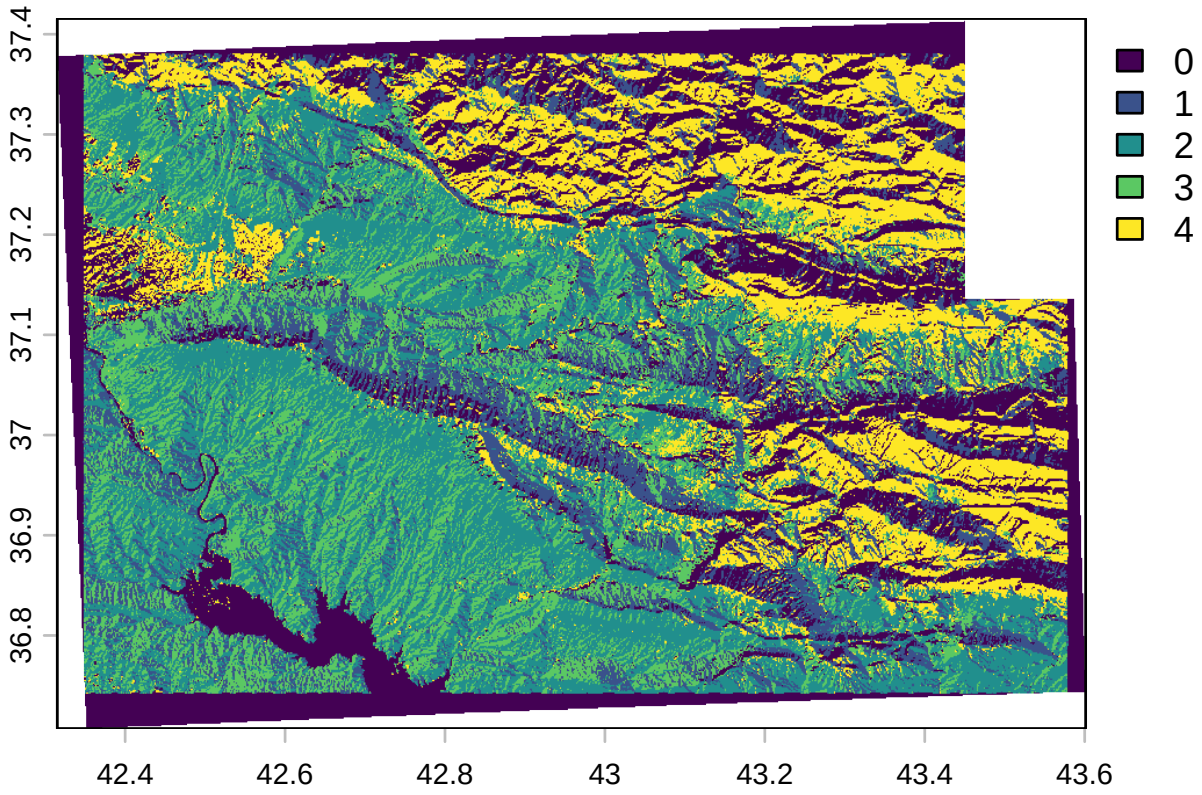

Figure S8: Soil prediction based on indexes (WGS84, UTM38N).

Soil erosion based on RUSLE model

The RUSLE map was previously computed in Chapter 1.

DEM, slope and topographic wetness index

The DEM is on a 30 x 30 m scale based on the ESA data (ESA and Airbus 2022). The slope under *SAGA GIS 7.8.2* with the function *Slope, Aspect, Curvature* with the method by **Zevenbergen, L.W., Thorne, C.R. (1987)**. For the topographic wetness index the *Topographic Wetness Index (One step)* command was used with *Multiflow direction* parameters

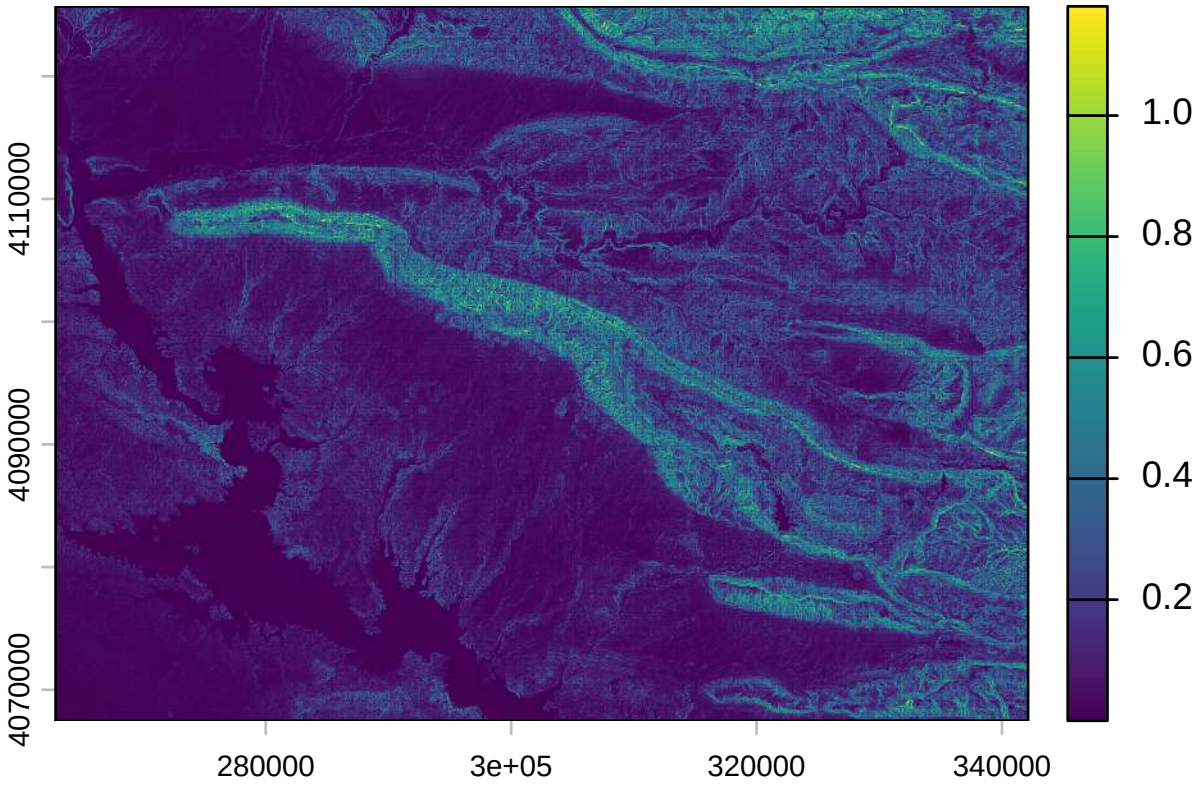

Figure S9: Slope (radians) based on the DEM (WGS84, UTM38N).

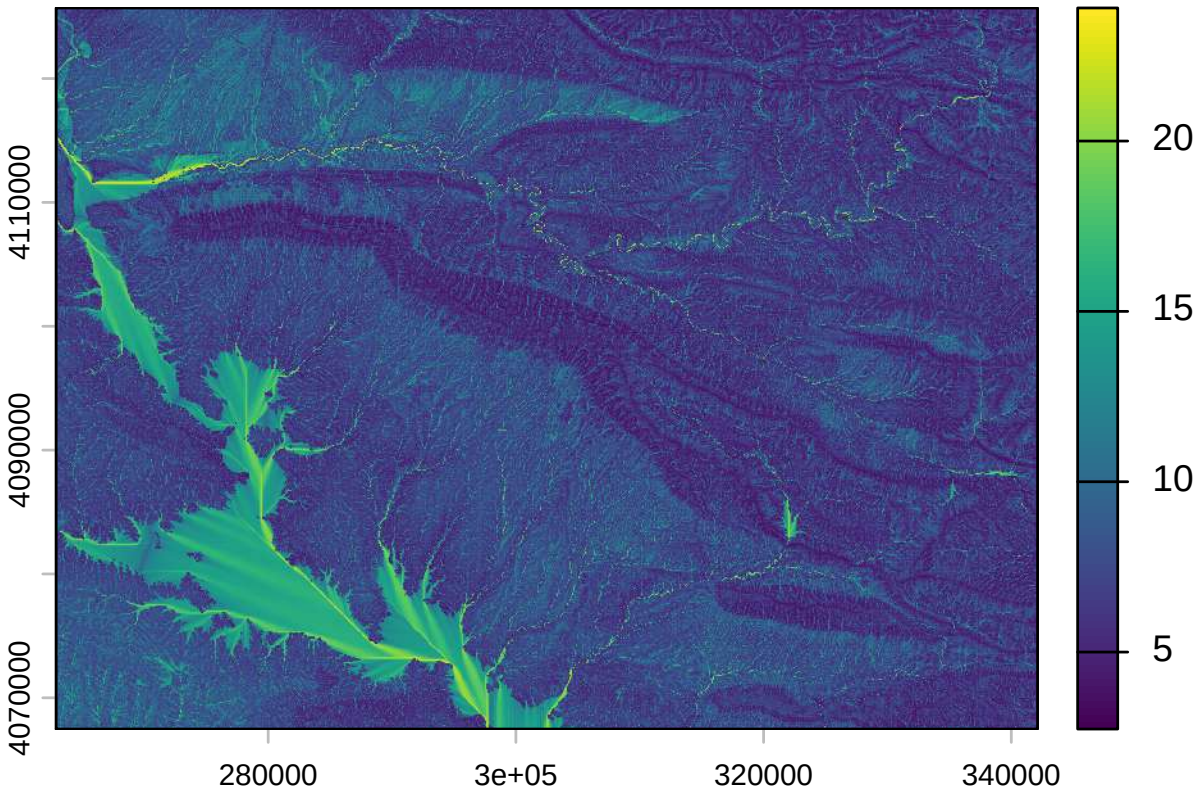

Figure S10: Topographic wetness index based on the DEM (WGS84, UTM38N).

Geomorphological map

The geomorphological map is computed based on (Forti et al. 2021). We created 17 different layers for each of the geomorphons present.

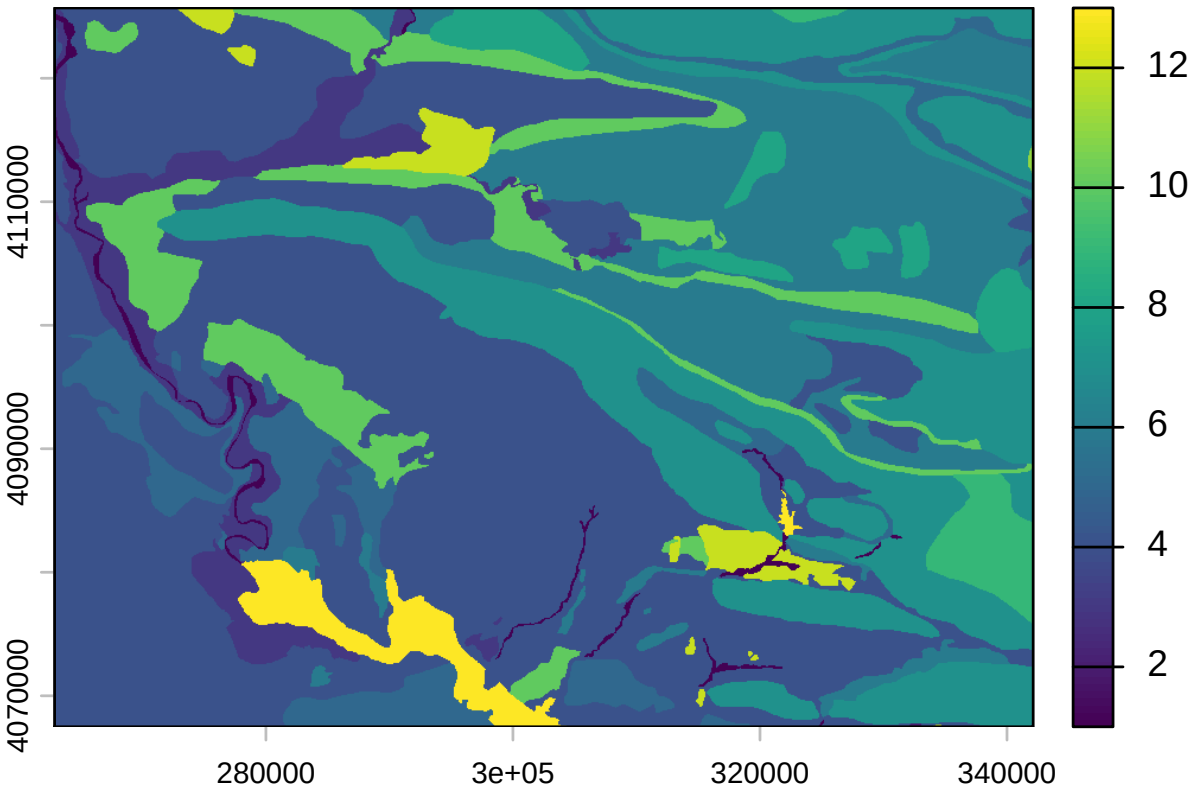

Figure S11: Geomorphological map (WGS84, UTM38N).

Conditioned Latin hypercube sampling script

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Sampling locations

Some sampling locations were located in inaccessible areas so not all the spots have been sampled. Only **101** in 2022 and **21** in 2023. Despite this lack of few samples sites, the distribution of the cLHS variables was similar to better than a classical random sampling method.

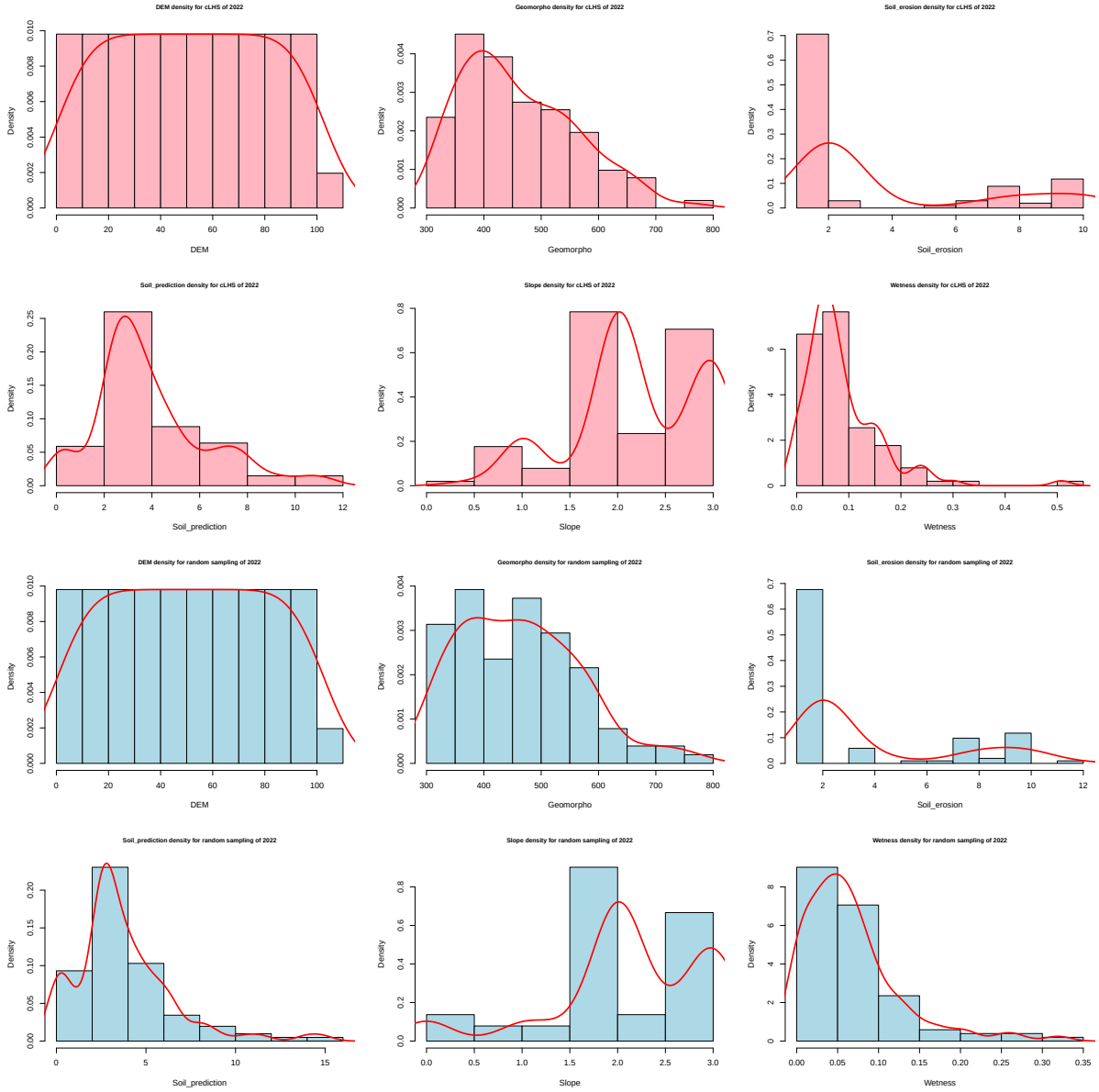

Figure S12: 2022 Campaign samples distribution.

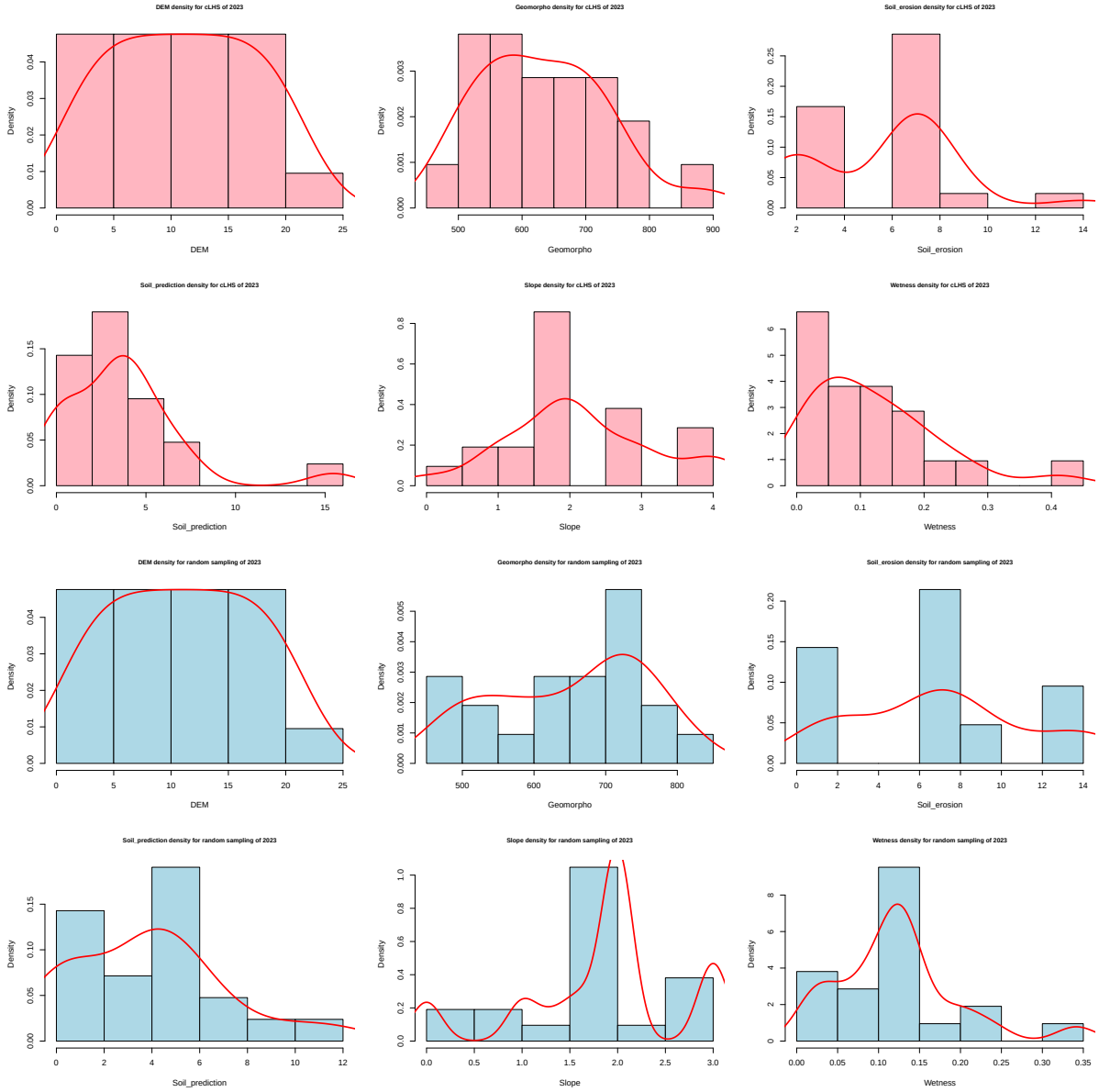

Figure S13: 2023 Campaign samples distribution.

Field information and sampling locations

Sites sampled

2017 and 2018 campaigns

During the EHAS survey of summer 2017 and 2028 P. Sconzo and T. Rentschler collected and analysed **29** samples on **16** different sites. These sampling sites are matching with archaeologists sites from the EAHS survey. They were selected based on a legacy sampling strategy on different part of the slope of the archaeological sites.

2022 and 2023 campaigns

During the two sampling campaigns of 2022-2023, a total of **124** sites were collected *via* the use of a conditioned Latin hypercube sampling.

Regional information and collected data

The collected information is only for the 2022 and 2023 campaigns, as the EHAS survey in 2017 and 2018 had a different sampling protocol only sampling surface soil (0 - 10 cm).

Landscape observations

Zakho region

The sampling campaigns took place in three distinct areas. The first one in the north reaches from Zakho in the west toward Mangesh in the east. It has a semi-wet to dry climate with over 400 mm per year precipitations. This region is dominated by pebbly-sandstone (Mukdadiyah formation Al-Mousawi, Fouad, and Sissakian (2007)) and conglomerate formation and colluvial deposits on the surface. Vegetation with dominating oaks (*quercus brantii*) and olive trees. *Wadis* and a major affluent of the Tigris, the *upper Khabur*, flow from east to west.

Simele/Selevani Upper plain

On the south of the Be'khair mountain, from the foothills towards the main east-west road from Dohuk to Dayrabun in the south, the Upper Simele plain forms another geographical unit. Temperature is higher, and precipitation is below 400 mm per year. There is very little vegetation, and no trees except cultivated species along the rivers and wadis. Surface structures are dominated by glacis, pediments and pebble deposits and relics of conglomerates or sandstone and claystone formations (Injana formation Sissakian, Hagopian, and Hasan (1995)).

Simile/Selevani Lower plain

The last region is the southern, or Lower, Simile plain, from the Dohuk Dayrabun road to the Mosul dam lake in the south. Temperatures are high, and precipitation is below 400 mm per year, with no vegetation except along rivers and wadis. Surface structures are dominated by badlands, colluvial deposits, sandstone (Injana formation Sissakian, Hagopian, and Hasan (1995)), claystone formations (Fatha formation Sissakian, Hagopian, and Hasan (1995)), and some terraces.

Sampling method

The samples were collected with the help of an auger and separated in five-depth increment :

- 0 - 10 cm with the top ploughed soil
- 10 - 30 cm the top and sub-surface soil
- 30 - 50 cm the sub-surface soil
- 50 - 70 cm the sub-surface and lower soil
- 70 - 100 cm the lower soil

In many cases, the bedrock was reached before 100 cm. Therefore, five samples were not collected in every site location. Samplings were saved in a small plastic bag before being air-dried for 48h, followed by wet sieving at 2 mm.

Collected informations

At each location, several pieces of information were registered and can be found in the deposit in the *field_observations* file. The different entries are:

- *Site_Name* the site's name according to the cLHS sampling
- *Altitude A.S.L.* the altitude above is the level in meters, measured with *Garmin, GPSMAP 60Cx* with max. 3 m accuracy (depending of satellite coverage)
- *X (WGS 84 UTM 38)* longitude in WGS84 UTM38N projection (epsg:32638), measured with *Garmin, GPSMAP 60Cx* with ± 10 - 3 m accuracy (depending of satellite coverage)
- *Y (WGS 84 UTM 38)* latitude in WGS84 UTM38N projection (epsg:32638), measured with *Garmin, GPSMAP 60Cx* with ± 10 - 3 m accuracy (depending of satellite coverage)
- *Texture* the texture of the different soil horizons
- *Colors* colors of the soils according to the *Munsell soil-color charts* 2009 edition
- *Number_horizon* short description of the different horizons.
- *Description* short description of the entire pedon.
- *Interpretation* is a short interpretation of the entire pedon.
- *Relief* is a short description of the relief surrounding the site.
- *Soil_depth* depth of the soil before bedrock in cm.

Additionally, photos were taken at each location and they are accessible on the deposit.

Fourier-transform infrared spectroscopy of soil samples

Protocol and devices

We analysed all of our **29** samples from 2017-2018 campaigns and **532** from 2022-2023 campaigns with mid-infrared spectroscopy to serve as predictors for the prediction model on the soil properties. These techniques are now well knowned and commonly used in soil science as they allow spare time and money on the different laboratory measurements (Wadoux et al. 2021; Viscarra Rossel et al. 2016; Ng et al. 2022; Ge, Wadoux, and Peng 2022; Stenberg et al. 2010; Bahrami, Danesh, and Bahrami 2022).

Before being analysed with FTIR the samples were ground under 1 µm with a *Pulverisette 5/4, classic line* (Fritsh, Idar-Oberstein, Germany) in a **250 ml** stainless steel hardness container (ISO: X105CrMo17) with **five** 20 mm sintered corrodium (99.7 % Al_2O_3) grinding balls. The settings were set at 350 turns per minute for 12 minutes in total.

The last step before FTIR analyse was to realise lenses from the samples with KBr pressling method. 250 mg of potassium brodime (KBr) and 1-1.3 mg of sample substance are mixed and then loaded in a hydraulic press under vacuum with around 10 - 11 tonnes and pressed for 1-2 minutes to form a transparent tablet, e.g. 10 mm in diameter and 1 mm thick.

This tablet was analysed under mid-infrared Fourier-transform spectroscopy with a *Vertex 80v* (Bruker OPTIK GmbH; Germany) under a control environment. The measurement resolution is 4 cm^{-1} in the interval of 375 - 4500 cm^{-1}, the spectrum is in absorbance and the source is MIR with the optic filter. To calibrate these samples, one control tablet made of 100% of KBr was measured at the beginning of each measurement session.

Raw spectra production

A script was written to produce a raw spectra in absorbance for 375 - 4500 cm^{-1} interval with a 4 cm^{-1} resolution and export it under a .txt format.

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Prepare the spectra regarding state of the art

Interval interferances

The MIR spectra present interference in different intervals such as 375 - 499 cm^{-1} area and 2451 - 2500 cm^{-1} area (Ng et al. 2018).

Outlier values

Some values of the spectra can also present high interference and therefore should be removed or at least be noticed (Ng et al. 2018). This concerned values over 2 or lower than -2. Here we export the different rows containing values over 2.

Convert into different spectra variation

We converted the MIR data into a total of **14** spectra transformations according to literature review (Ludwig et al. 2023; Ng et al. 2018)

The spectra in absorbance

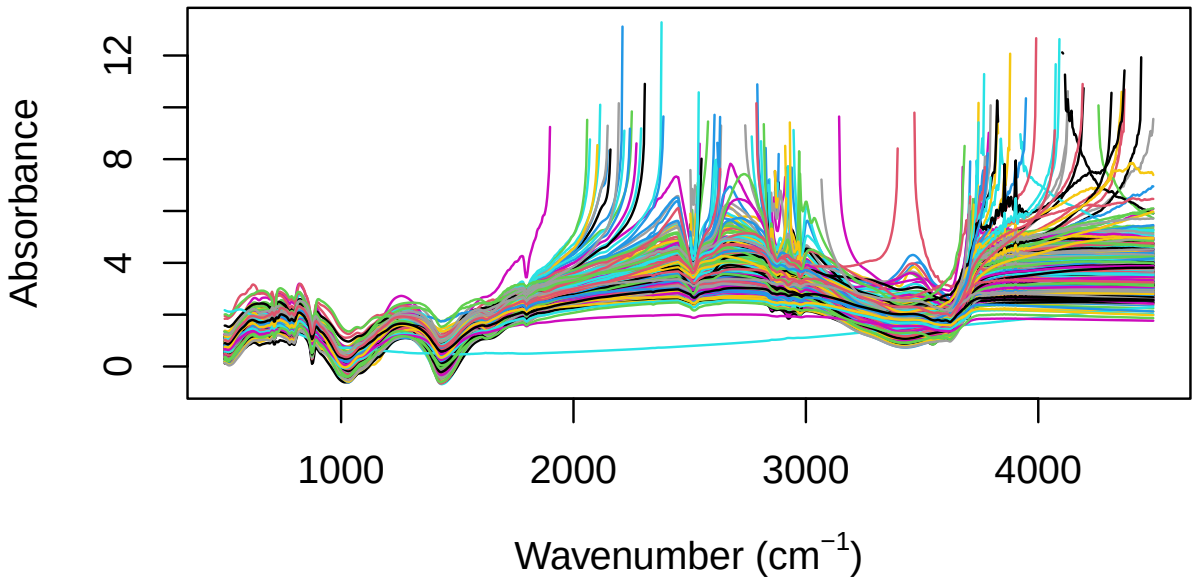

Figure S14: Absorbance spectra in wavenumber.

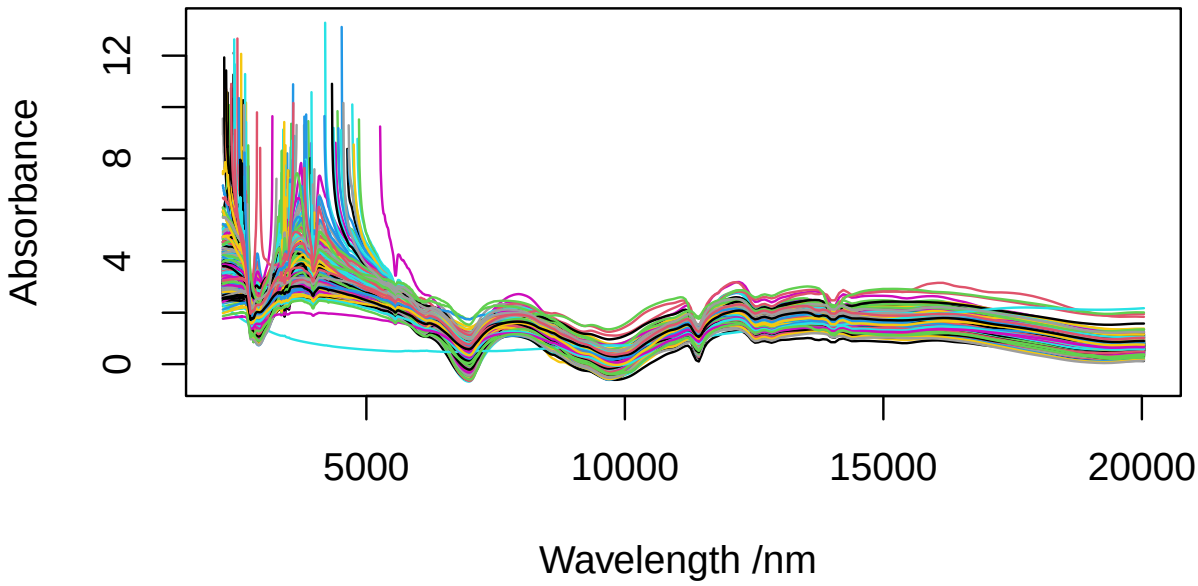

Figure S15: Absorbance spectra in wavelength.

The spectra in reflectance

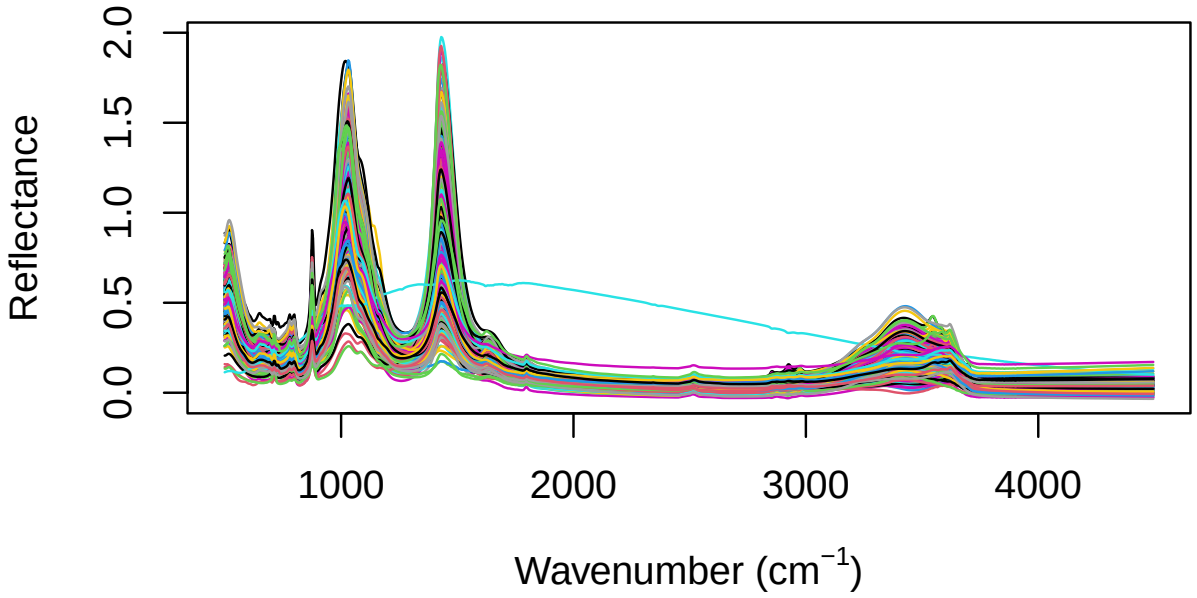

Figure S16: Reflectance spectra in wavenumber.

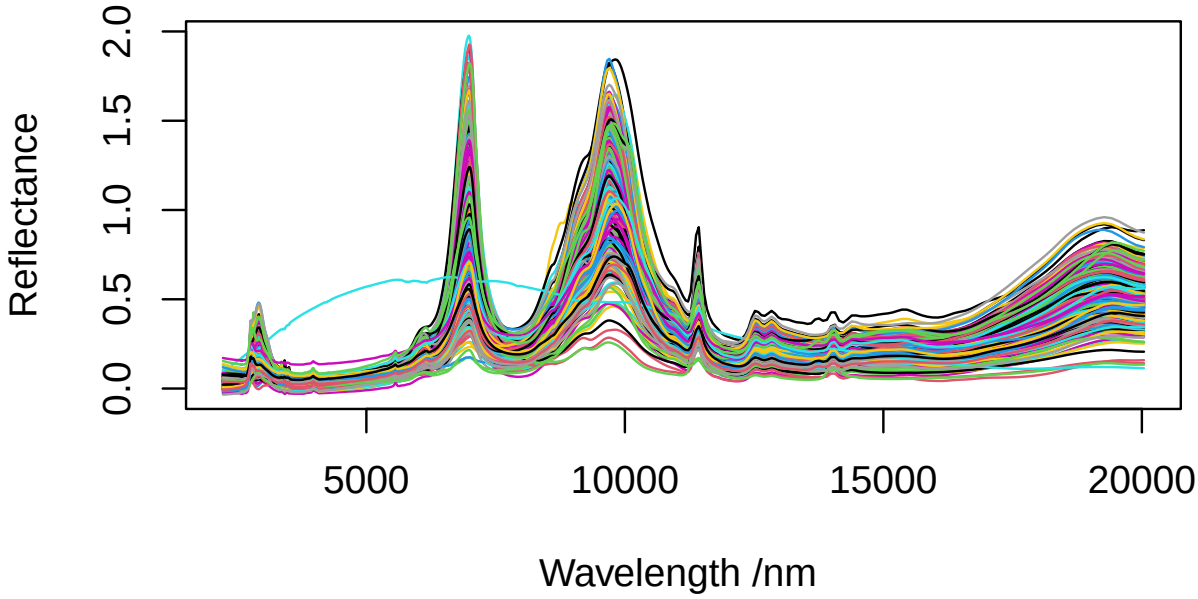

Figure S17: Reflectance spectra in wavelength.

Kennard Stone sampling

The Kennard stone sampling is used to select a wide range of a population that will represent the diversity of the individuals (Kennard and Stone 1969). This sampling strategy has proven to be efficient in the soils spectrometry context (Ramirez-Lopez et al. 2014). We selected different numbers of samples for each depth increment with an over-representation of the topsoil (0 - 10 cm). All the samples from 2017 - 2018 were analysed, so no sampling strategy has been applied to them.

Table S5: Selected samples with Kennard Stone.

Depth (cm)	Year	Number
0 - 10	2022	30
10 - 30	2022	5
30 - 50	2022	5
50 - 70	2022	5
70 - 100	2022	5
0 - 10	2023	10
10 - 30	2023	4
30 - 50	2023	6
50 - 70	2023	6
70 - 100	2023	3

In the selection of samples, you can choose the year and depth increment. Here, we show a selection for the 2022 campaign at 0 - 10 cm depth. The table *Samples_info* use the information collected during the field campaign and the *Field_observations* file. The different entries are:

- **Lab_ID** Laboratory number gave at the samples when enter into TÃ¼bingen Soil Science and Geomrophology laboratory inventory.
- **Site_name** the site's name according to the CLHS sampling.
- **Depth_cm** depth increment of the sampling (0 - 10 - 30 - 50 - 70 - 100 cm).
- **X_WGS84** longitude in WGS84 (epsg:4326), measured from *Garmin, GPSMAP 60Cx* with ± 10 - 3 m accuracy (depending of satellite coverage).
- **Y_WGS84** latitude in WGS84 (epsg:4326), measured from *Garmin, GPSMAP 60Cx* with ± 10 - 3 m accuracy (depending of satellite coverage).

Laboratory measurement of the selected soil samples

Protocol

Following the selection of samples with the Kennard Stone sampling strategy we had a total of **108** samples (29 for 2017 - 2018, 50 for 2022 and 29 for 2023) to measure in the laboratory. First, the samples were air dried (35 - 45°C) for 24 h, root fragments were removed, and sieved (< 2 mm). Additionally, the samples measured with the elementary particles analyser, were ground under 1 µm with a *Pulverisette 5/4, classic line* (Fritsh, Idar-Oberstein, Germany) in a **250 ml** hardness stainless steel container (ISO: X105CrMo17) with **five** 20 mm sintered corrodium (99.7% Al_2O_3) grinding balls.

Properties

All the results from the laboratory measurement are accessible at: https://doi.org/10.1594/PANGAEA.973701.

We measured a total of **8** properties containing chemical, physical and biological attributes:

- **pH** measured with Potassium chloride (KCl) solution with *ProfiLine pH 3310* and a *WTW SenTix 81* pH electrode (Fisher Scientific, Strabourg, France).
- **$CaCO_3$** carbonate calcium calculate in purcent with a calcimeter *08.53* (Royal Eijkelkamp, Giesbeek, Netherlands).
- **Nt** total nitrate calculated in percent with a *Vario EL III* (Elementar, Hanau, Germany).
- **Ct** total carbon calculated in percent with a *Vario EL III* (Elementar, Hanau, Germany).
- **St** total sulfur calculated in percent with a *Vario EL III* (Elementar, Hanau, Germany).
- **OC** total organic calculated in percent with a *Vario EL III* (Elementar, Hanau, Germany). It might be refereed also as **SOC** or **Corg**.

- **EC** electro-conductivity measured in micro-siemens per centimeters with a *Cond 330i/340i* (WTW, Weilheim in Oberbayern, Germany).
- Texture measured in percent either with wet sieving for the sand or with a *SediGraph III 5120* combined with an autosampler *MasterTech MT 052* for lower fractions (Micromeritics, Norcross, USA).

Non mesured samples

Due to the low amount of sample material collected, some measurements were not possible to realise. All the 2017 - 2018 had not enough material (< 5 gr.), after texture analysis, to measure their electro-conductivity and samples **41614** and **53881** had less than 10 gr. collected, which was not enough for texture measurement.

Texture mesurment

The texture was divided texture into diverse classes and we also measured the mean weight diameter (**MWD**) in mm with the following formula.

$$MWD = Xi * Wi/100$$

Where:

- Xi the average diameter in mm.
- Wi the percentage of aggregate.

It can occurs that the total of the measured texture is not perfectly equal to 100%.

Table S6: Texture classes.

Texture class	Diameter in mm
Coarse sand	2 - 0.63
Medium sand	0.63 - 0.2
Fine sand	0.2 - 0.125
Fine fine sand	0.125 - 0.063
Total sand	**2 - 0.063**
Coarse silt	0.063 - 0.02
Medium silt	0.02 - 0.0063
Fine silt	0.0063 - 0.002
Total silt	**0.063 - 0.002**
Coarse clay	0.002 - 0.00063
Medium and fine clay	0.00063 - 0.000063
Total clay	**0.002 - 0.000063**

Texture class	Diameter in mm
Total silt and clay	0.063 - 0.000063
Total sand, silt and clay	**2 - 0.000063**

Properties of the measured samples

Vario EL III analyser is not able to measure Nt content lower than 0.03% and St content lower than 0.05%. Therefore, the Sulfur was not predicted as its values were to low.

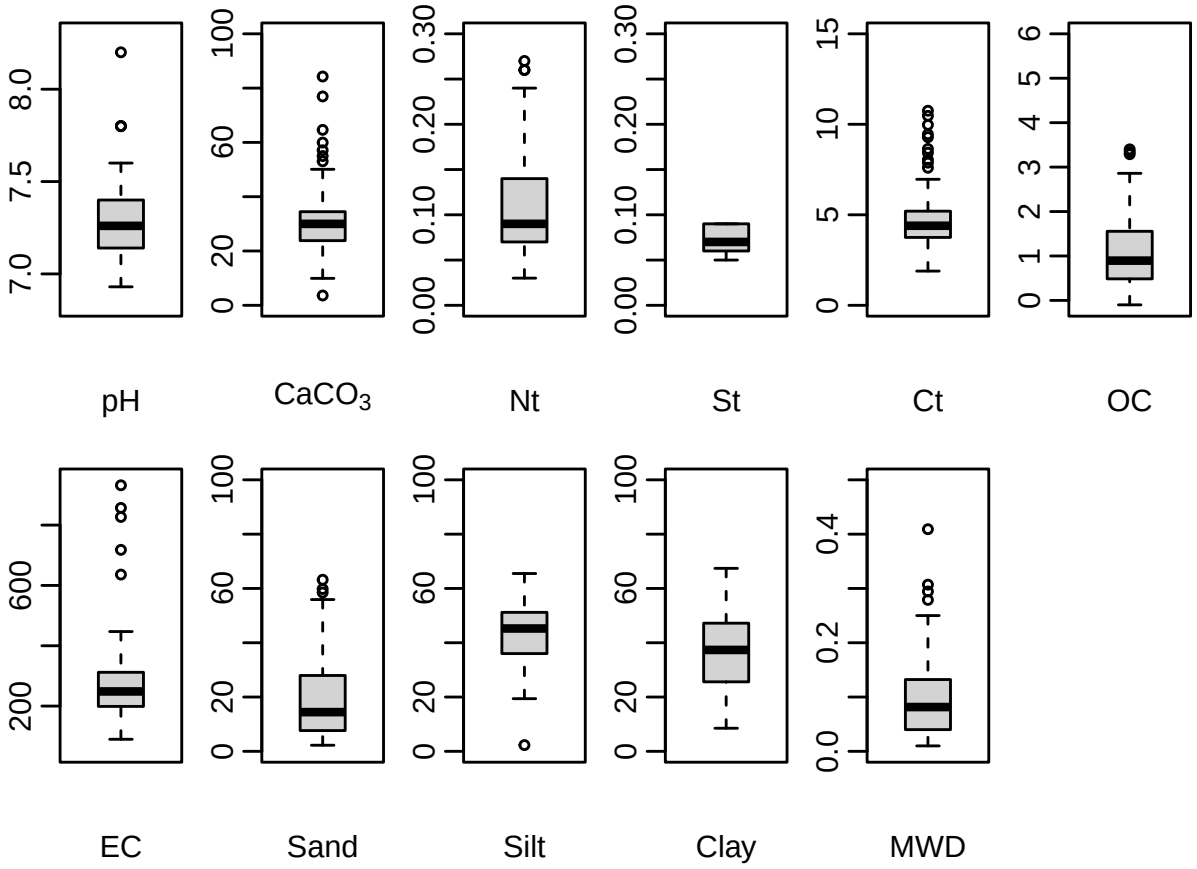

Figure S18: Boxplots of the measured values in laboratory.

We ploted the soil texture according to (WRB 2006) classification for each soil depth increment.

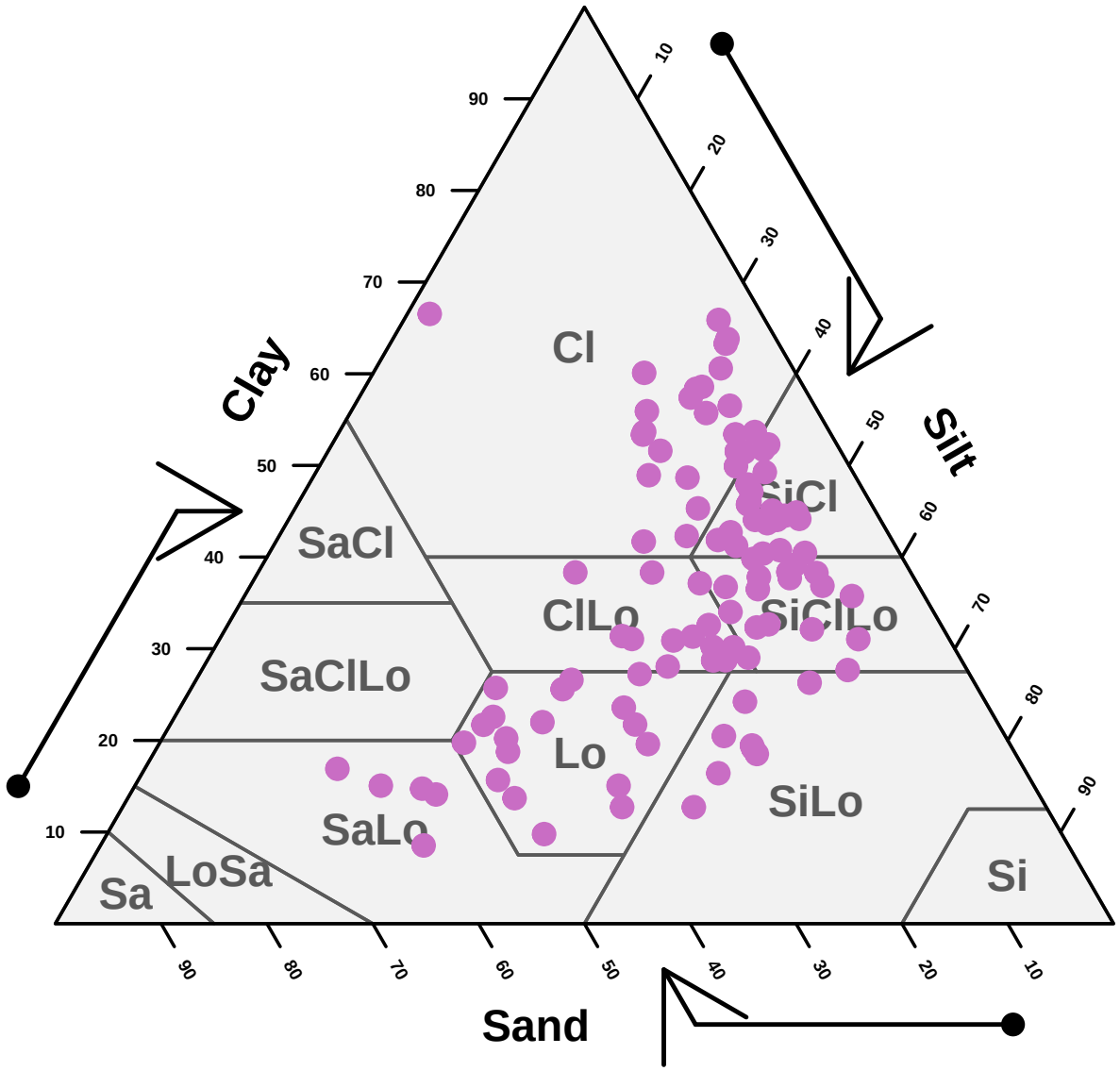

Figure S19: USDA texture of the measured samples for all depths.

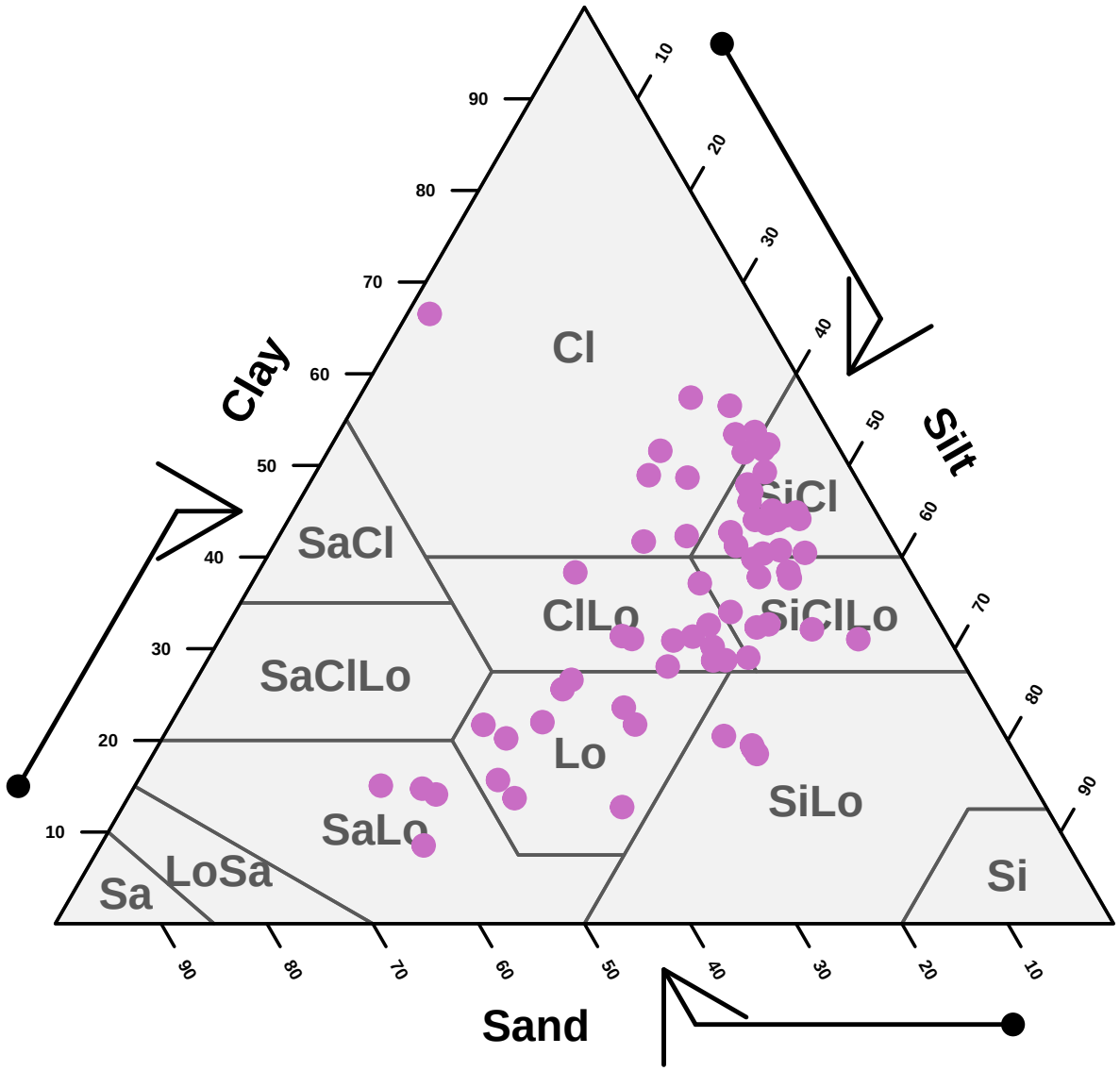

Figure S20: USDA texture of the measured samples for 0 - 10 cm samples.

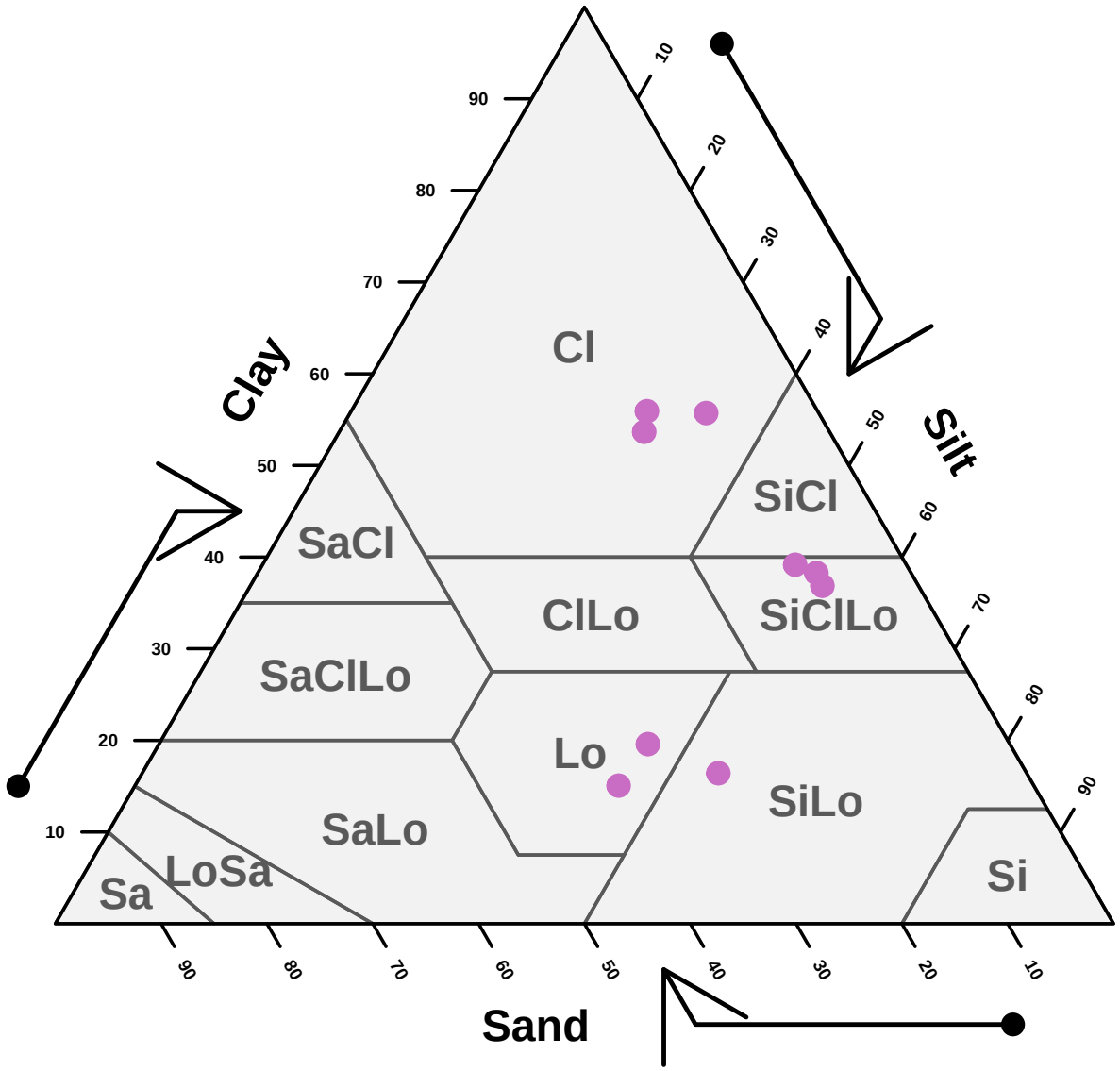

Figure S21: USDA texture of the measured samples for 10 - 30 cm samples.

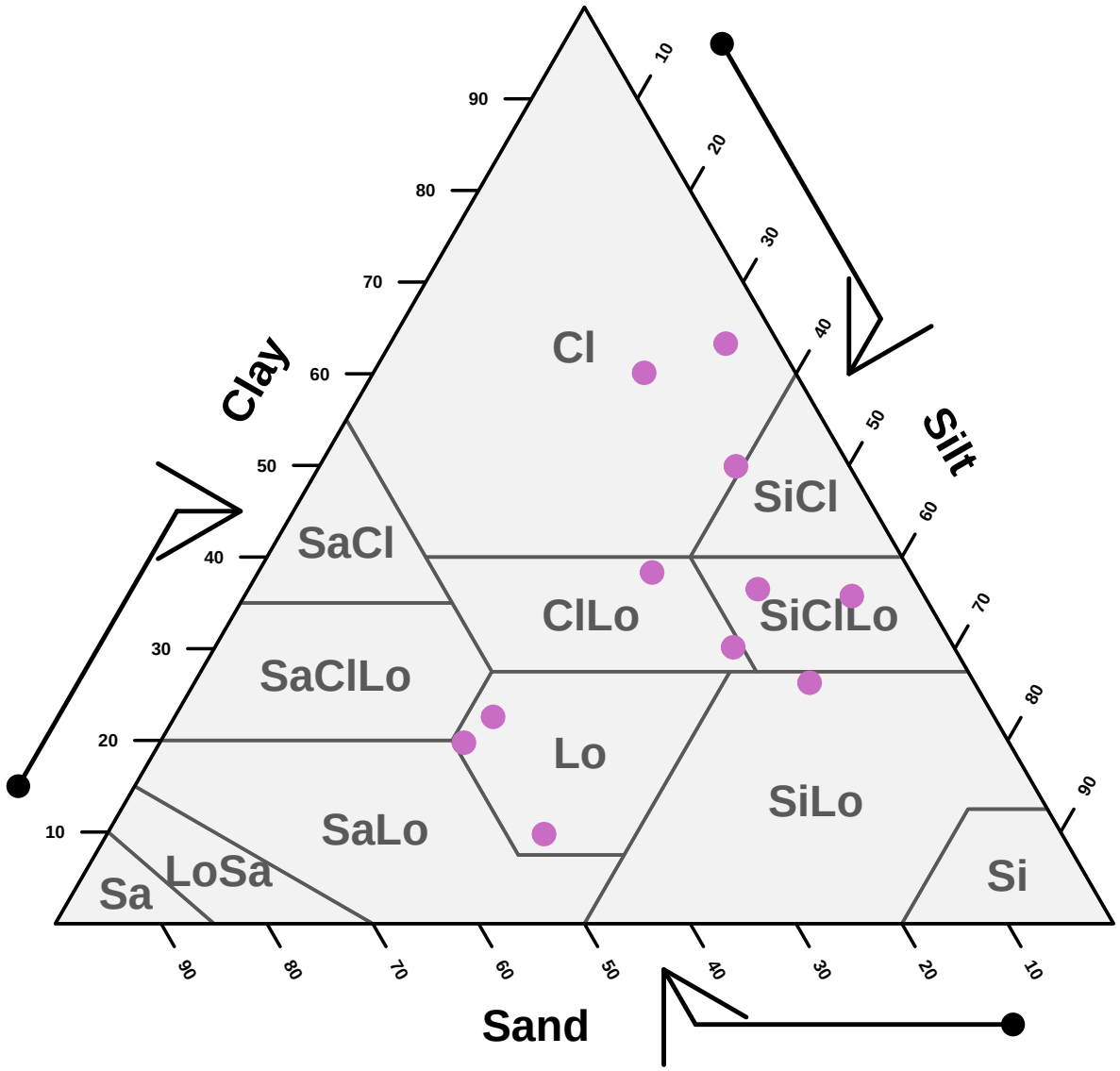

Figure S22: USDA texture of the measured samples for 30 - 50 cm samples.

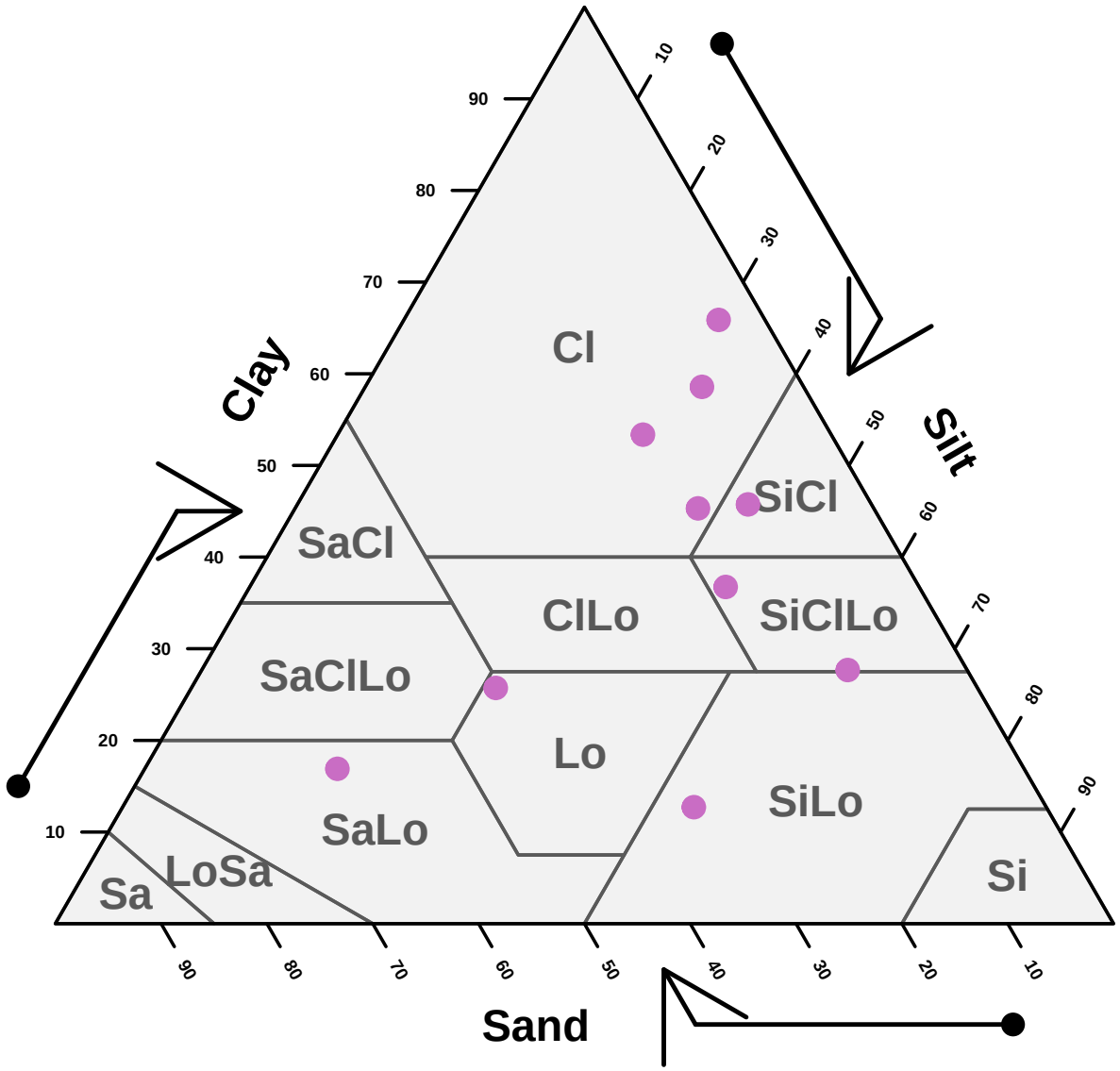

Figure S23: USDA texture of the measured samples for 50 - 70 cm samples.

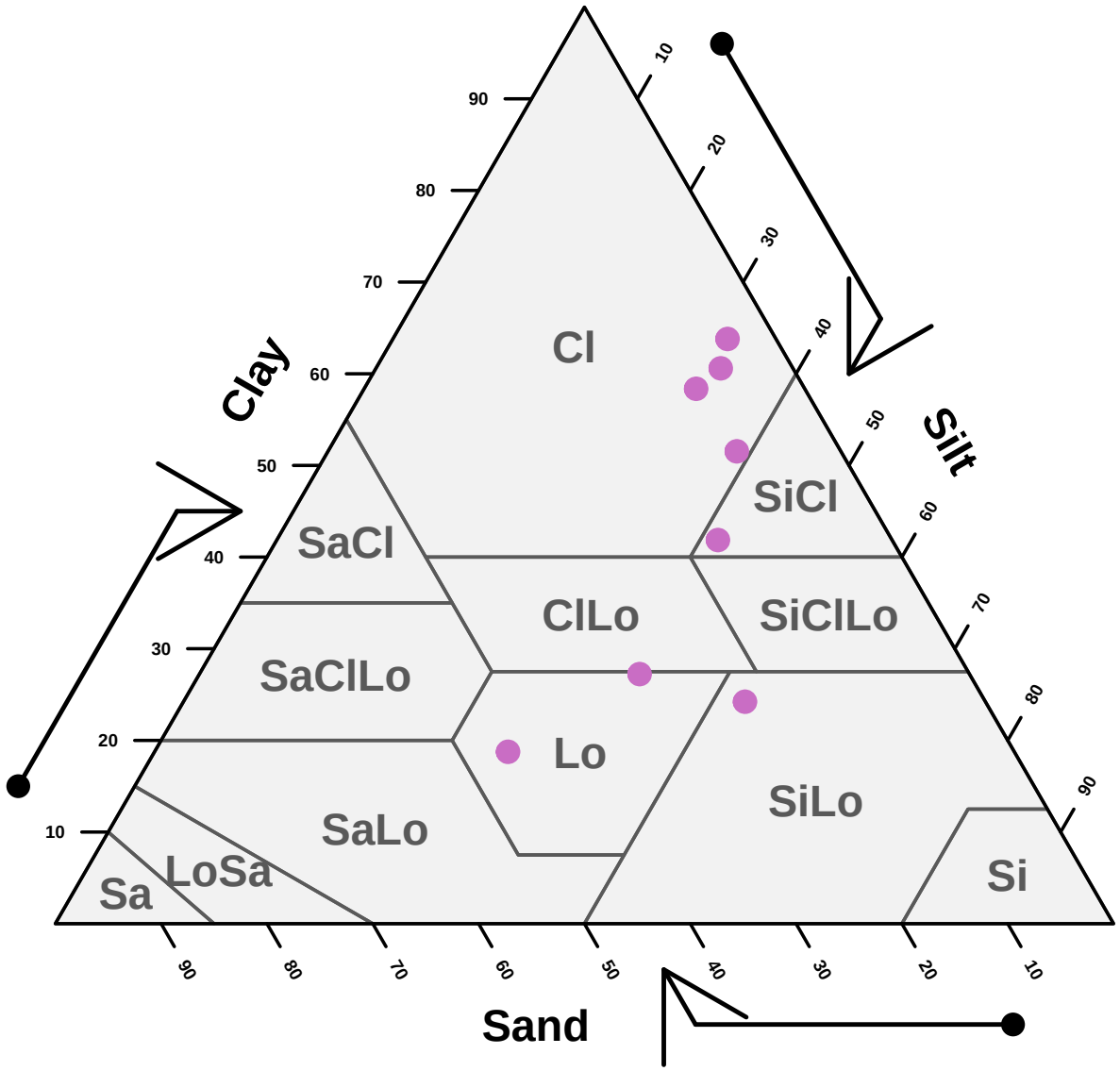

Figure S24: USDA texture of the measured samples for 70 - 100 cm samples.

Spectra model prediction

We used only the raw spectra and three spectra transformations from the **15** spectra produced in the FTIR part. This transformation are improving the quality of the prediction model (Ludwig et al. 2023). In total we had four spectra:

- **Raw spectra**.

- **SG-2-11** the Savitzky-Golay with a polynomial order of 2 and a window size of 11.
- **Moving average of 11**. **-SNV-SG** standard normal variate transformation on the Savitzky-Golay with a polynomial order of 2 and a window size of 11.

Model implementation

We implemented the prediction model in *Python* language. A sanity check was performed to assess the conformity of the data set *Sanity_Check_code*.

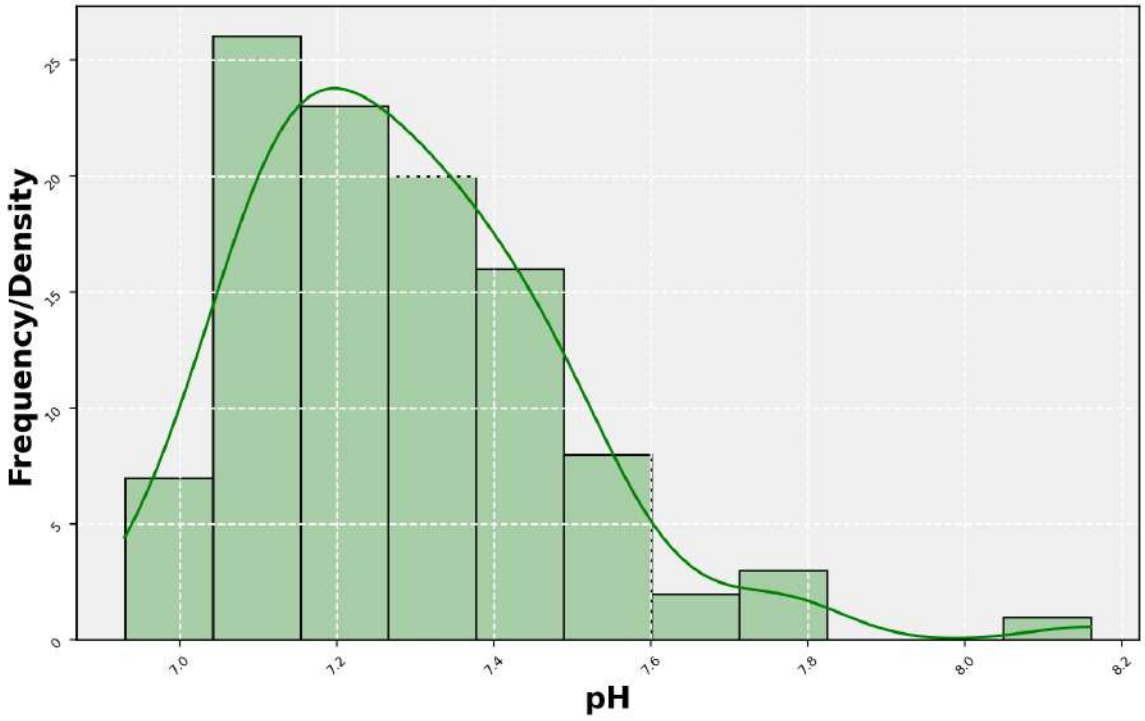

Figure S25: Density plot of pH.

We split the data into an 85% training set and a 15% testing set.

As EC values had outliers we removed all samples with values higher than 400 µS/cm: **53802** (827 µS/cm), **53871** (857 µS/cm), **53872** (932 µS/cm), **53902** (718 µS/cm), **53938** (447 µS/cm), **53954** (426 µS/cm) and **55852** (636 µS/cm).

All the properties were predicted using the Cubist model under *Python* environment *Cubist_Kfold_code*. A **5** time stratified k-fold was used with a total of **9** combinations for the hyperparameter. The stratified k-fold helps have a higher diversity of sampling in the training and testing set.

Table S7: Hyperparamters of the Cubist model.

n_rules	n_committees
20	5
30	5
40	5
20	10
30	10
40	10
20	15
30	15
40	15

Model evaluation

All spectra tranformations metrics

Table S8: Cubist model evaluation metrics on raw spectra.

Metric	ME	R2	RMSE	RPIQ
pH	-0.1130862	0.4401298	0.1479898	1.2423942
CaCO3	-1.3252205	0.8882458	4.1193275	2.3192395
Nt	-0.0210327	0.6229612	0.0543926	1.4595850
Ct	-0.2208946	0.8913525	0.5595367	2.3885632
OC	-0.0612059	0.5264148	0.6699846	0.9218613
EC	-30.6617828	0.1141583	66.3535662	0.9523699
Sand	-6.3994920	0.7674734	7.3599863	2.3624346
Silt	3.2742514	0.2136148	9.3070178	0.8782774
Clay	3.5598760	0.6930271	7.9695160	2.3912120
MWD	-0.0183279	0.4745772	0.0539841	1.2472350

Table S9: Cubist model evaluation metrics on mooving average 11 transformed spectra.

Metric	ME	R2	RMSE	RPIQ
pH	-0.0822356	0.4621653	0.1450483	1.221859
CaCO3	-0.0115416	0.8820737	4.2315510	2.389729
Nt	-0.0207909	0.6334873	0.0536280	1.410630
Ct	-0.0333944	0.9101929	0.5087145	2.228431
OC	-0.0276244	0.5112682	0.6806143	1.087405
EC	-16.8330812	0.1914366	63.3932824	1.114983
Sand	-6.9387918	0.7336180	7.8775860	2.300951
Silt	0.4148447	0.1086026	9.9089712	0.875066
Clay	6.7544251	0.7609876	7.0322156	2.343124
MWD	-0.0284105	0.4724332	0.0540942	1.331620

Table S10: Cubist model evaluation metrics on SG transformed spectra.

Metric	ME	R2	RMSE	RPIQ
pH	0.0139322	0.4081296	0.1521604	0.9712110
CaCO3	-0.0466458	0.9012224	3.8727870	2.7690478
Nt	0.0136756	0.3532399	0.0712392	0.4787131
Ct	-0.0795040	0.8415988	0.6756130	1.8099295
OC	-0.0751953	0.2343362	0.8518925	0.3591207
EC	37.7198879	0.0463005	68.8481029	0.9539687
Sand	-2.6170488	0.7616948	7.4508777	2.4868337
Silt	-4.7215761	0.3247059	8.6246080	0.9129198
Clay	-3.5665841	0.7175530	7.6445239	2.4276119
MWD	0.0299509	0.4394541	0.0557593	1.1867736

Table S11: Cubist model evaluation metrics on SNV-SG transformed spectra.

Metric	ME	R2	RMSE	RPIQ
pH	-0.1487360	0.3990151	0.1533275	0.8891613
CaCO3	-1.1519722	0.8888660	4.1078803	2.7618729
Nt	-0.0046014	0.8257789	0.0369741	1.7119021
Ct	0.0367057	0.9123964	0.5024348	2.7212452
OC	0.1736385	0.6126784	0.6059009	1.1525632
EC	-5.9358389	0.2942953	59.2239976	0.9349021
Sand	2.0420108	0.8106039	6.6424182	2.6565763
Silt	0.1929490	0.3783928	8.2746742	0.9772474
Clay	-1.7827777	0.6928479	7.9718426	2.2699483
MWD	-0.0022866	0.4592173	0.0547675	1.1640412

Selected transformed spectra

Regarding the results of the predictions, we selected for each variables the following spectra:

- **pH** Raw spectra.
- **$CaCO_3$** SG.2.11 spectra.
- **Nt** SNV-SG spectra.
- **Ct** SNV-SG spectra.
- **OC** SNV-SG spectra.
- **EC** SNV-SG spectra.
- **Sand** SNV-SG spectra.
- **Silt** SNV-SG spectra.
- **Clay** SG.2.11 spectra.
- **MWD** Raw spectra.

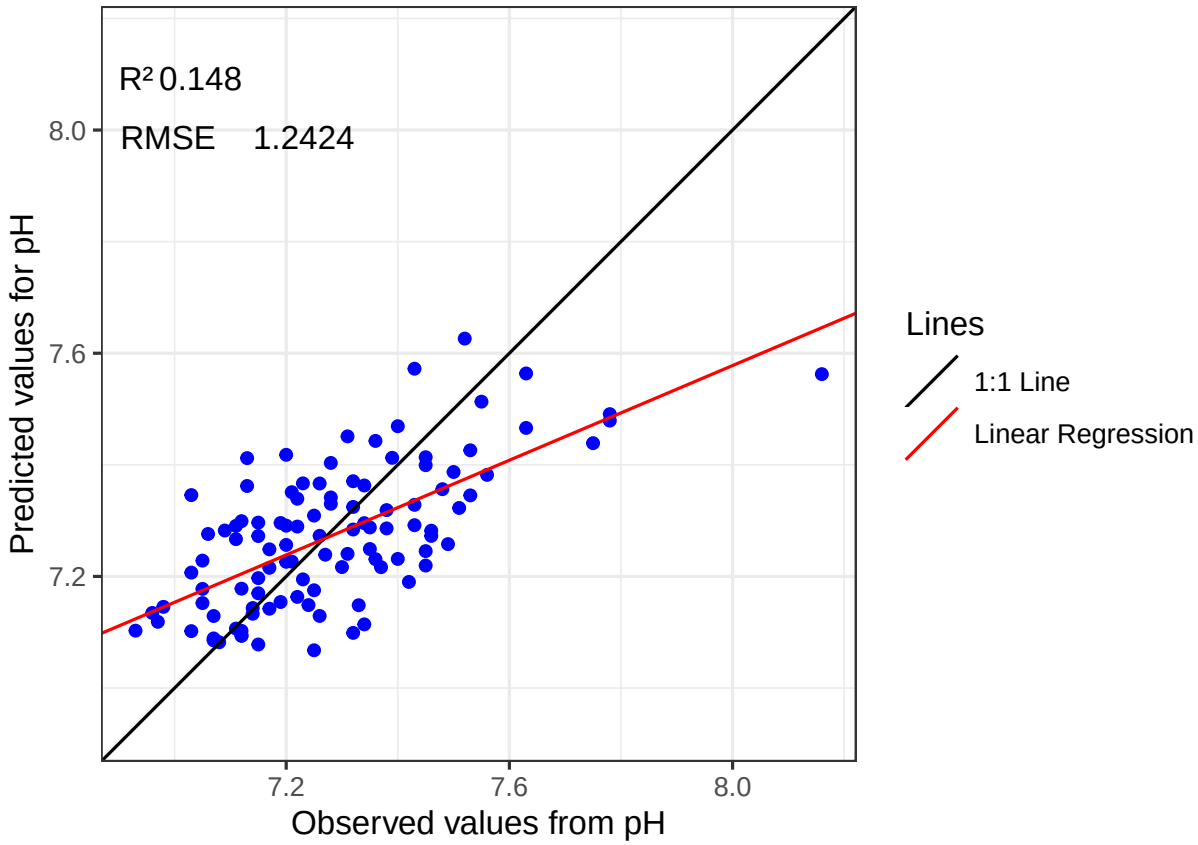

Figure S26: Regression curve of the predicted pH values.

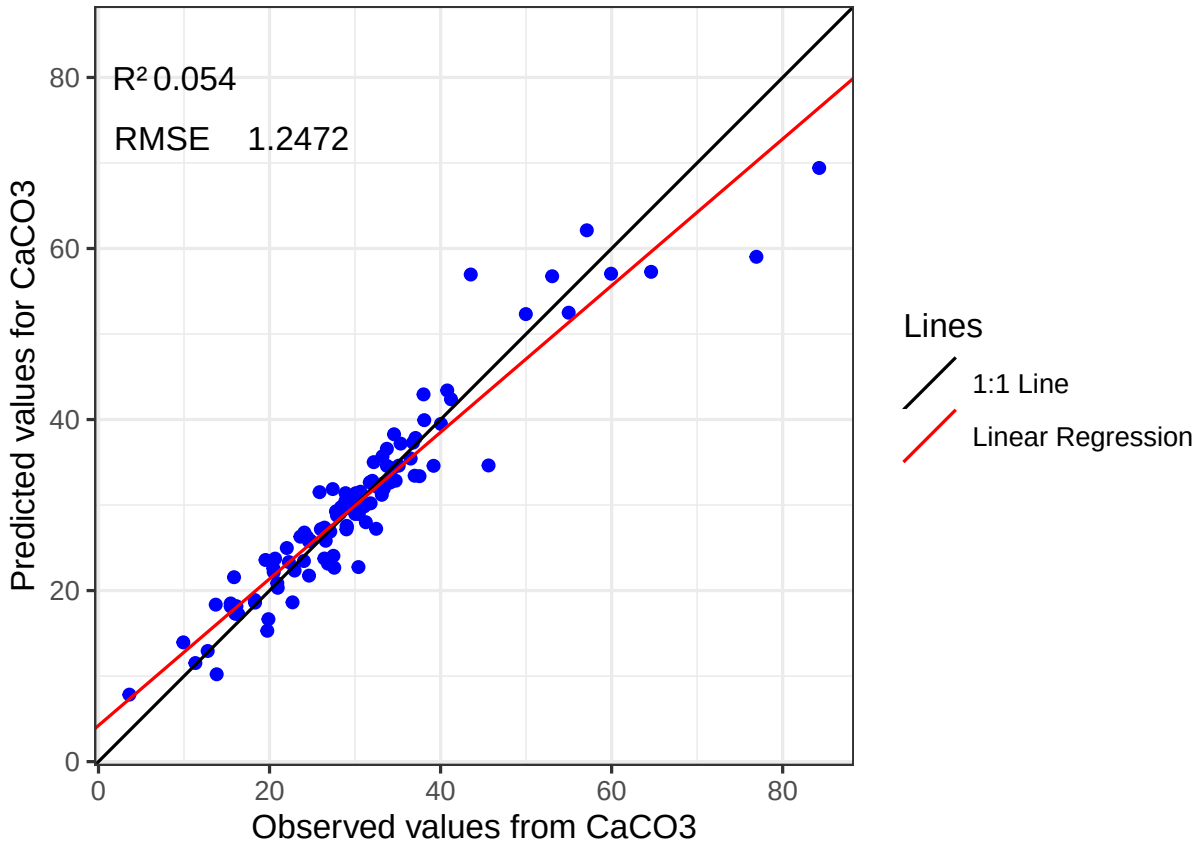

Figure S27: Regression curve of the predicted CaCo3 values.

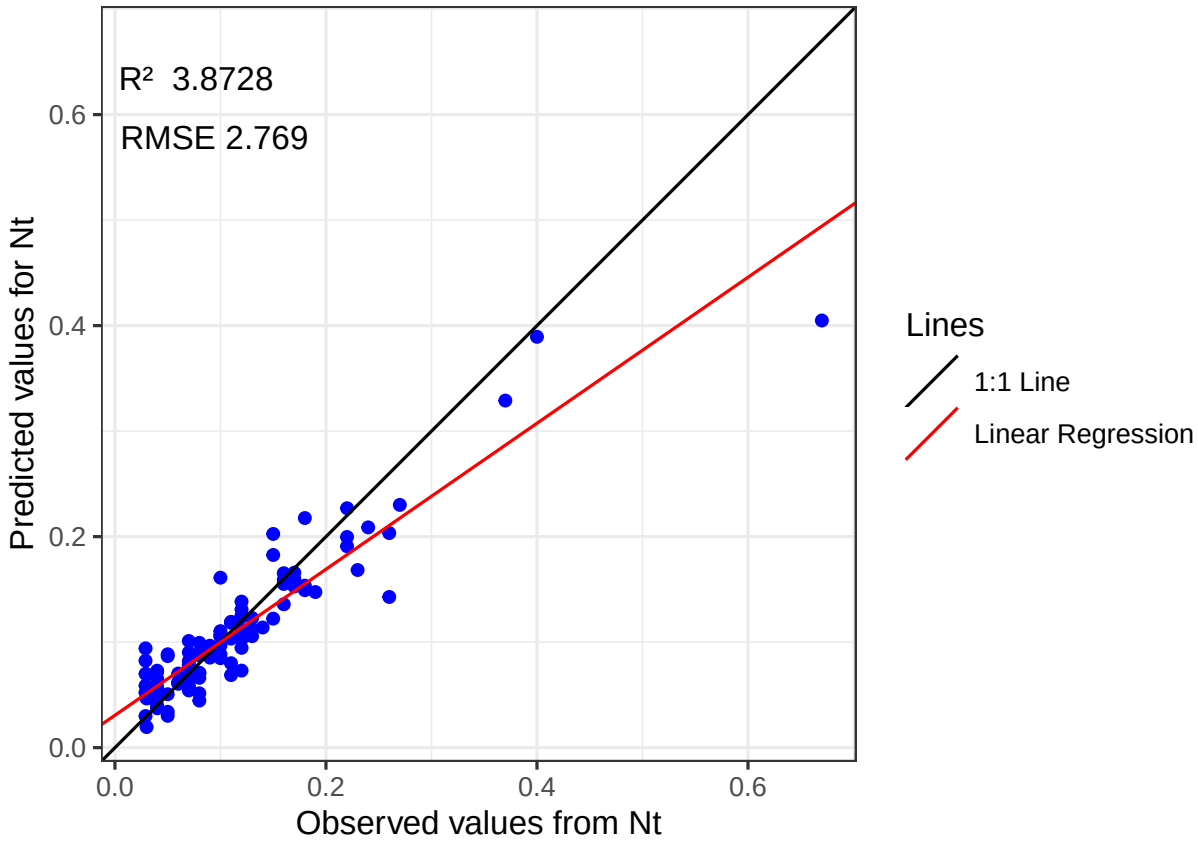

Figure S28: Regression curve of the predicted Nt values.

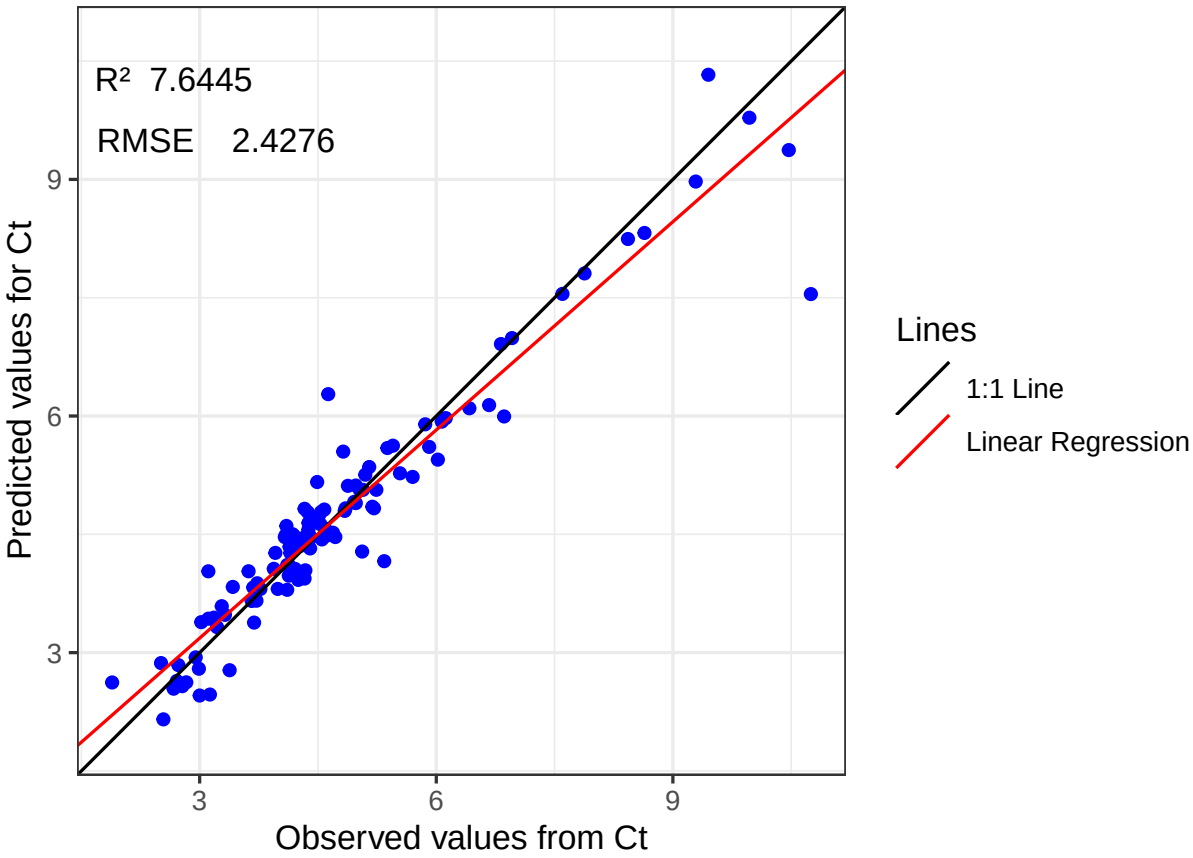

Figure S29: Regression curve of the predicted Ct values.

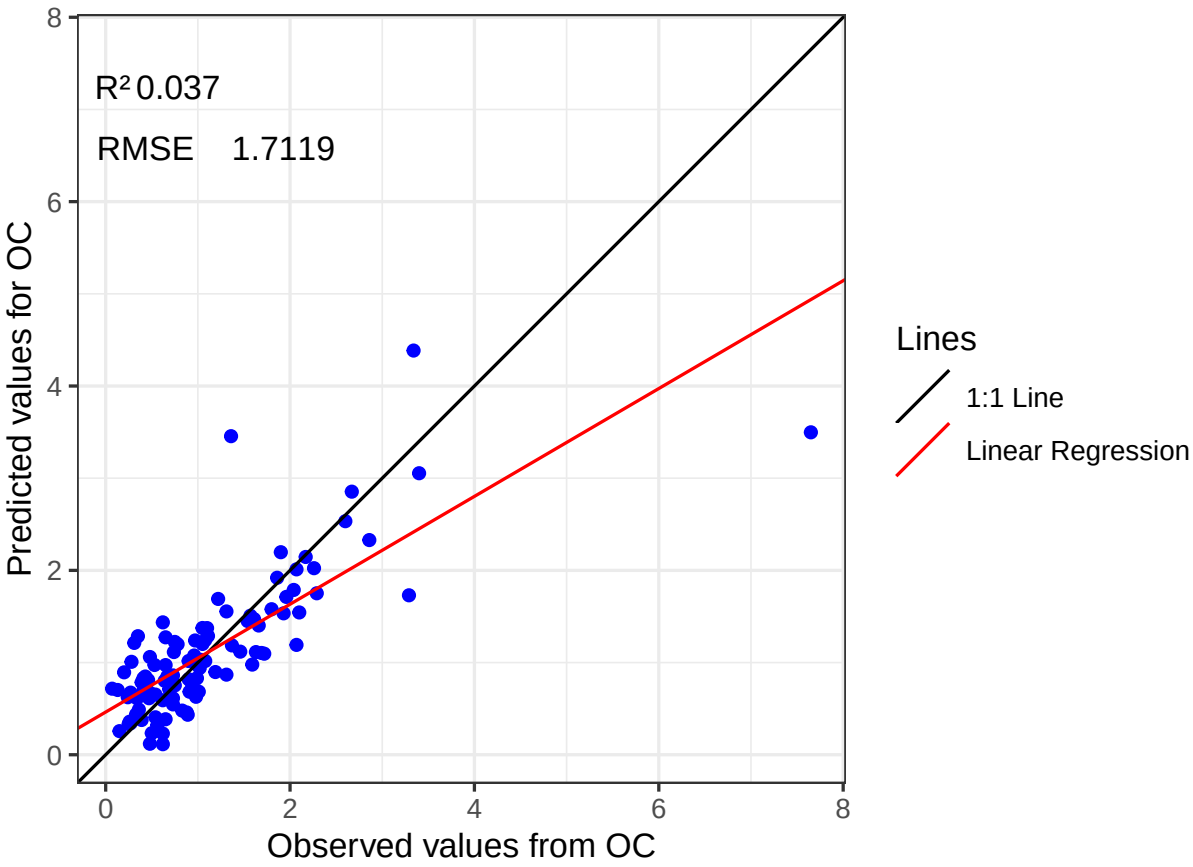

Figure S30: Regression curve of the predicted OC values.

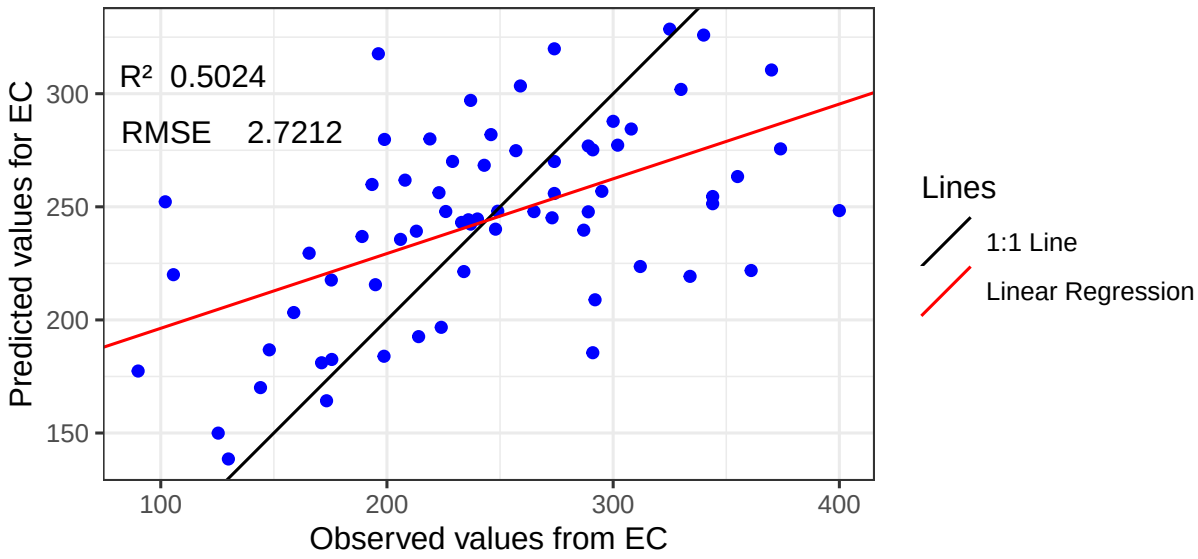

Figure S31: Regression curve of the predicted EC values.

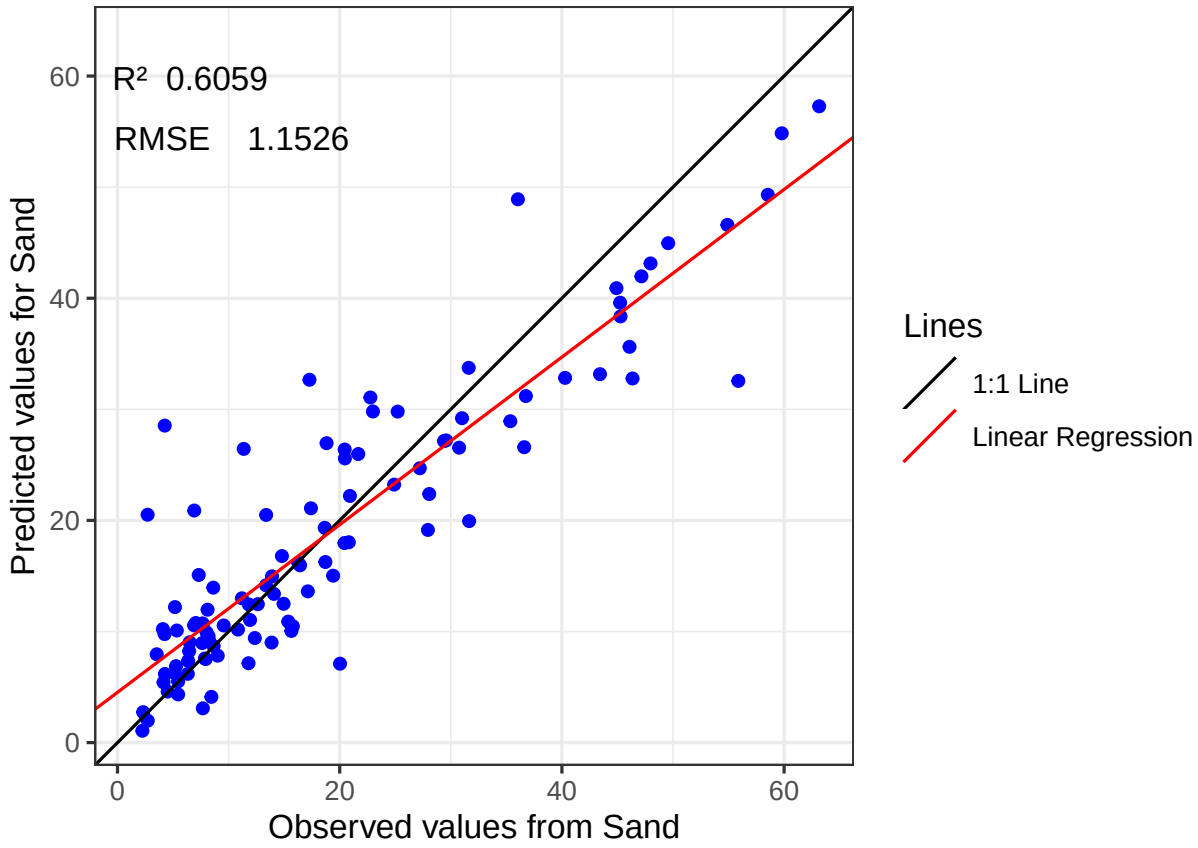

Figure S32: Regression curve of the predicted Sand values.

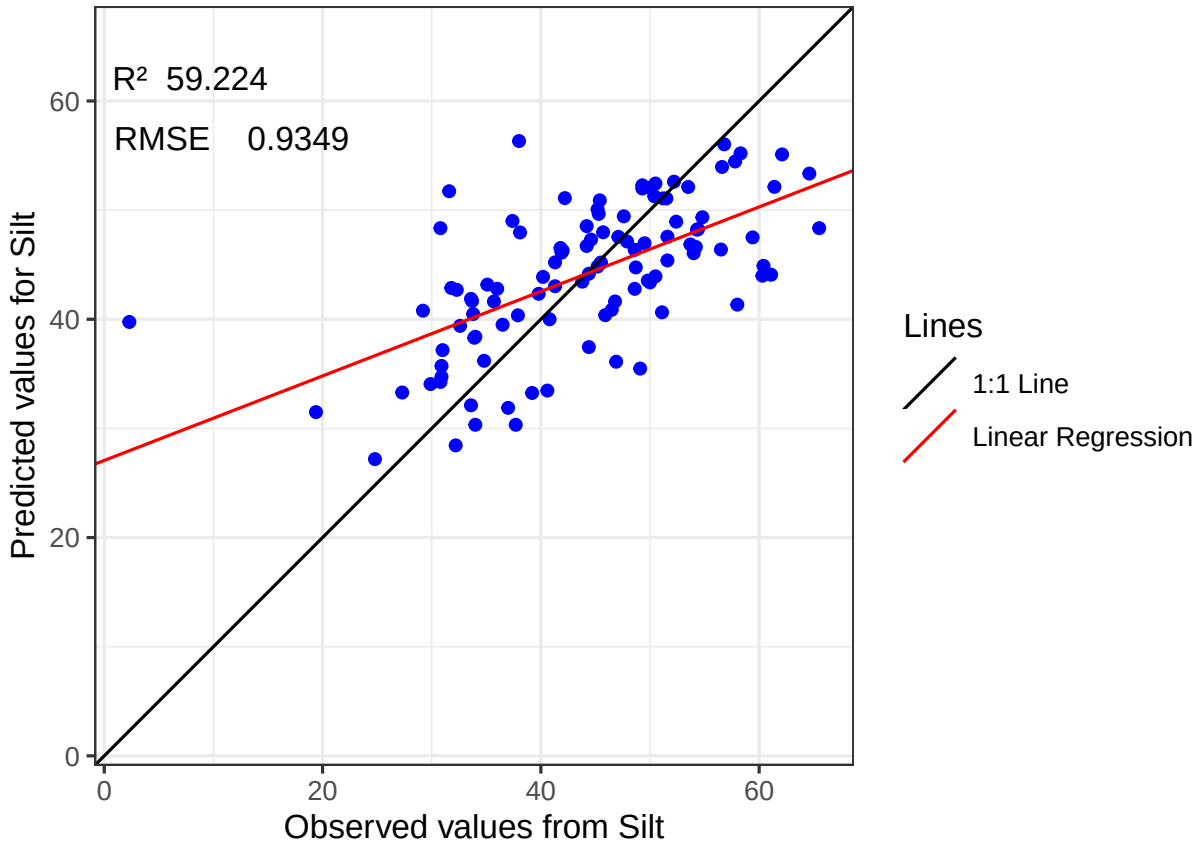

Figure S33: Regression curve of the predicted Silt values.

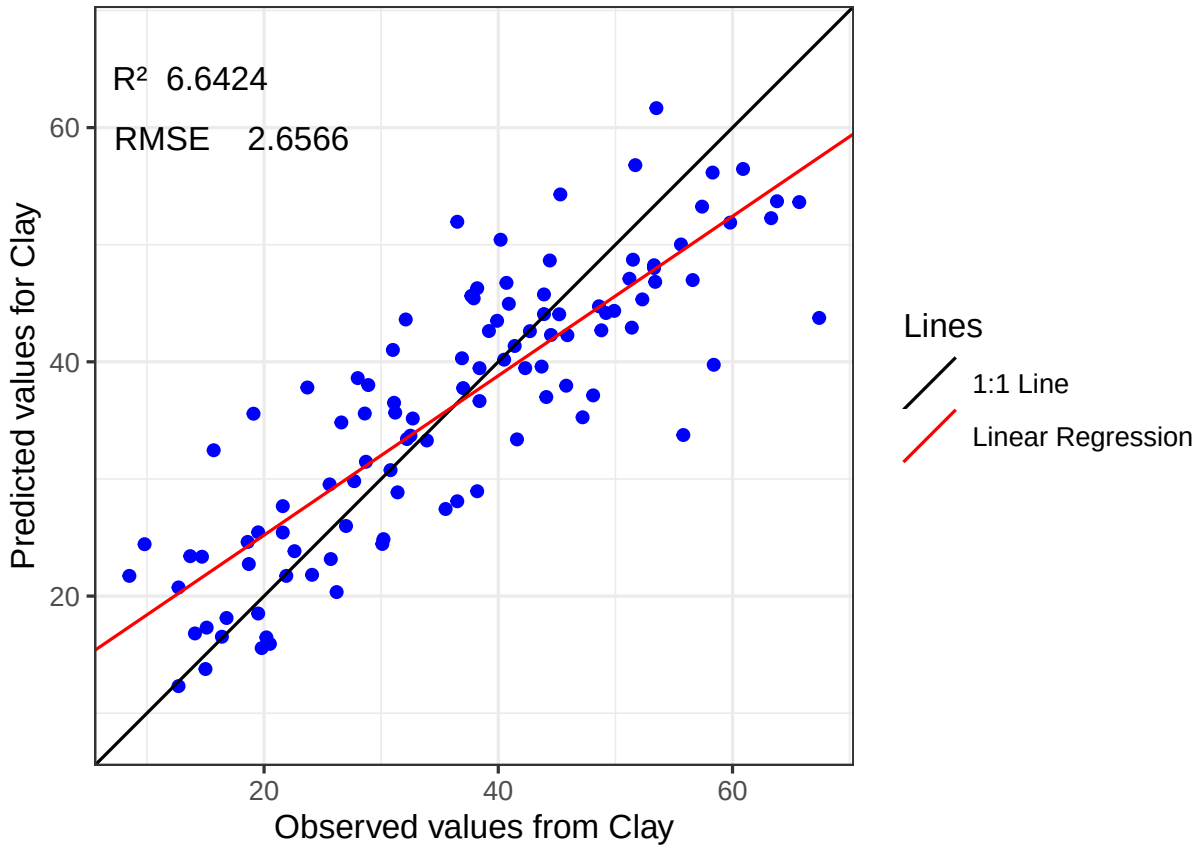

Figure S34: Regression curve of the predicted Clay values.

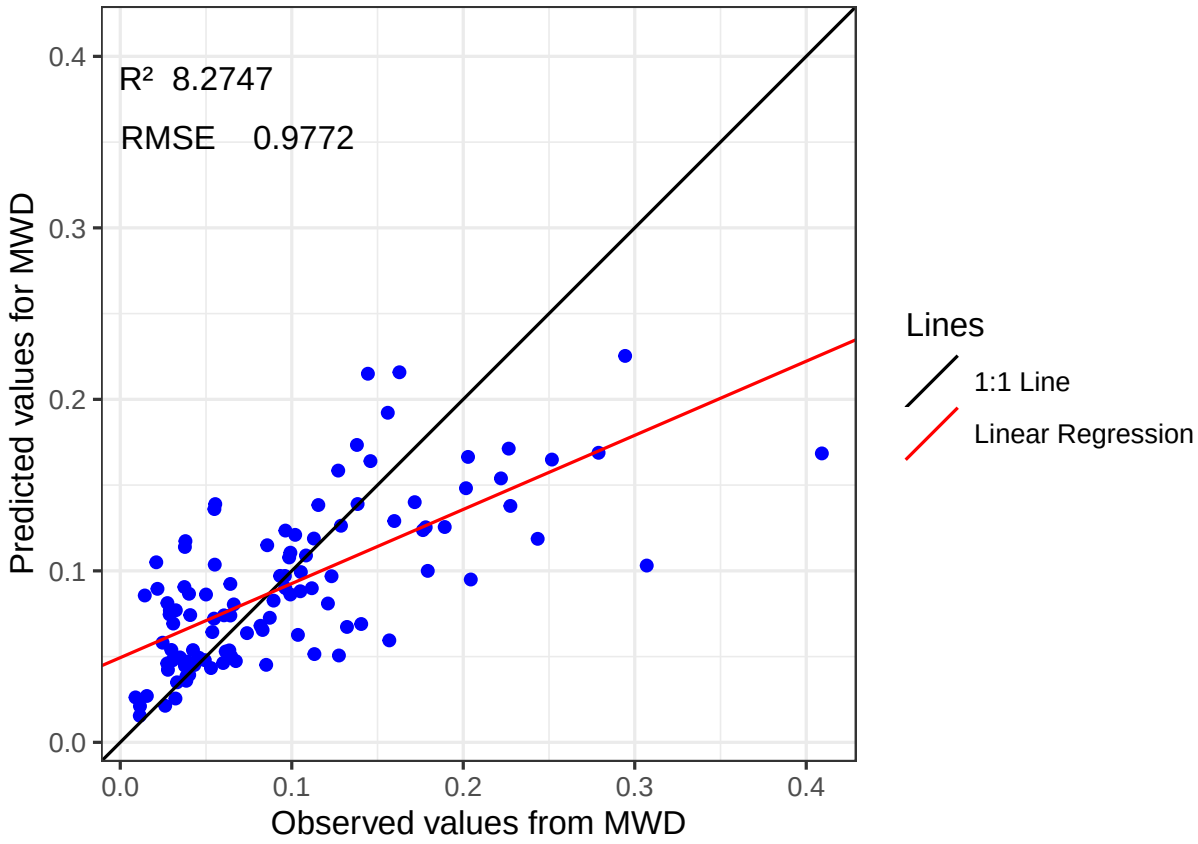

Figure S35: Regression curve of the predicted MWD values.

Predicted values

Full predicted values table

The full table of 2022 and 2023 samples predicted is visible below or can be found on the deposit as *Mir_spectra_prediction* file or at: https://doi.org/10.1594/PANGAEA.973700.

We removed the sample with ID **55864** which is an outlier.

Boxplot of the predicted values

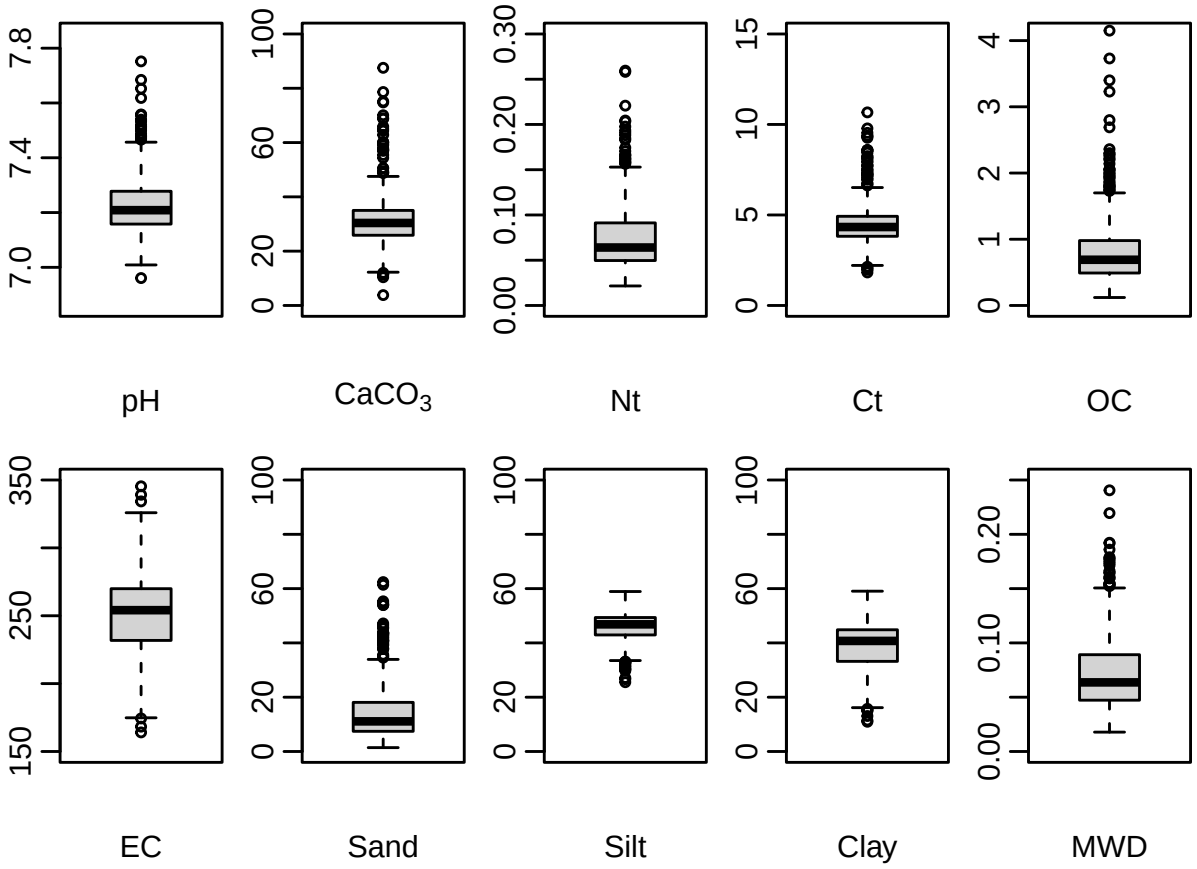

Figure S36: Boxplots of all the predicted samples.

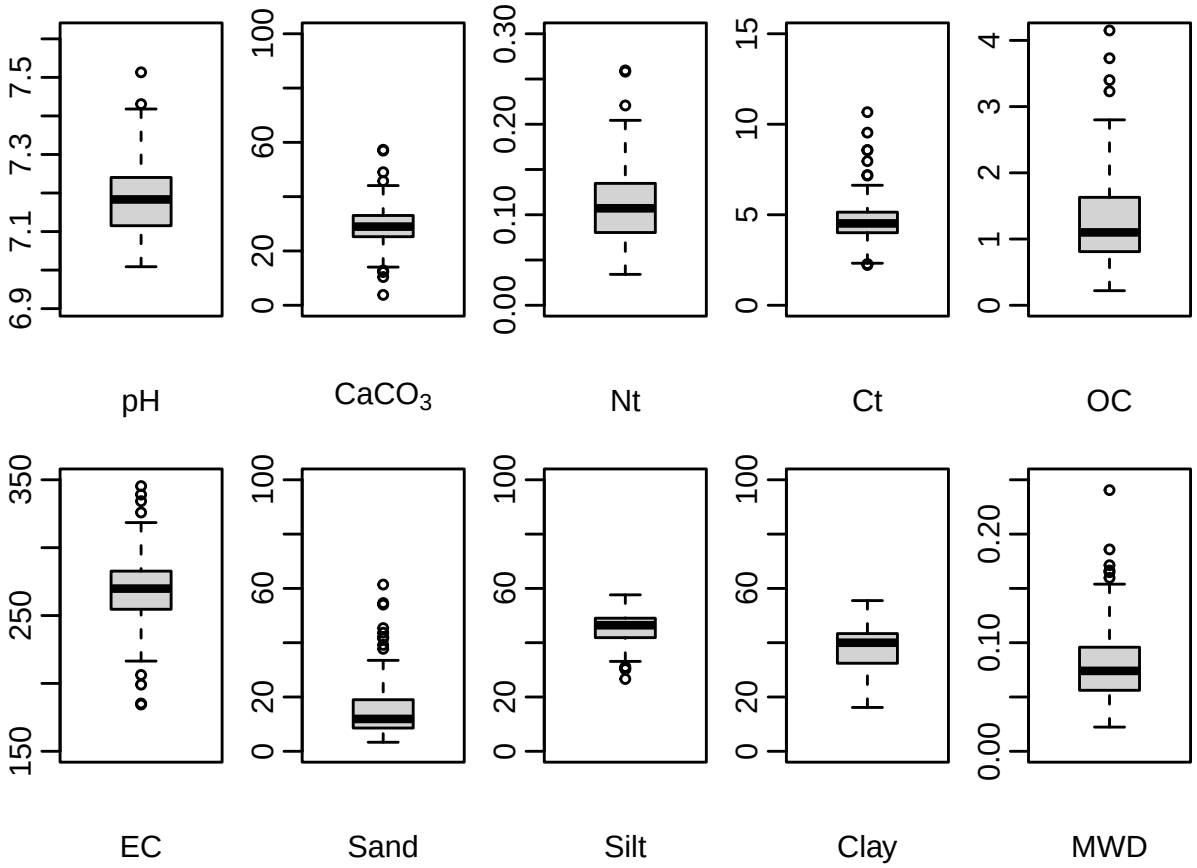

Figure S37: Boxplots of the 0 - 10 cm predicted samples.

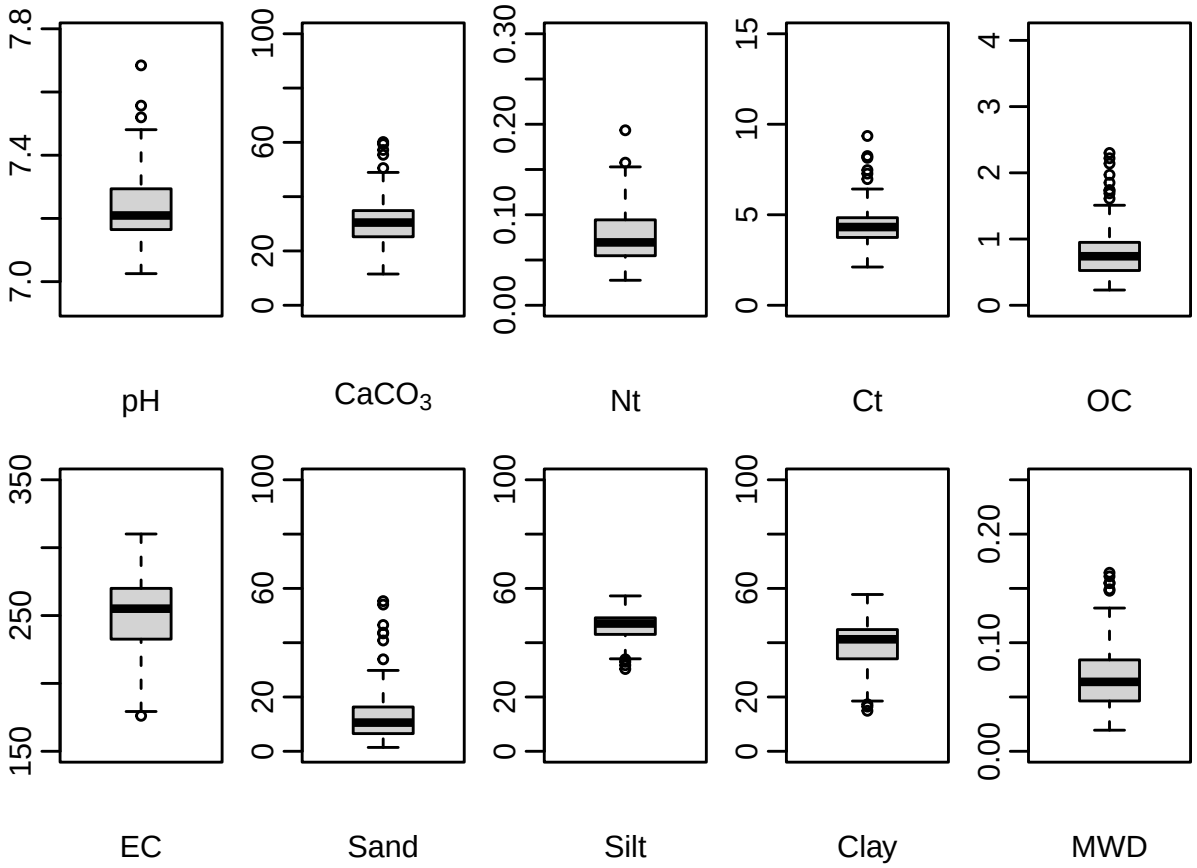

Figure S38: Boxplots of the 10 - 30 cm predicted samples.

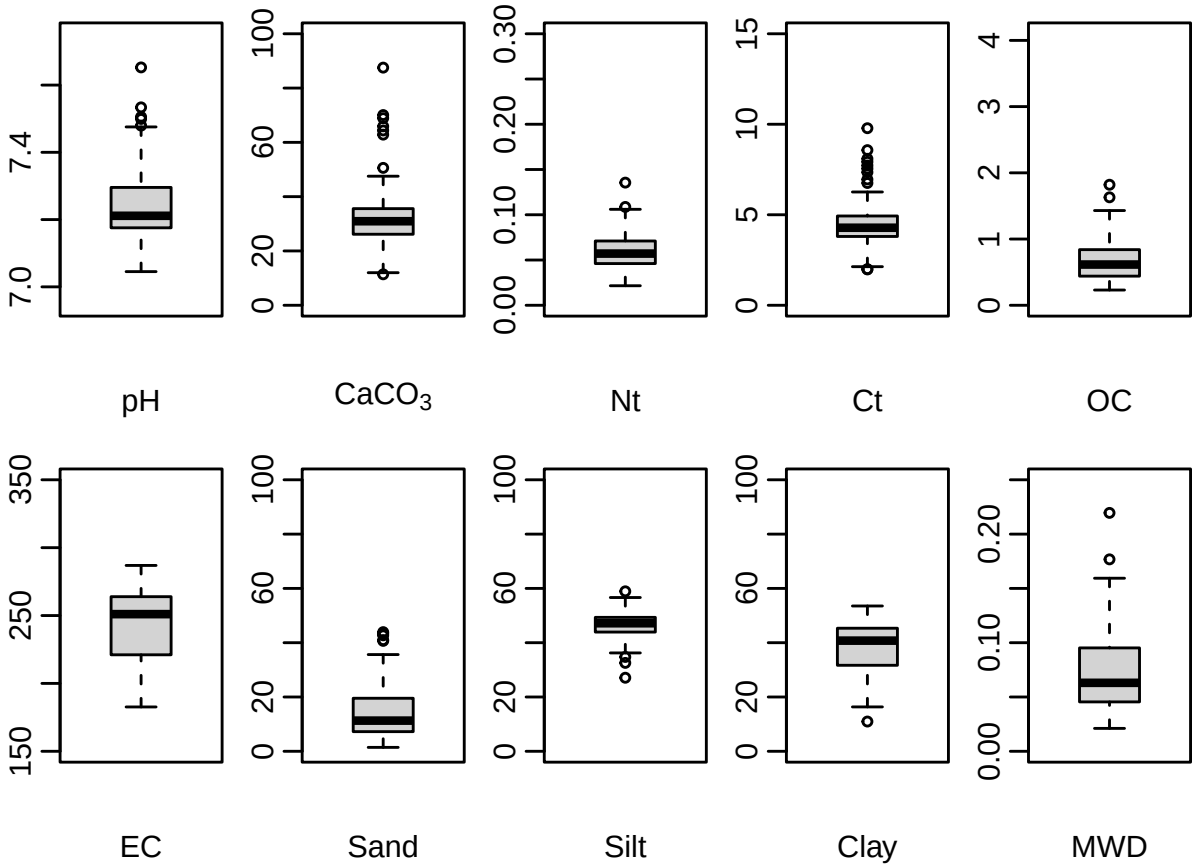

Figure S39: Boxplots of the 30 - 50 cm predicted samples.

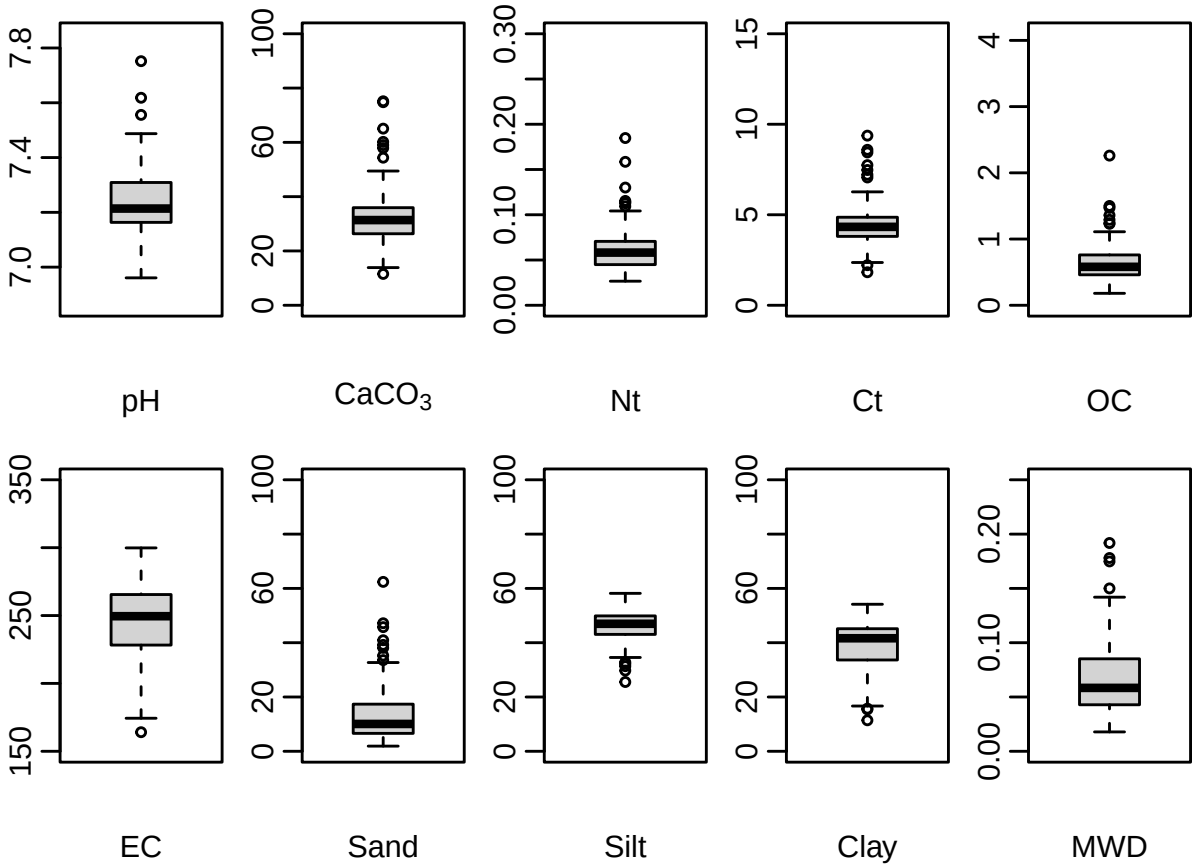

Figure S40: Boxplots of the 50 - 70 cm predicted samples.

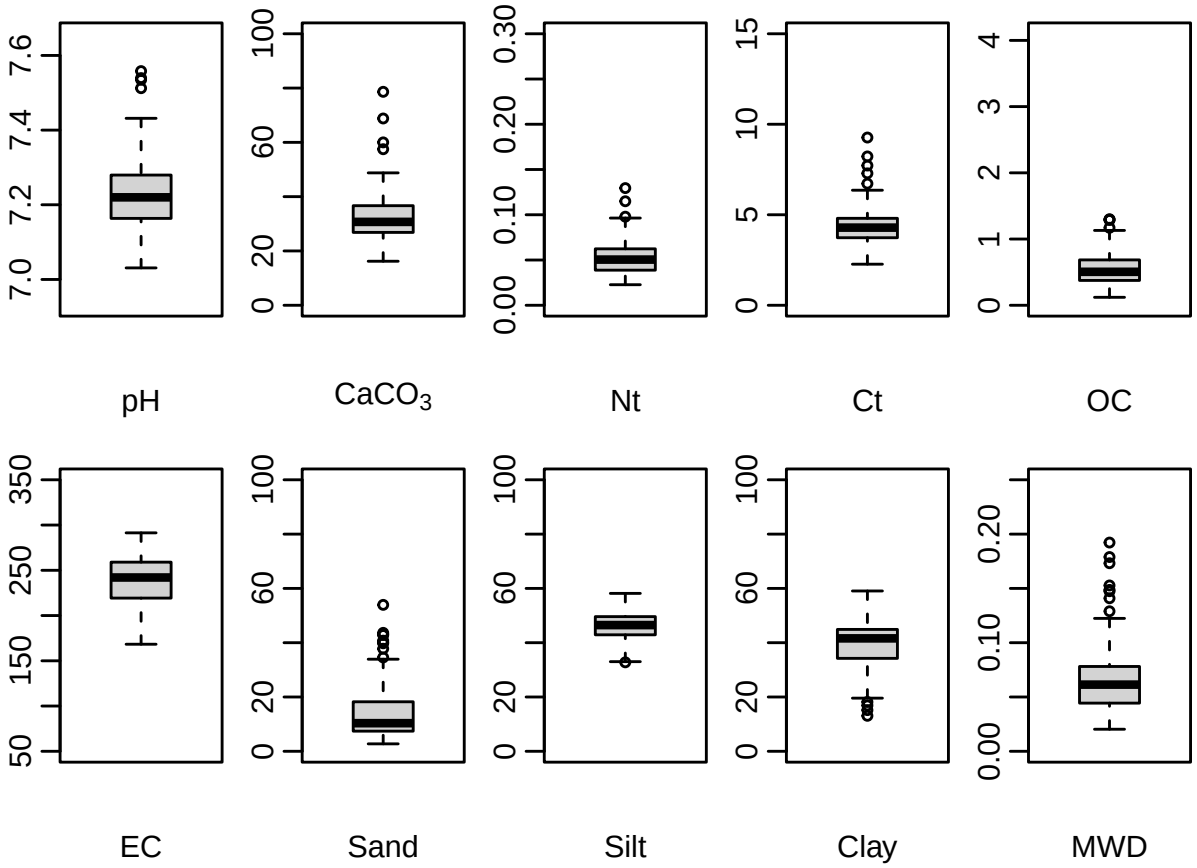

Figure S41: Boxplots of the 70 - 100 cm predicted samples.

Texture of the predicted samples

We ploted the soil texture according to (WRB 2006) classification.

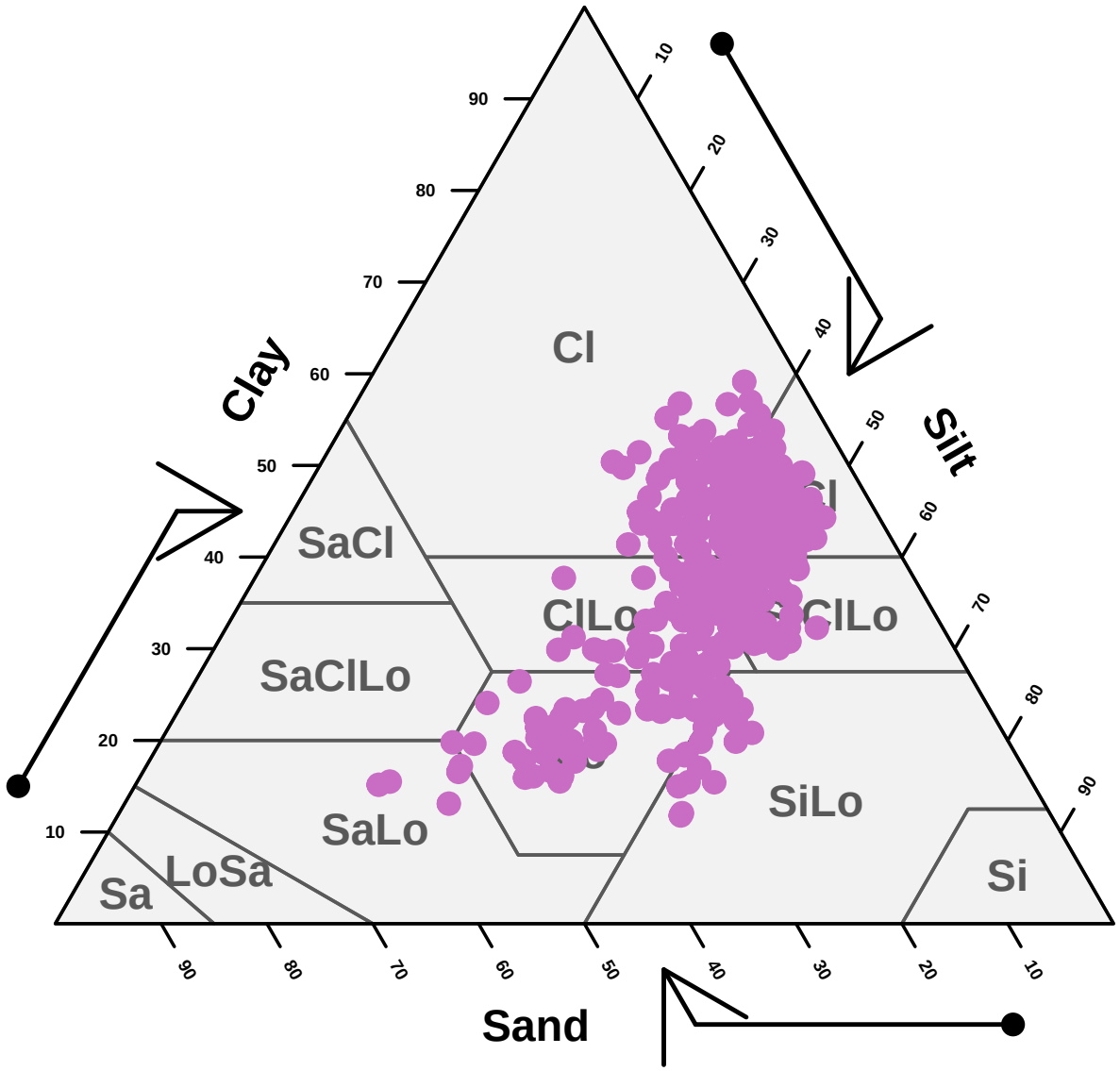

Figure S42: USDA texture of the predicted samples for all depths.

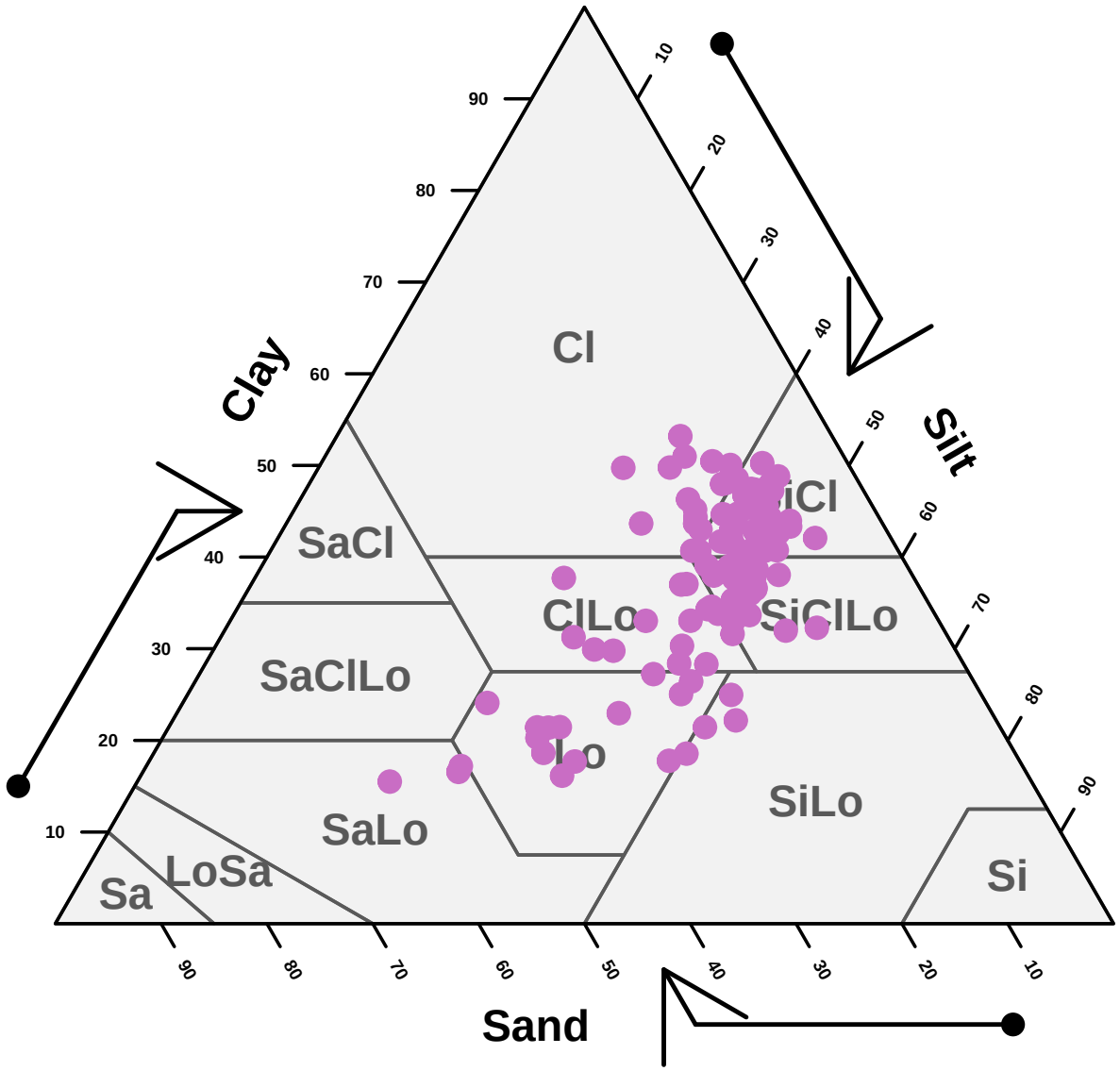

Figure S43: USDA texture of the predicted samples for 0 - 10 cm samples.

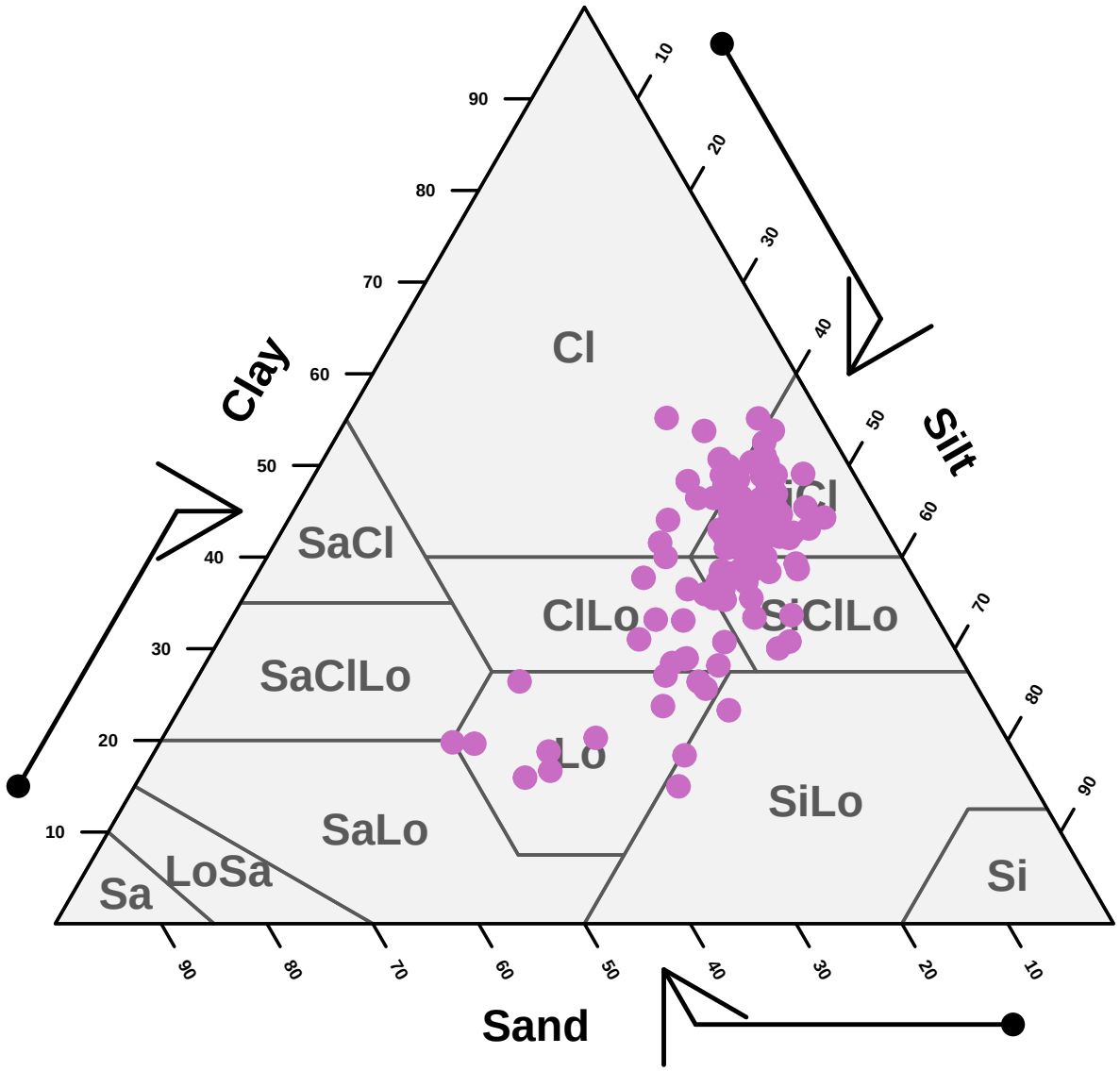

Figure S44: USDA texture of the predicted samples for 10 - 30 cm samples.

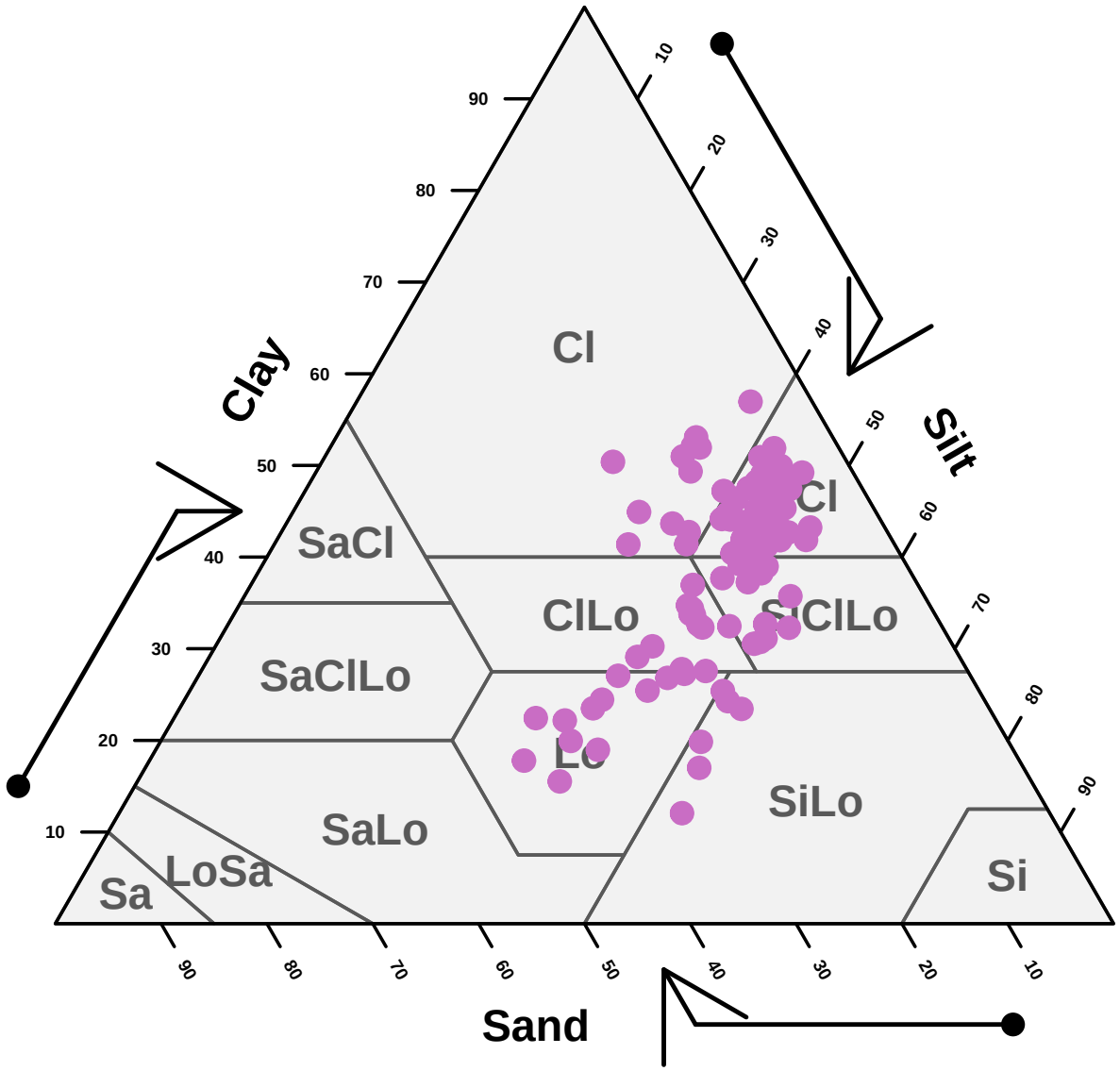

Figure S45: USDA texture of the predicted samples for 30 - 50 cm samples.

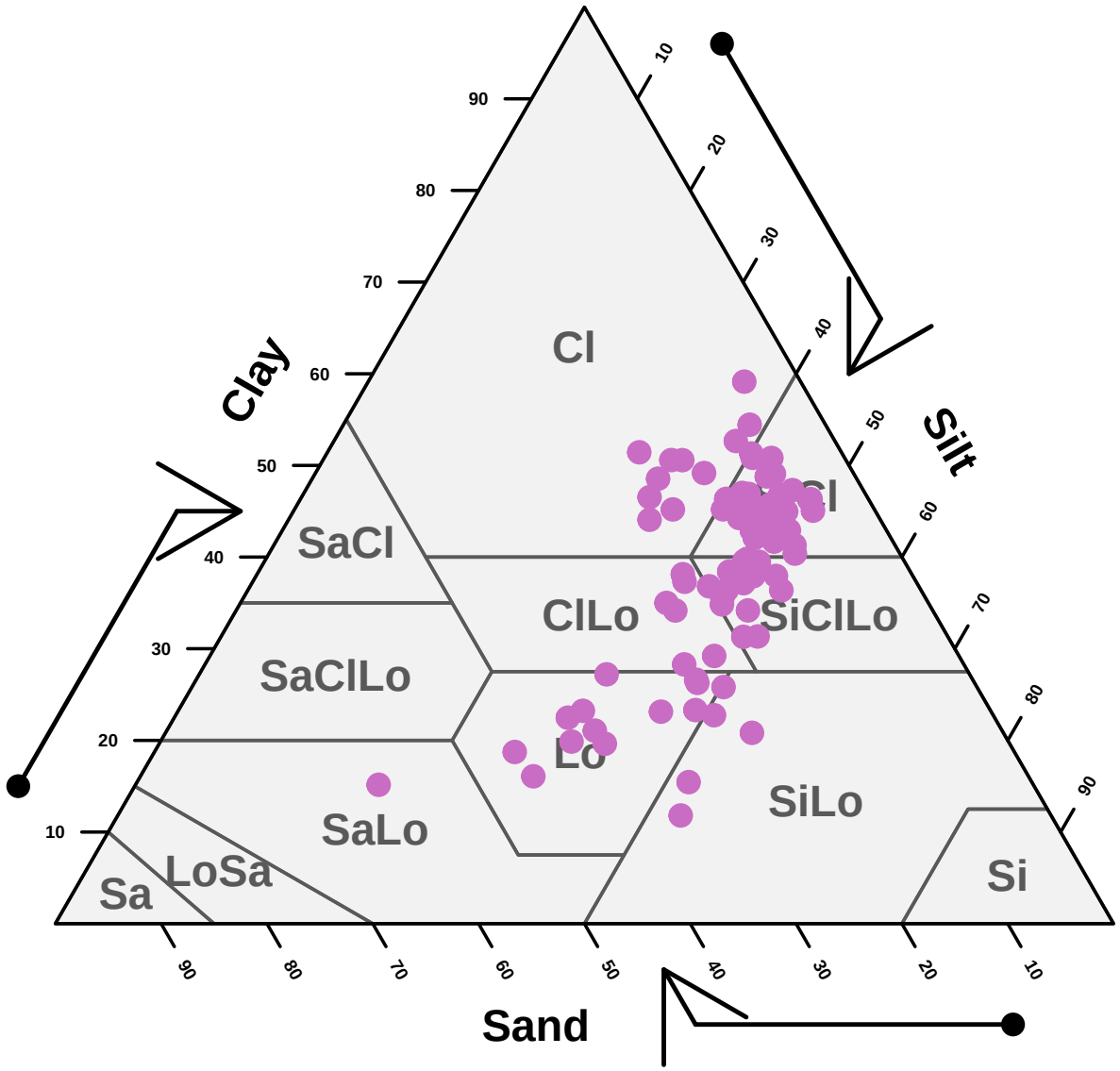

Figure S46: USDA texture of the predicted samples for 50 - 70 cm samples.

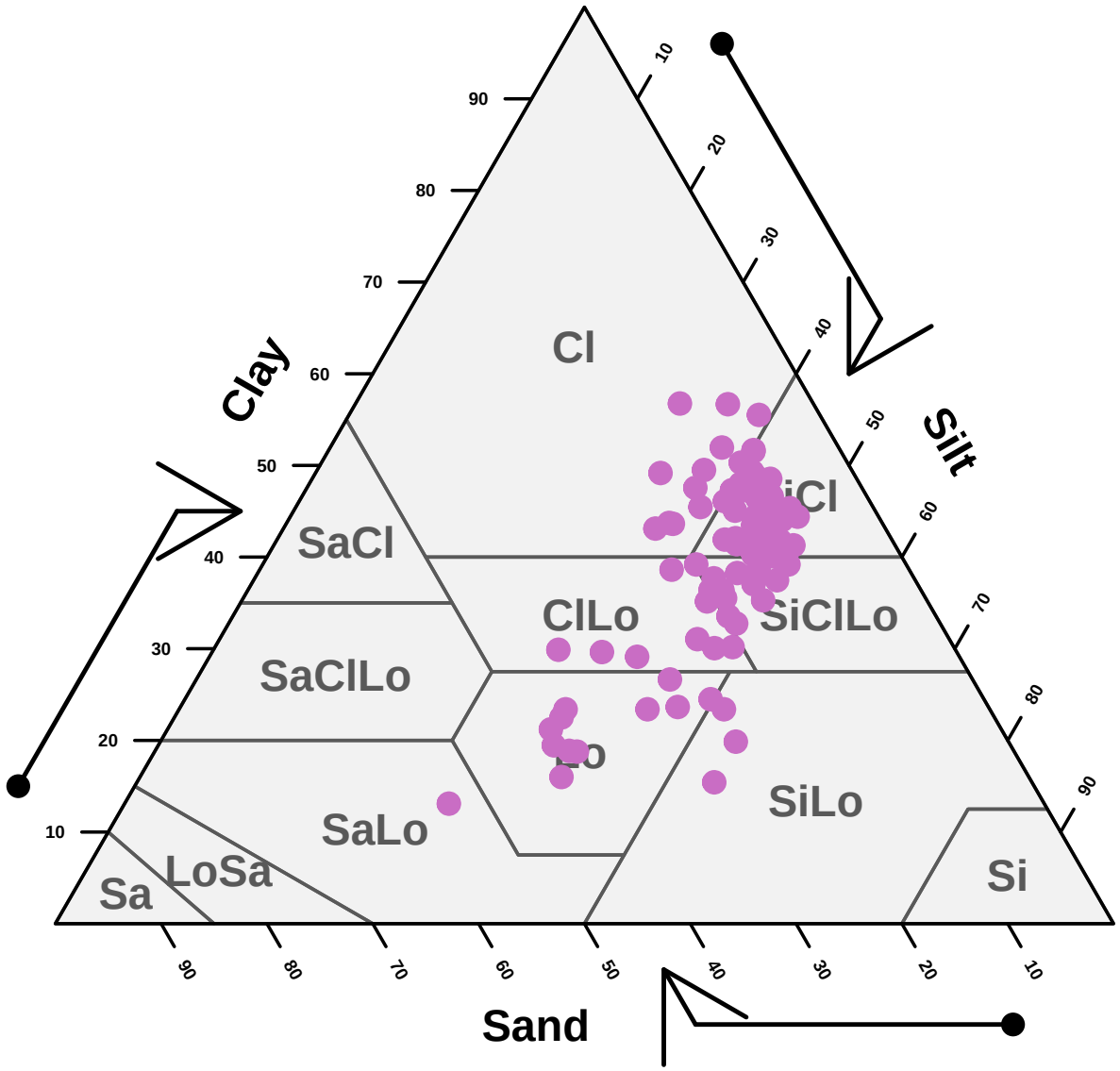

Figure S47: USDA texture of the predicted samples for 70 - 100 cm samples.

Observed vs. predicted values

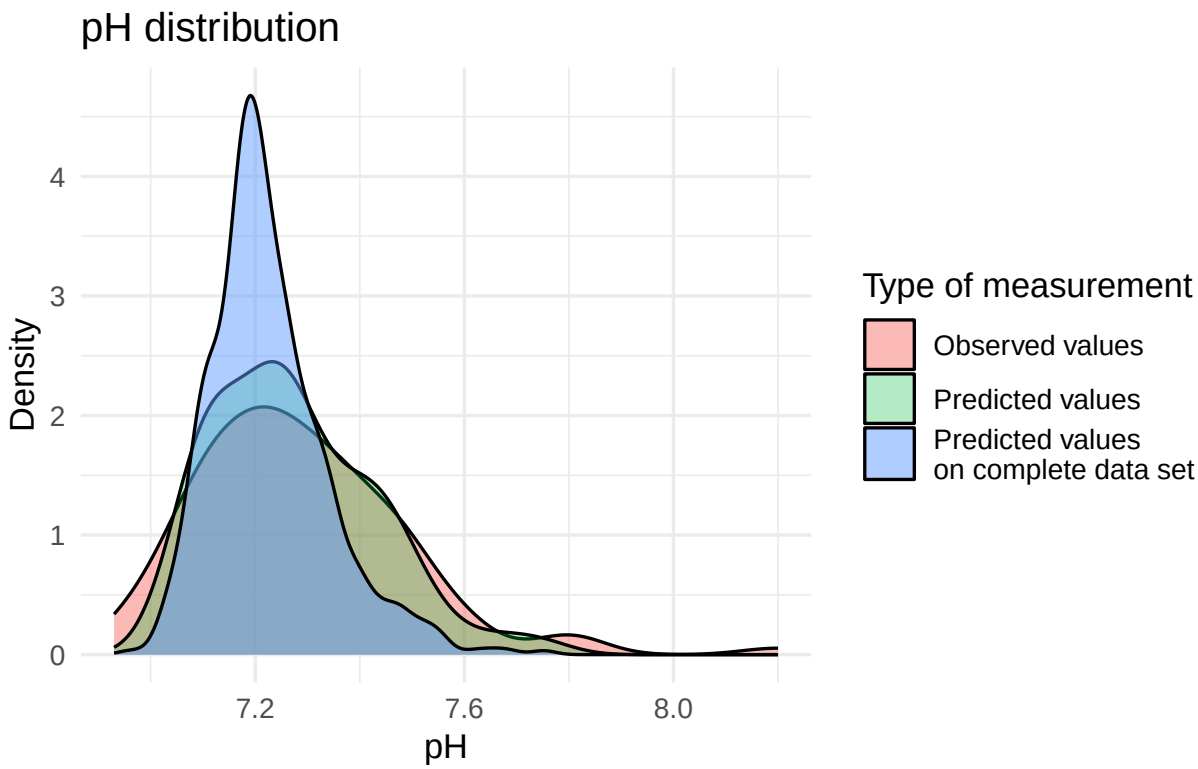

Figure S48: Density plot of pH values.

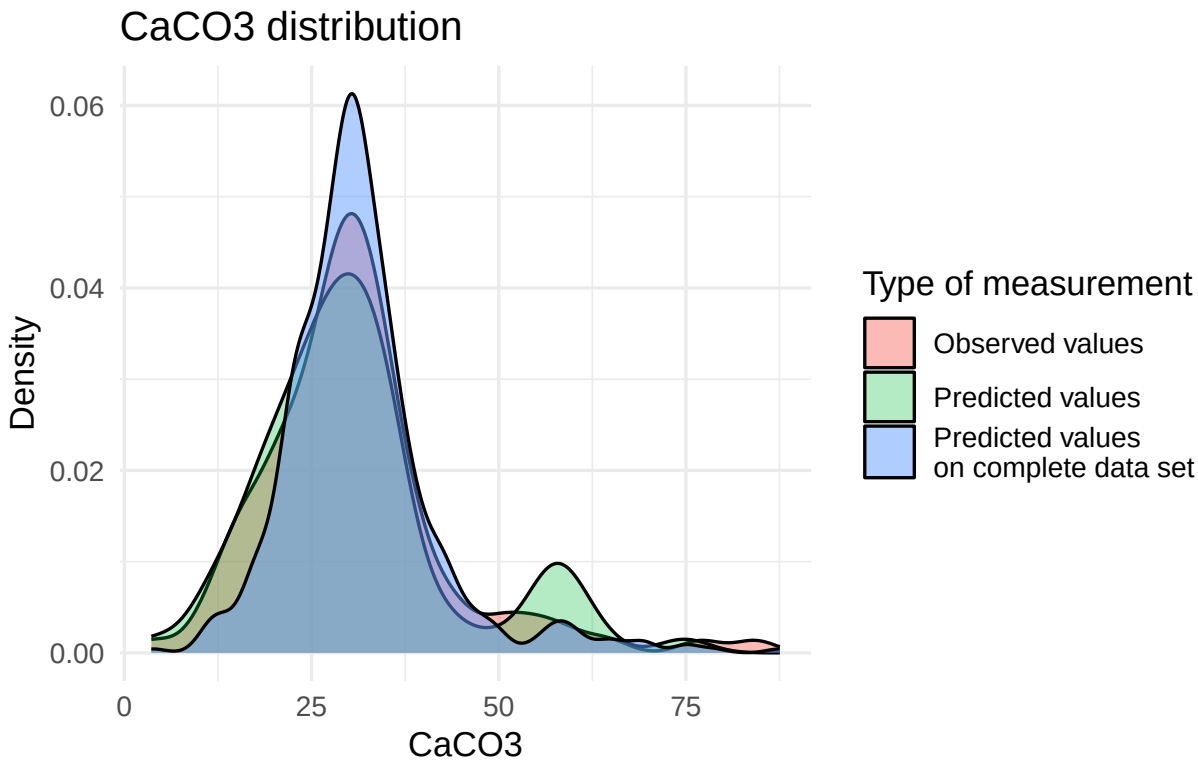

Figure S49: Density plot of CaC03 values.

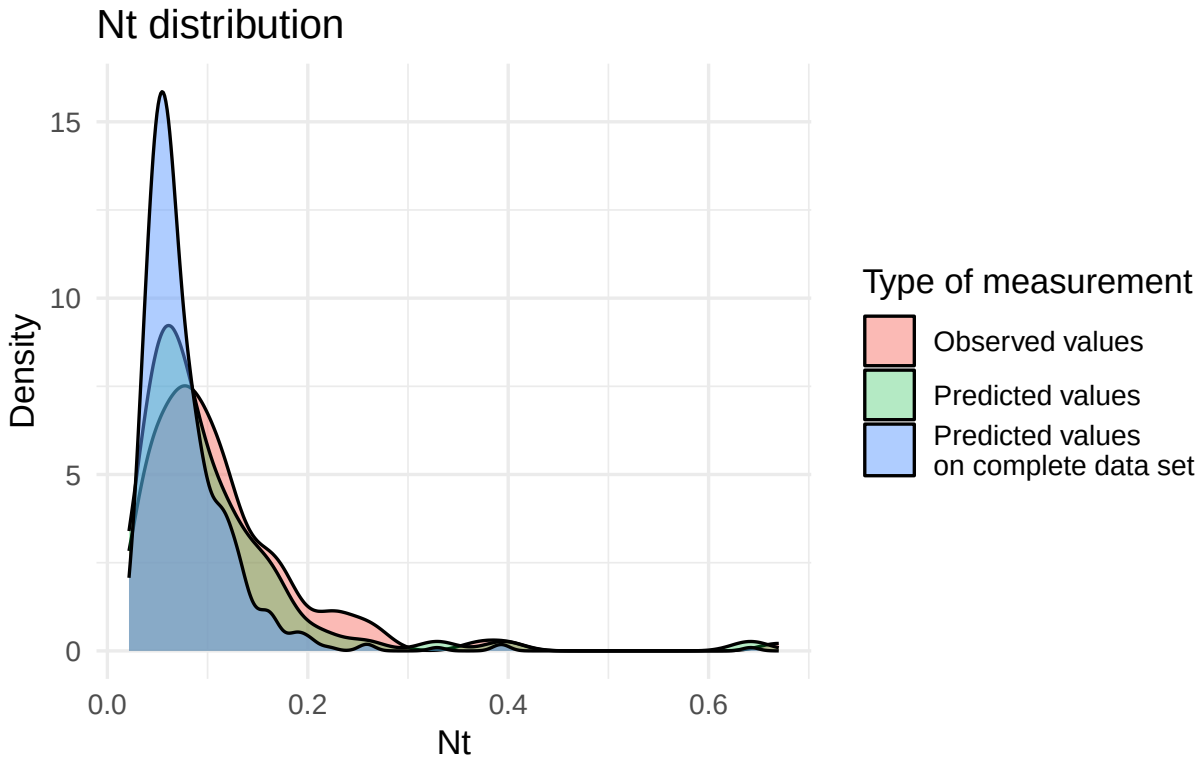

Figure S50: Density plot of Nt values.

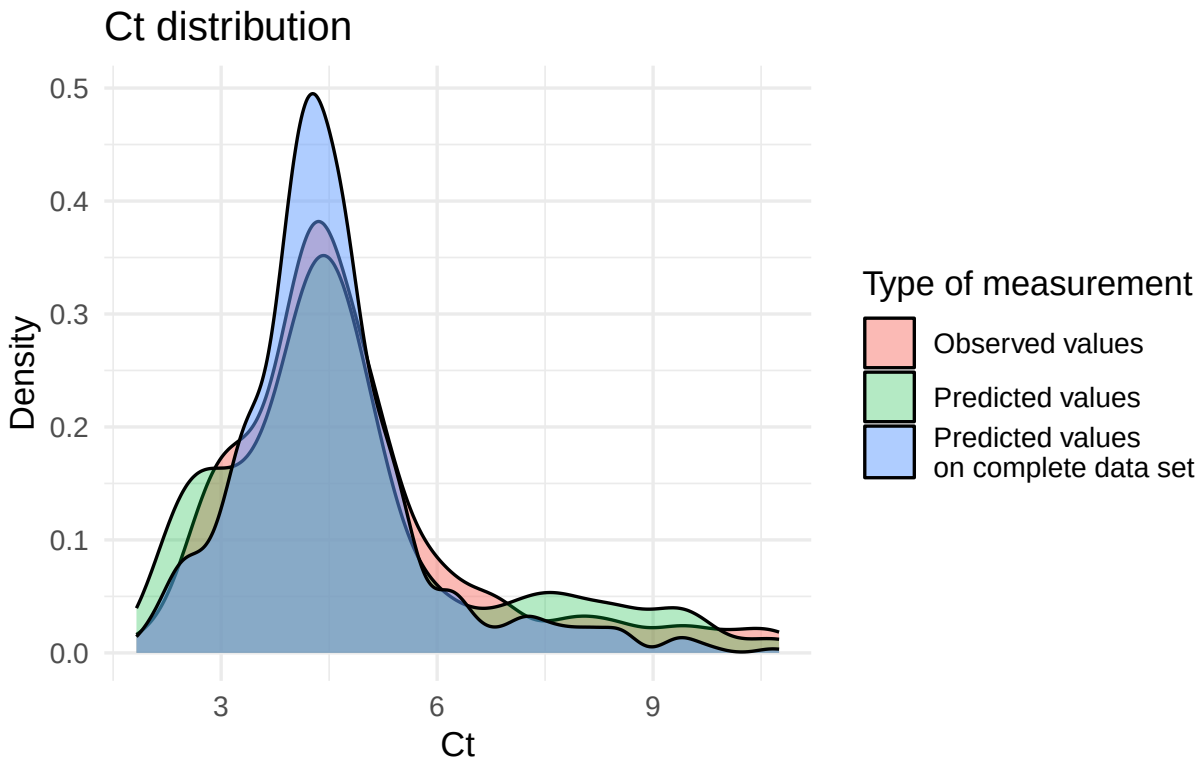

Figure S51: Density plot of Ct values.

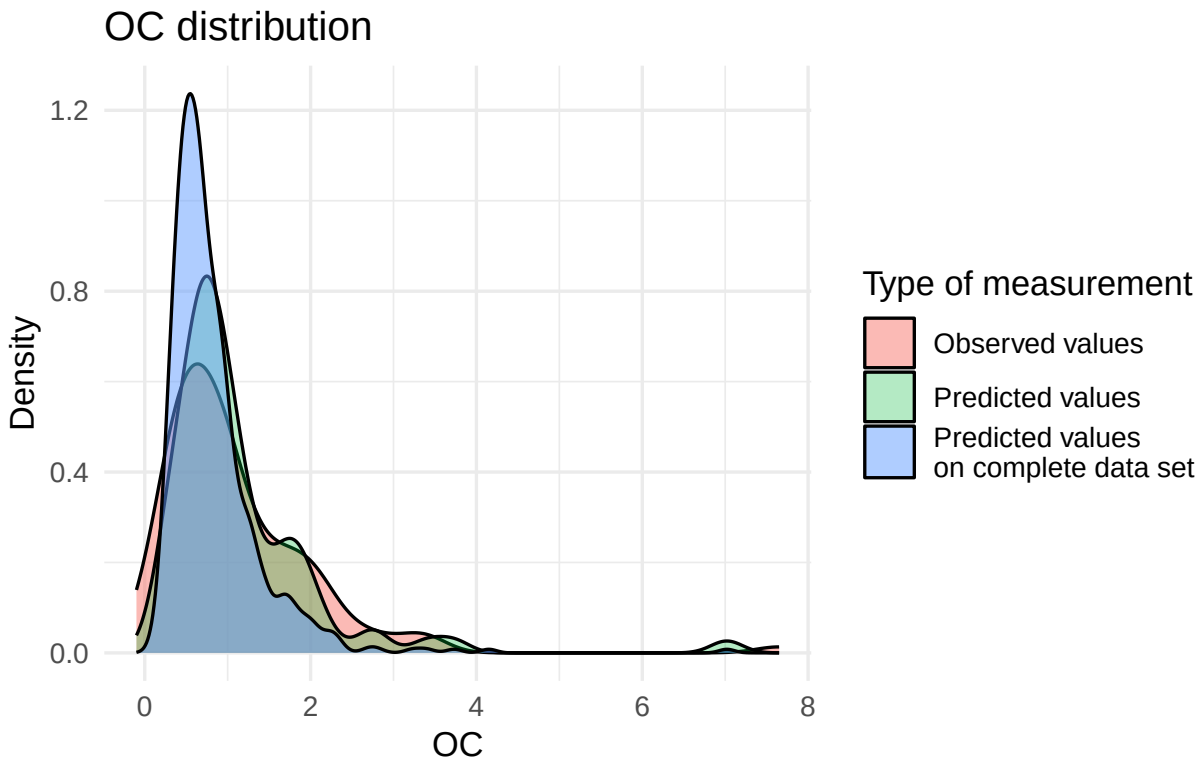

Figure S52: Density plot of OC values.

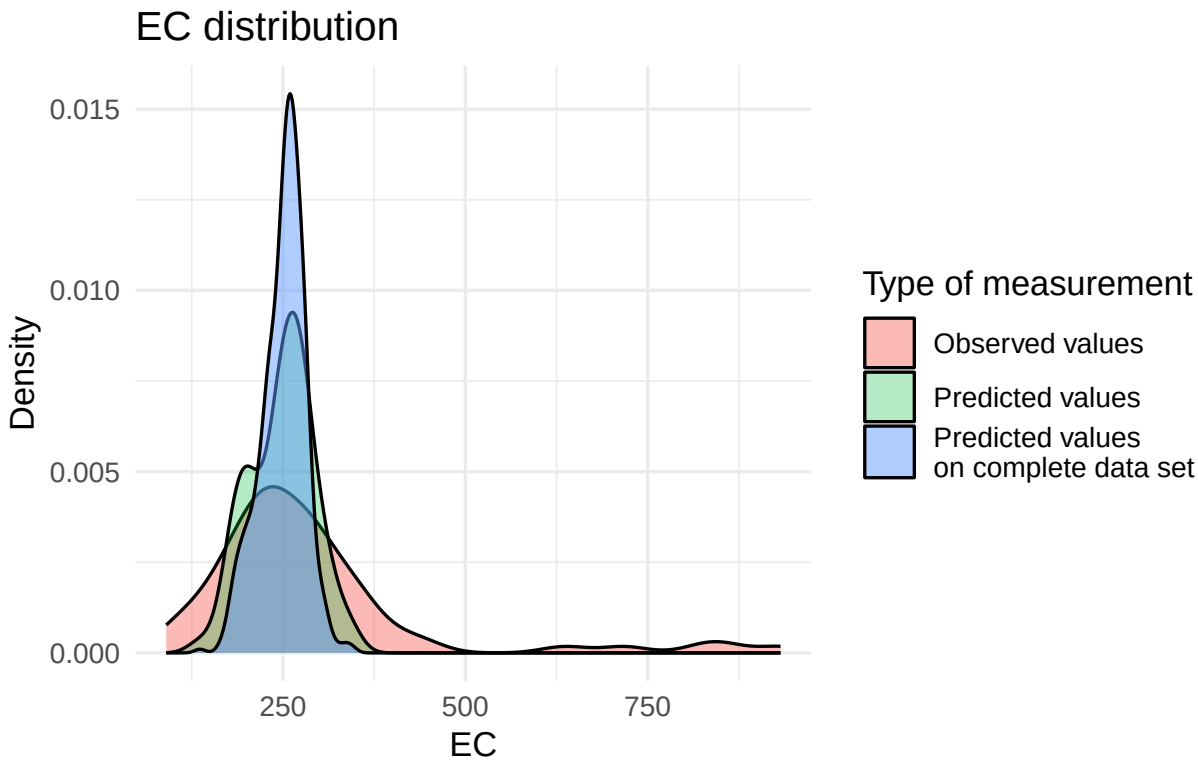

Figure S53: Density plot of EC values.

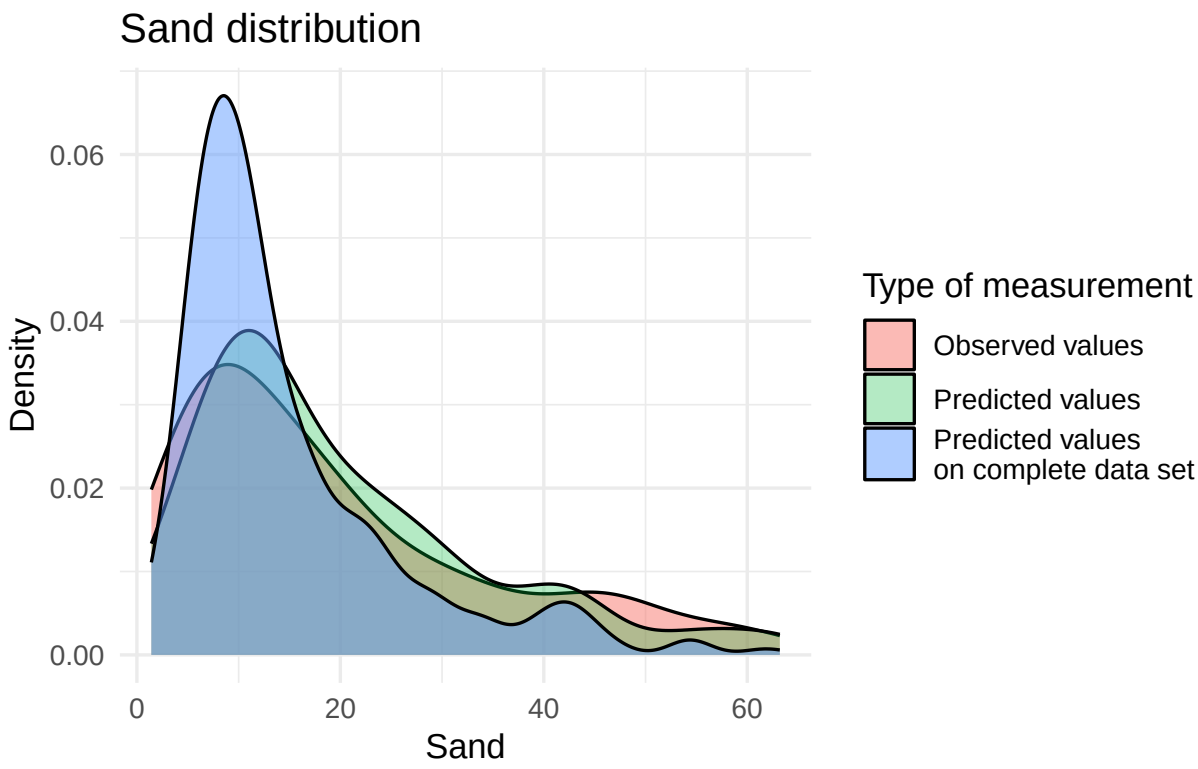

Figure S54: Density plot of Sand values.

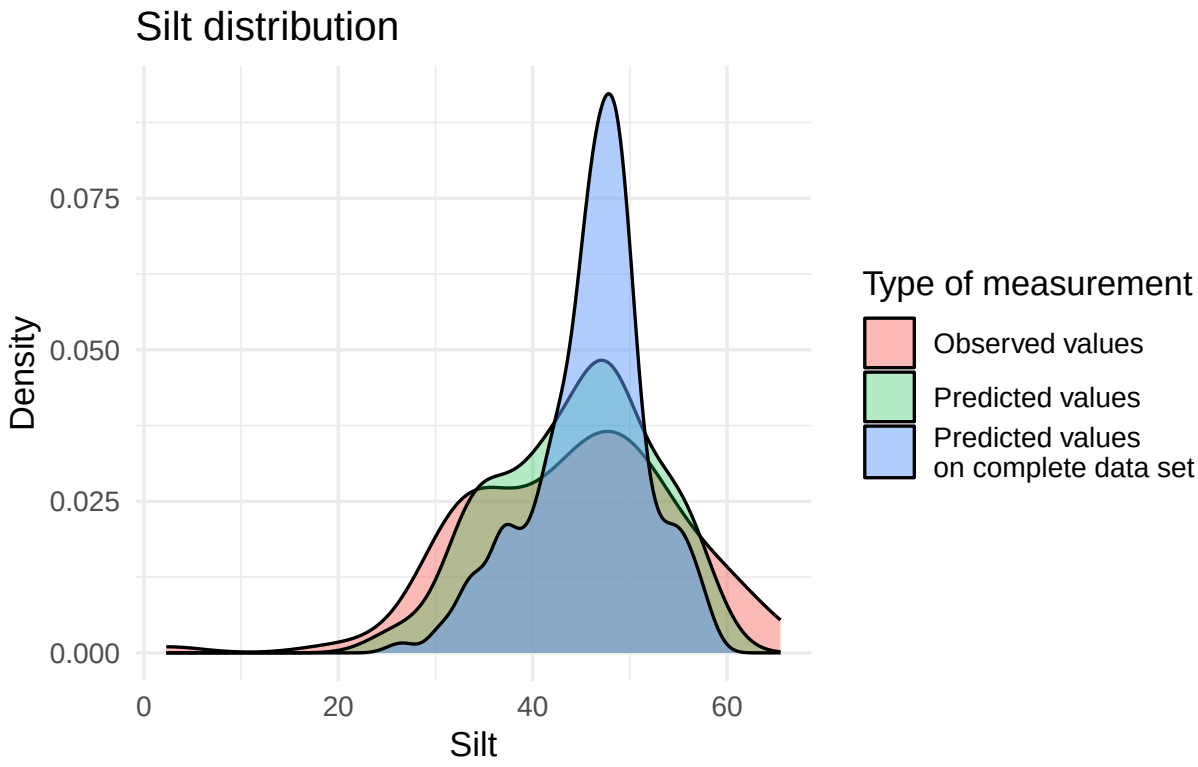

Figure S55: Density plot of Silt values.

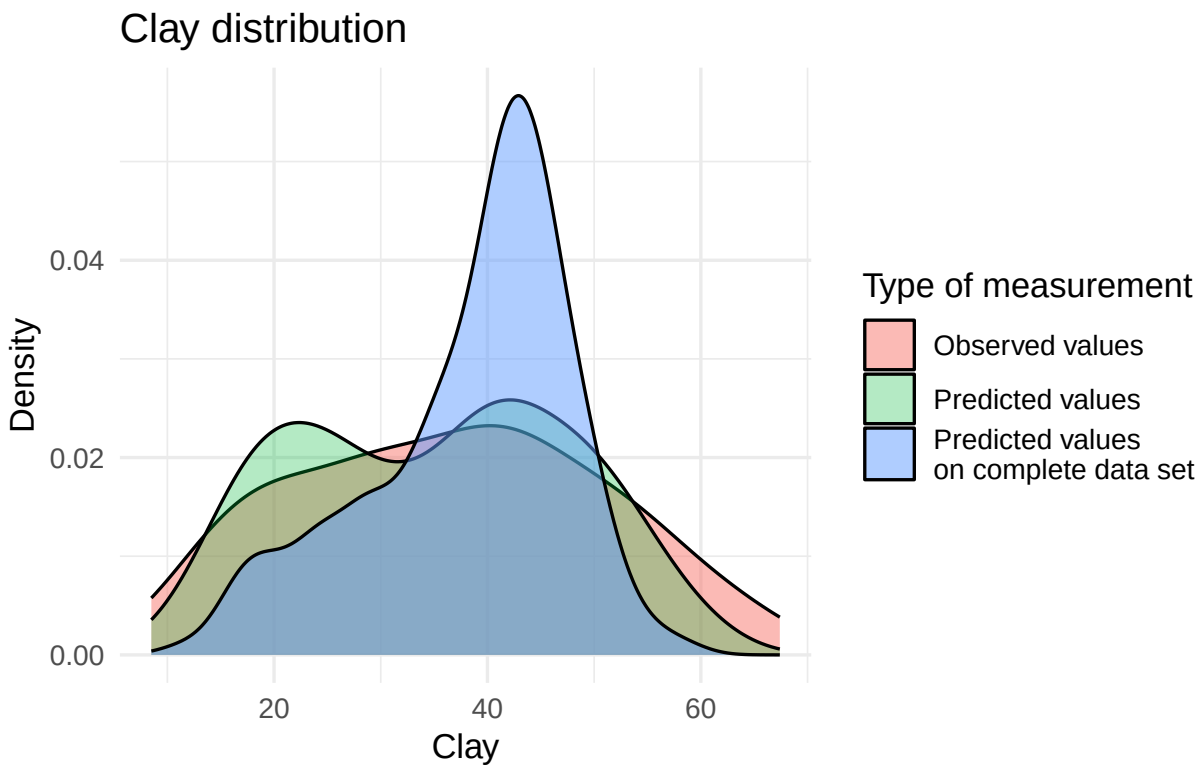

Figure S56: Density plot of Clay values.

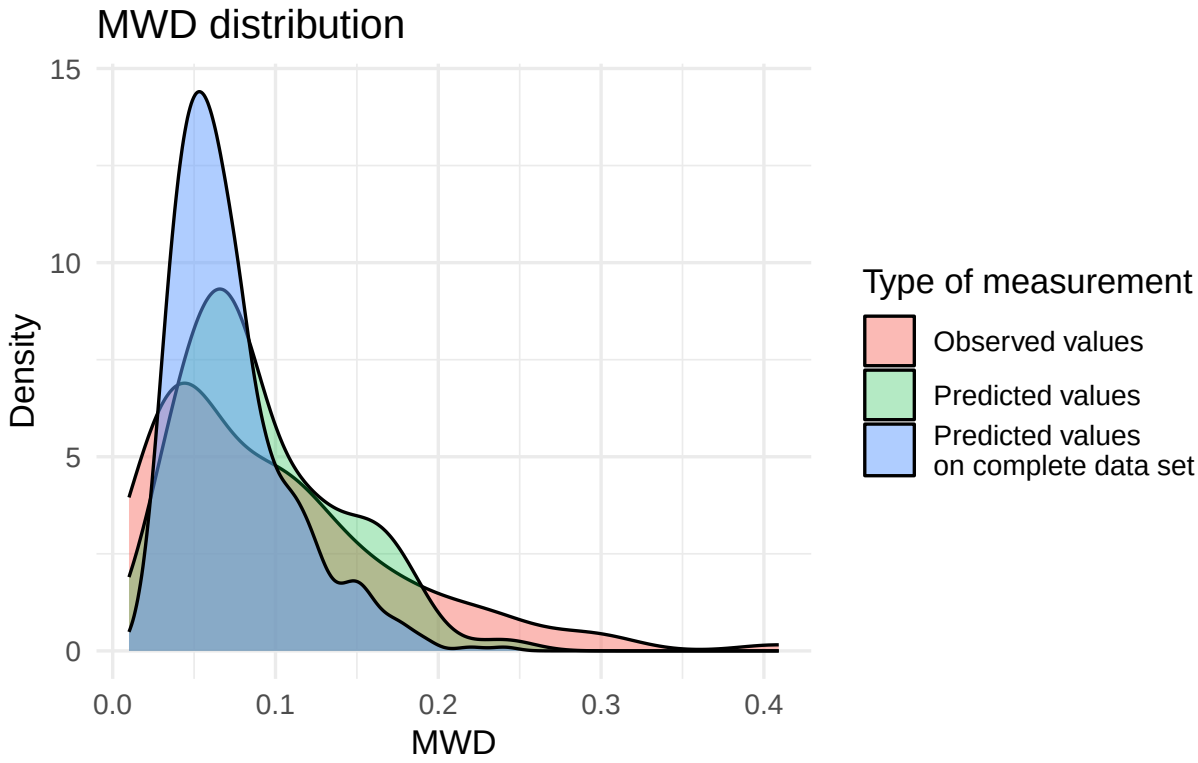

Figure S57: Density plot of MWD values.

Digital soil mapping

Introduction

Purpose

We present here the methodology for the digital soil mapping based on the soil values from the spectra prediction.

Covariates

We used a set of 85 covariates mainly based on Zolfaghari Nia et al. (2022). All the data listed below are available freely online excepted for the digitized maps (geology, geohydrology and geomorphology).

Table S12: Data used in the production of the digital soil mapping.

Name/ID	Original resolution (m)	Type/Unit
Landsat 8 Blue/LA.1	30	0.45 - 0.51 µm
Landsat 8 Green/LA.2	30	0.53 - 0.59 µm
Landsat 8 NDVI/LA.3	30	-
Landsat 8 NDWI/LA.4	30	-
Landsat 8 NIR/LA.5	30	0.85 - 0.88 µm
Landsat 8 Panchromatic/LA.6	15	0.52 - 0.90 µm
Landsat 8 Red/LA.7	30	0.64 - 0.67 µm
Landsat 8 SWIR1/LA.8	30	1.57 - 1.65 µm
Landsat 8 SWIR2/LA.9	30	2.11 - 2.29 µm
Landsat 8 EVI/LA.10	30	-
Landsat 8 SAVI/LA.11	30	-
Landsat 8 NDMI/LA.12	30	-
Landsat 8 CORSI/LA.13	30	-
Landsat 8 Brightness index/LA.14	30	-
Landsat 8 Clay index/LA.15	30	-
Landsat 8 Salinity index/LA.16	30	-
Landsat 8 Carbonate index/LA.17	30	-
Landsat 8 Gypsum index/LA.18	30	-
MODIS EVI/MO.1	300	-
MODIS LST day/MO.2	1000	°C
MODIS LST night/MO.2	1000	°C
MODIS NDVI/MO.4	300	-
MODIS NIR/MO.5	300	0.841 - 0.876 µm
MODIS Red/MO.6	300	0.62 - 0.67 µm
MODIS SAVI/MO.7	300	Meters
MODIS Brightness index/MO.8	300	35 classes
Distance rivers/OT.1	30	17 classes
Geology/OT.2	1 : 300 000	11 classes
Geomorphology/OT.3	1 : 300 000	mm
Landuses/OT.4	10	mm
PET sum/OT.5	750	Kj m^{-2}
Prec. sum/OT.6	1000	°C
SRAD sum/OT.7	1000	m s^{-1}
Diff Max. Min. Temp./OT.8	1000	0.492 - 0.496 µm
Wind sum/OT.9	1000	0.559 - 0.560 µm
Sentinel 2 Blue/SE.1	10	-
Sentinel 2 Green/SE.2	10	-
Sentinel 2 NDVI/SE.3	20	0.833 - 0.835 µm
Sentinel 2 NDWI/SE.4	20	0.665 - 0.664 µm

Name/ID	Original resolution (m)	Type/Unit
Sentinel 2 NIR/SE.5	10	0.738 - 0.739 µm
Sentinel 2 Red/SE.6	10	0.739 - 0.740 µm
Sentinel 2 RedEdge1/SE.7	20	0.779 - 0.782 µm
Sentinel 2 RedEdge2/SE.8	20	1.610 - 1.613 µm
Sentinel 2 RedEdge3/SE.9	20	2.185 - 2.202 µm
Sentinel 2 SWIR1/SE.10	20	0.943 - 0.945 µm
Sentinel 2 SWIR2/SE.11	20	-
Sentinel 2 water vapor/SE.12	90	-
Sentinel 2 EVI/SE.13	20	-
Sentinel 2 TVI/SE.14	20	-
Sentinel 2 SAVI/SE.15	20	-
Sentinel 2 LSWI/SE.16	20	-
Sentinel 2 Clay index/SE.17	20	-
Sentinel 2 Brightness index/SE.18	20	-
Sentinel 2 Salinity index/SE.19	20	-
Sentinel 2 Carbonate index/SE.20	20	Radians
Sentinel 2 Gypsum index/SE.21	20	-
Aspect/TE.1	30	-
Channel network base level/TE.2	30	-
Channel network distance/TE.3	30	Meters
Connexity/TE.4	30	-
DEM/TE.5	30	-
Flow accumaltion/TE.6	30	-
General curvature/TE.7	30	-
MrRTF/TE.8	30	Radians
MrVBF/TE.9	30	-
Negative openness/TE.10	30	-
Normalized height/TE.11	30	Radians
Plan curvature/TE.12	30	-
Positive openness/TE.13	30	-
Profile curvature/TE.14	30	Radians
Slope height/TE.15	30	-
Slope/TE.16	30	-
Standardized height/TE.17	30	-
Surface landform/TE.18	30	-
Terrain ruggedness index/TE.19	30	-
Terrain texture/TE.20	30	-
TPI/TE.21	30	-
TWI/TE.22	30	-
Total catchment area/TE.23	30	0.45 - 0.51 µm

We create an Python API to access the Google Earth Engine platform. You can read to the 'ipynb' file for more info.

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Terrain

For the DEM and all the derivatives, we used *SAGA GIS 9.3.1* software and all the specificties of the batch process are detailed in the code. The last *LS factor* corresponds to the *Total catchment area* covariate.

Remote sensing images and indexes

The Landsat 8 images were collected *via* a Google Earth Engine script on a period covering 2020, the median of the composite image from *Tier 1 TOA* collection was used. The Sentinel 2 image were collected *via* a Google Earth Engine script on a period covering 2021, the median of the composite image from *MultiSpectral Instrument Level-2A* collection was used.The land surface temperature (LST) and other MODIS component were computed also on Google Earth with a time covering from 2020 to 2021. the median of the *MODIS Terra* collection was used. The javascript codes for scraping these images are available in the supplementary file inside the "7 - DSM/code" folder. We computed the following indexes: Normalized Difference Vegetation Index (**NDVI**); Normalized Difference Water Index (**NDWI**); Enhanced Vegetation Index (**EVI**); Soil Adjusted Vegetation Index (**SAVI**); Transformed Vegetation Index (**TVI**); Normilized Difference Moisture Index (**NDMI**); COmbined Specteral Response Index (**COSRI**); Land Surface Water Index (**LSWI**).

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

Rouse et al. (1974)

$$NDWI = \frac{Green - NIR}{Green + NIR}$$

McFeeters (1996)

$$EVI = G\frac{NIR - Red}{(NIR + 6Red - 7.5Blue) + L}$$

Where G is 2.5 and L is 1 A. Huete, Justice, and Liu (1994)

$$SAVI = (1 + L)\frac{NIR - Red}{NIR + Red + L}$$

Where L is 0.5. A. R. Huete (1988)

$$TVI = \sqrt{NDVI + 0.5}$$

Deering (1975)

$$NDMI = \frac{NIR - SWIR1}{NIR + SWIR1}$$

Gao (1996)

$$COSRI = \frac{Blue + Green}{Red + NIR} * NDVI$$

Fernández-Buces et al. (2006)

$$LSWI = \frac{NIR - SWIR1}{NIR + SWIR1}$$

Chandrasekar et al. (2010)

$$Brightness\ index = \sqrt{Red^2 + NIR^2}$$

Khan et al. (2001)

$$Simple\ Ratio\ Clay\ index = \frac{SWIR1}{SWIR2}$$

Bousbih et al. (2019)

$$Salinity\ index = \frac{SWIR1 - SWIR2}{SWIR1 - NIR}$$

Also called NIR-SWIR index (NSI) Abuelgasim and Ammad (2019)

$$Carbonate\ index = \frac{Red}{Green}$$

Boettinger et al. (2008)

$$Gypsum\ index = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2}$$

Nield, Boettinger, and Ramsey (2007)

Soil properties

The soil 10 variables measurements for the five soil depth increment came from the predictions of the previous chapter and can be found online at https://doi.org/10.1594/PANGAEA.973700.

Preparation of the data

All raster were sampled to 30 x 30 m tiles to match the DEM. We used **bilinear** method excepted for the discrete maps (geology and geomorphology) where we used **ngb** resampling. The heavier data from GEE were also called from the GoogleDrive and WorldClim directly collected from 'geodata' package. The the vector data were also transformed in raster beeing resampled.

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Soil digital mapping preparation

Preparation of the environment

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Prepare the data

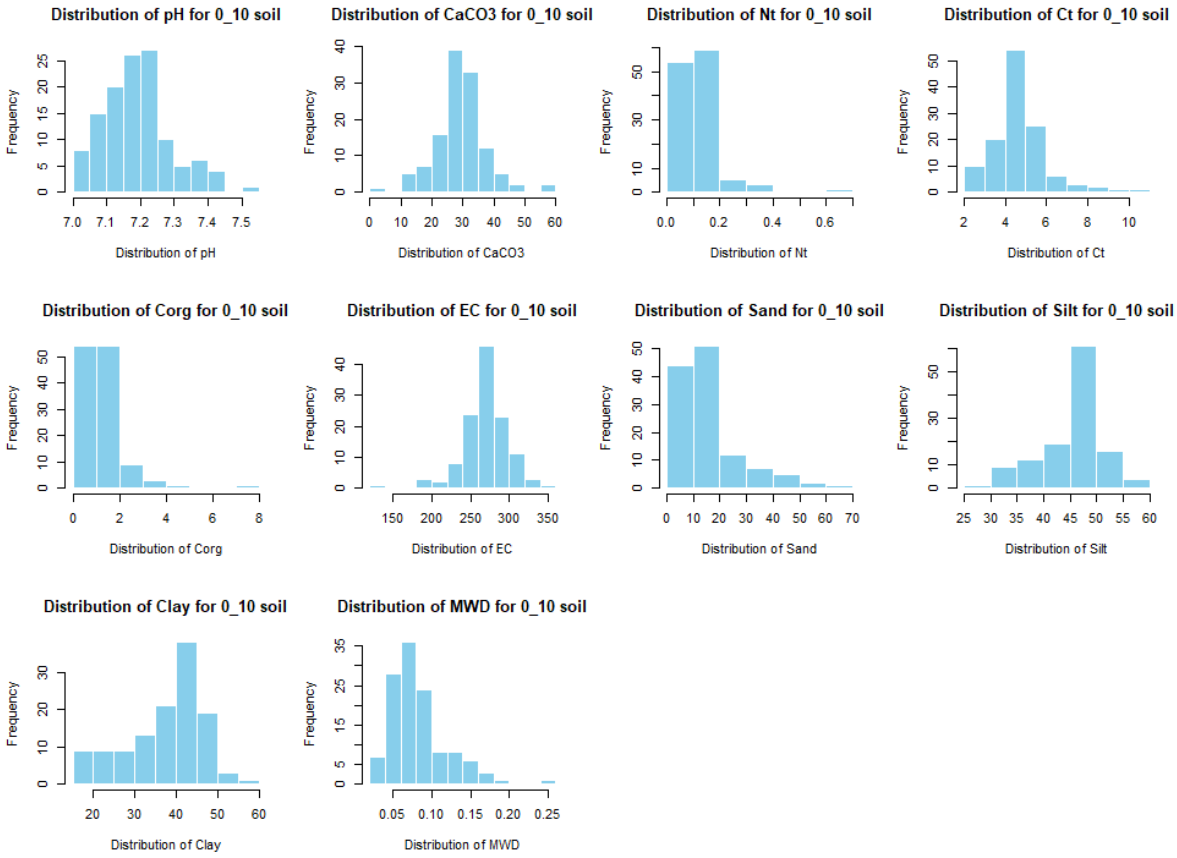

Figure S58: Histogram for 0 - 10 cm.

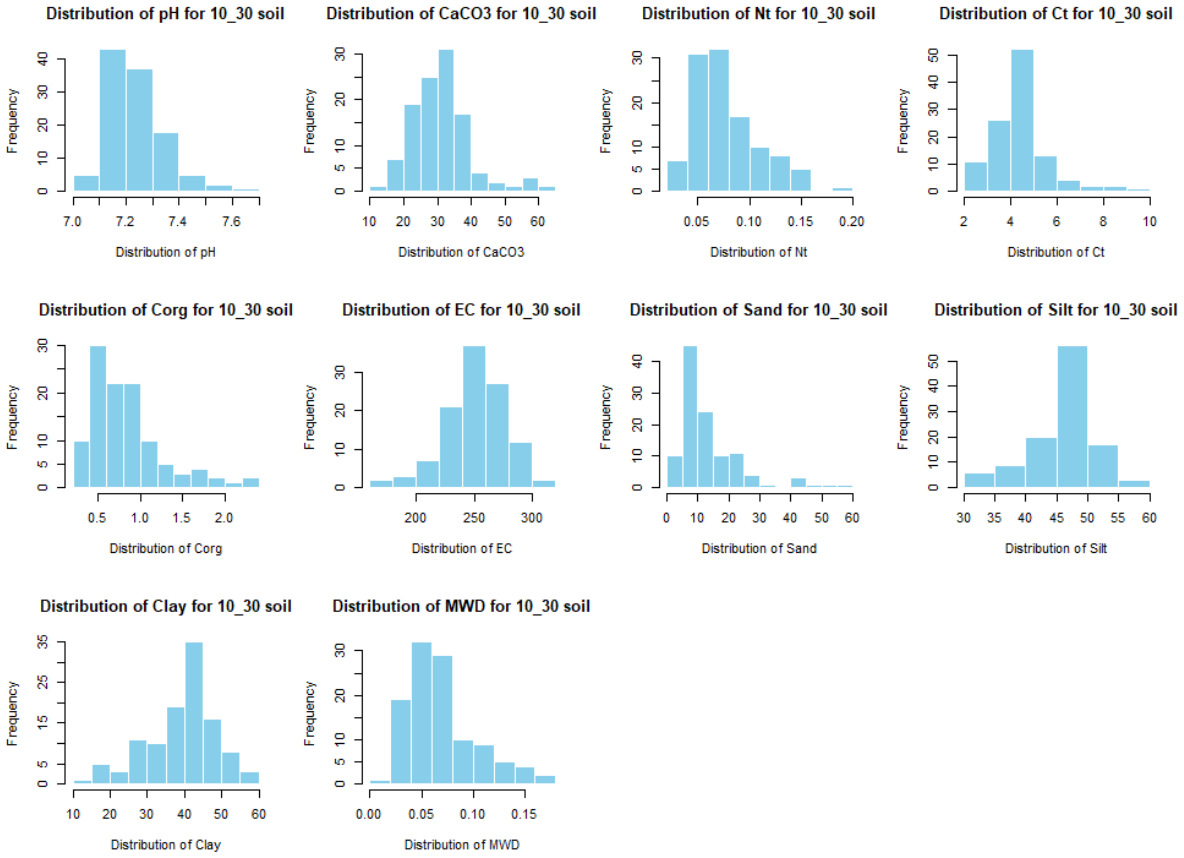

Figure S59: Histogram for 10 - 30 cm.

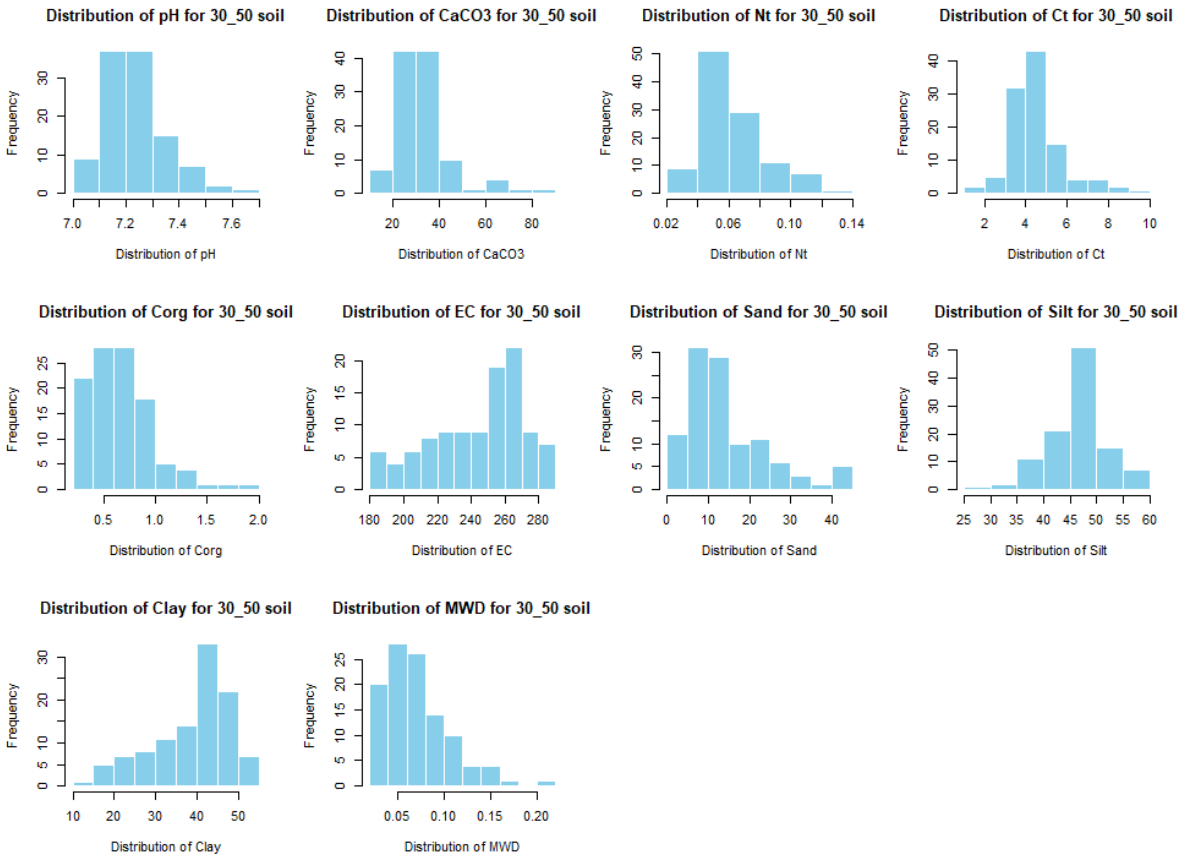

Figure S60: Histogram for 30 - 50 cm.

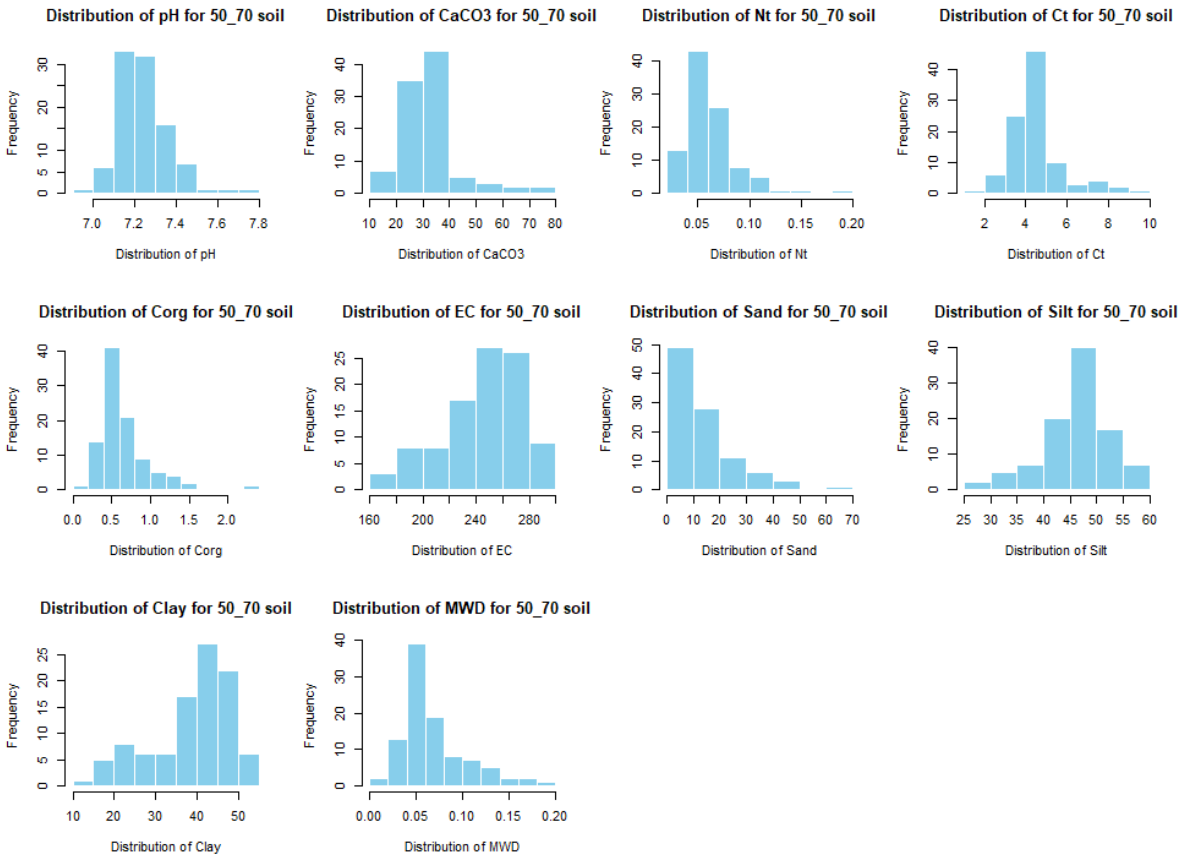

Figure S61: Histogram for 50 - 70 cm.

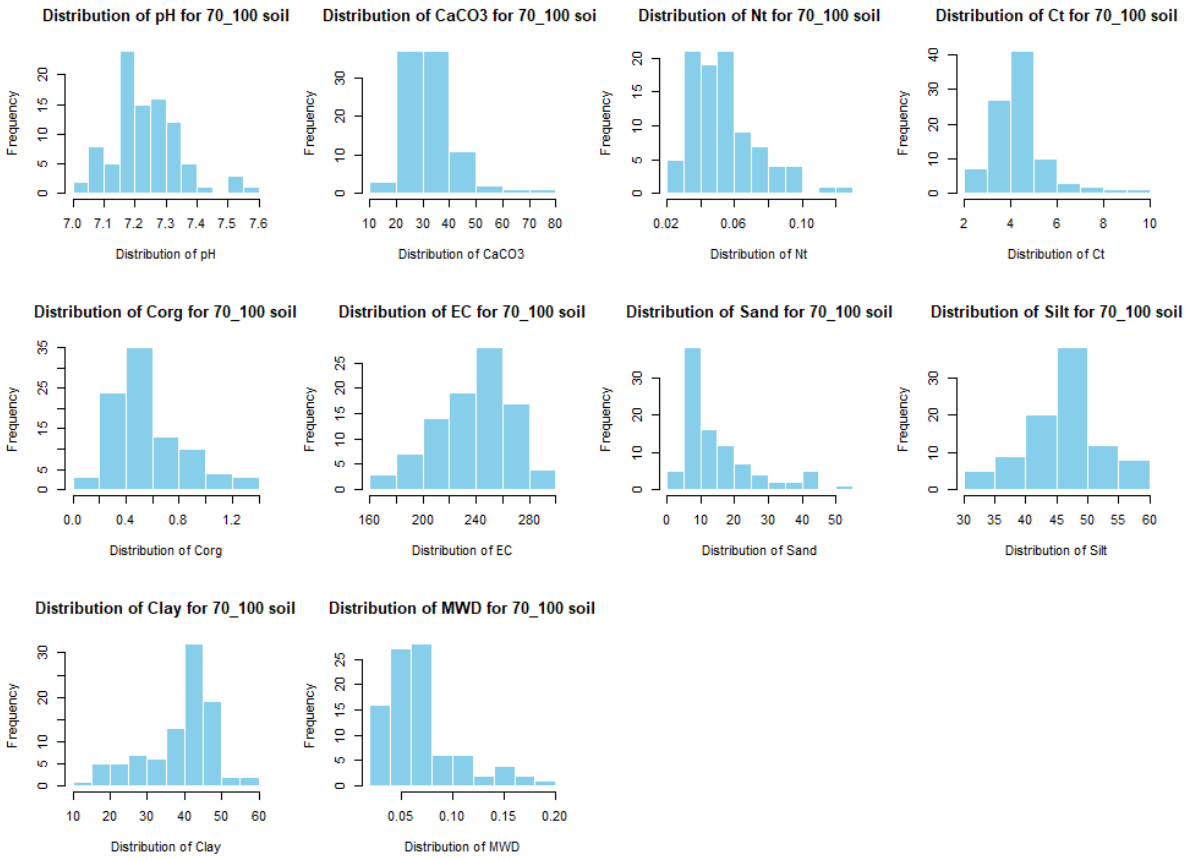

Figure S62: Histogram for 70 - 100 cm.

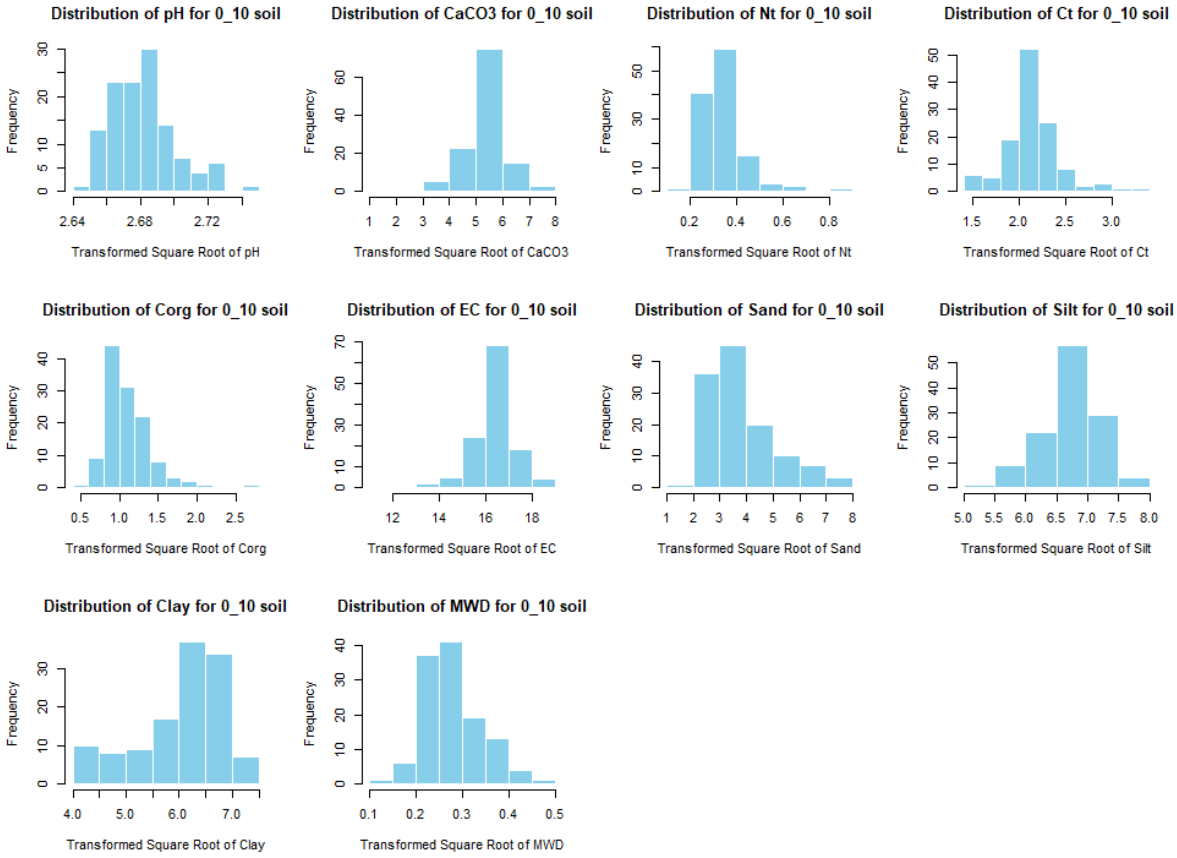

Figure S63: Histogram SQRTfor 0 - 10 cm.

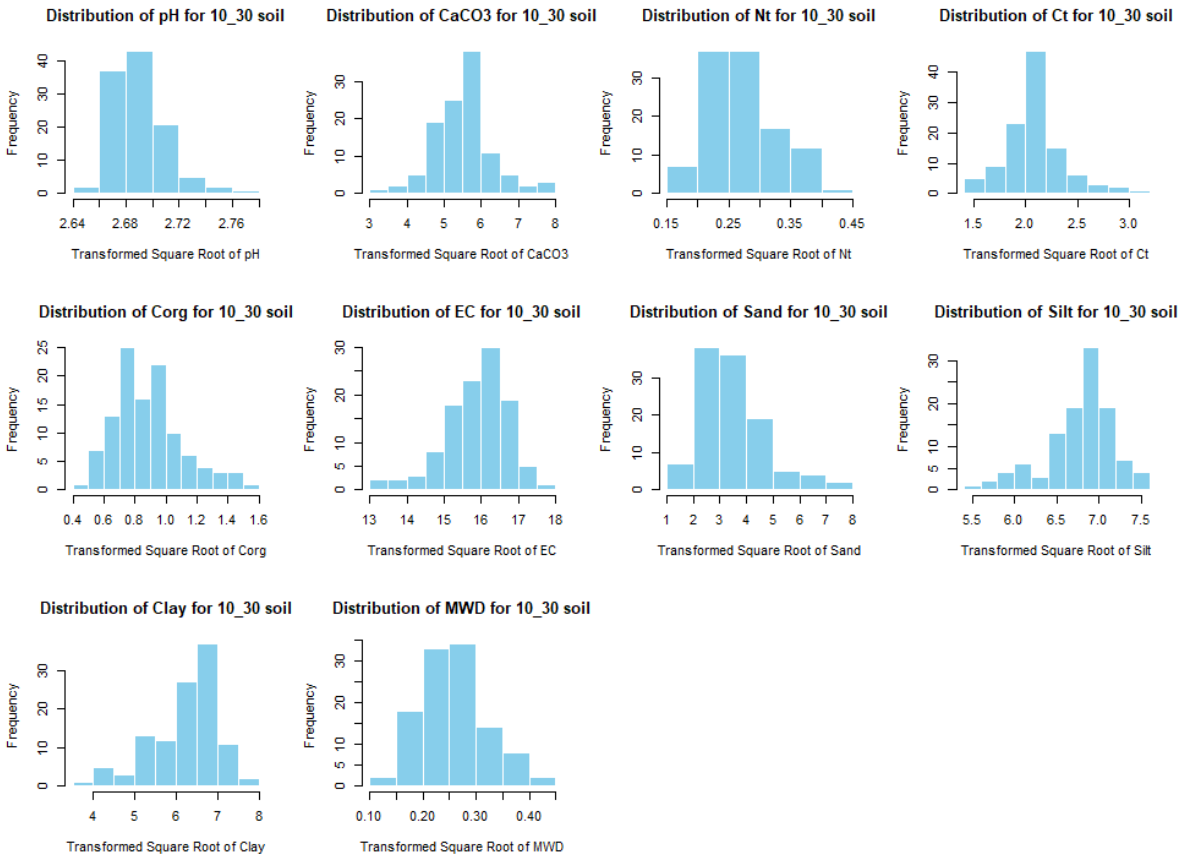

Figure S64: Histogram SQRT for 10 - 30 cm.

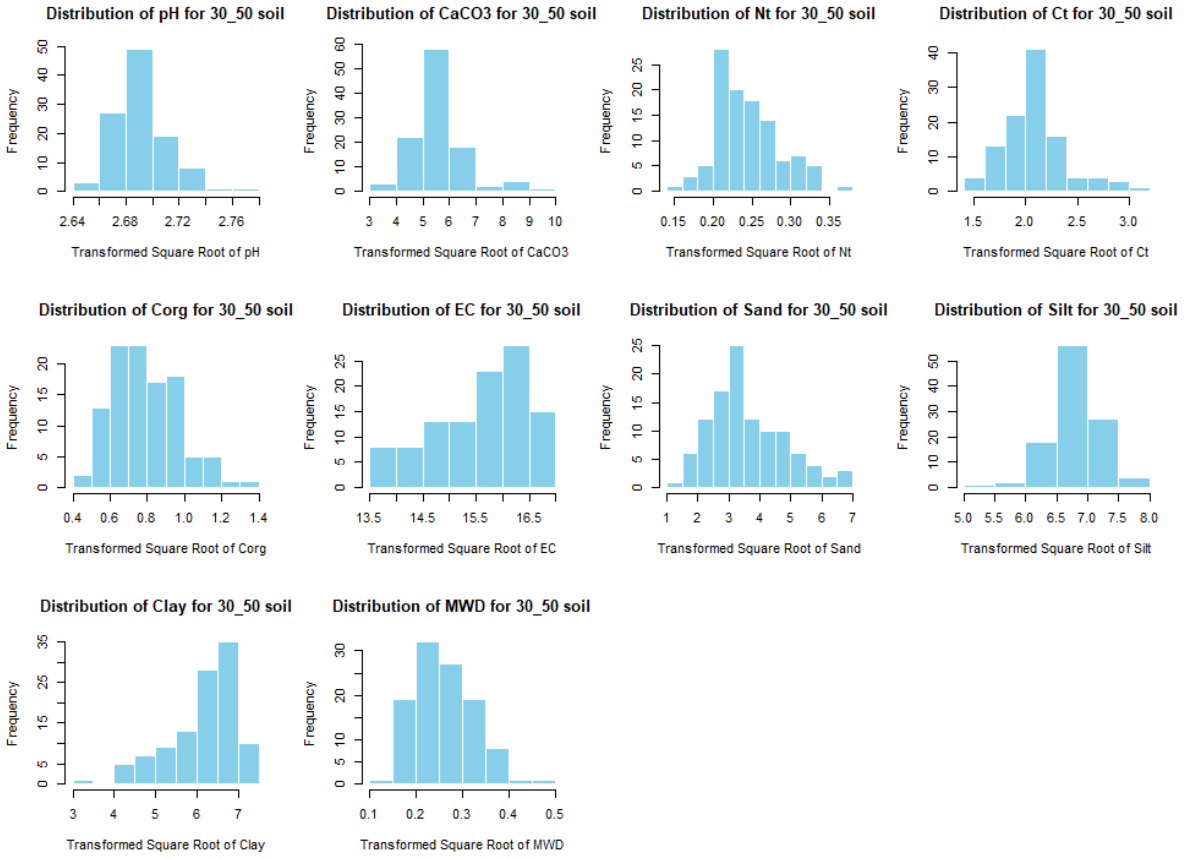

Figure S65: Histogram SQRT for 30 - 50 cm.

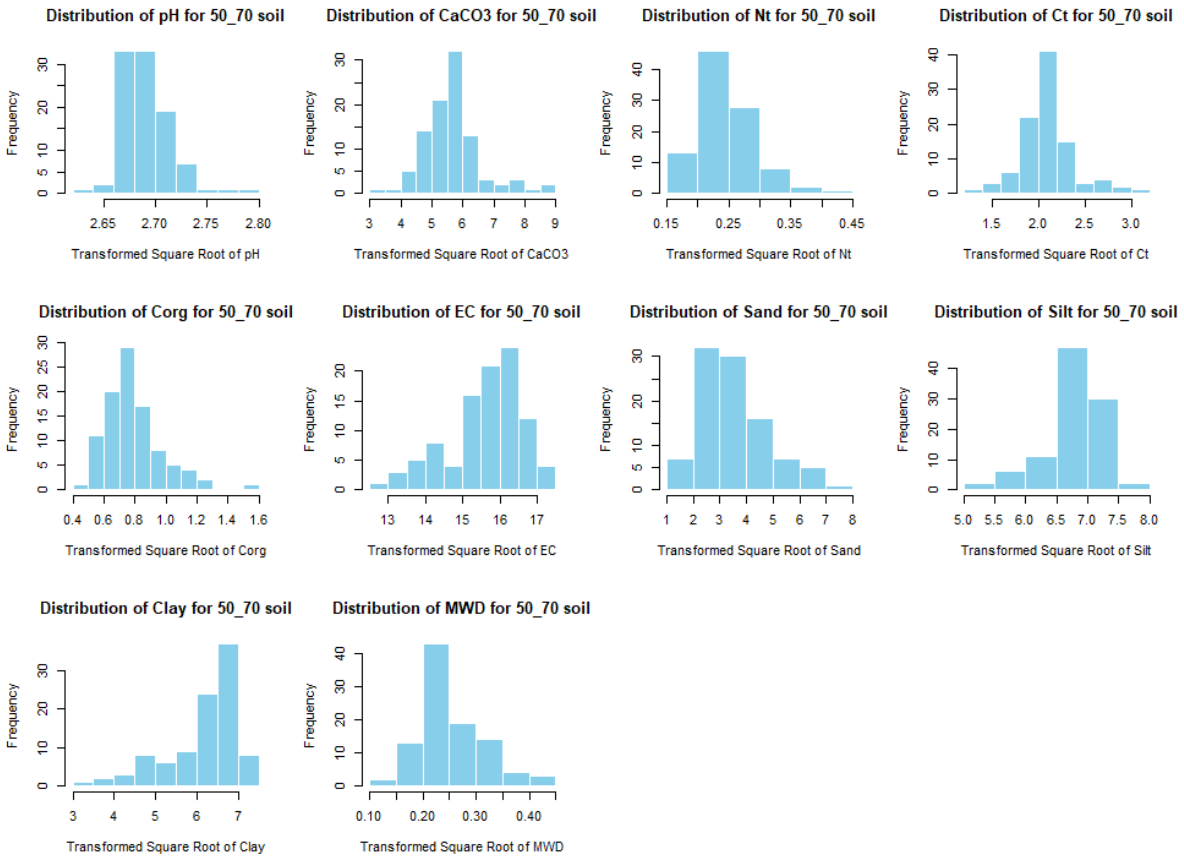

Figure S66: Histogram SQRT for 50 - 70 cm.

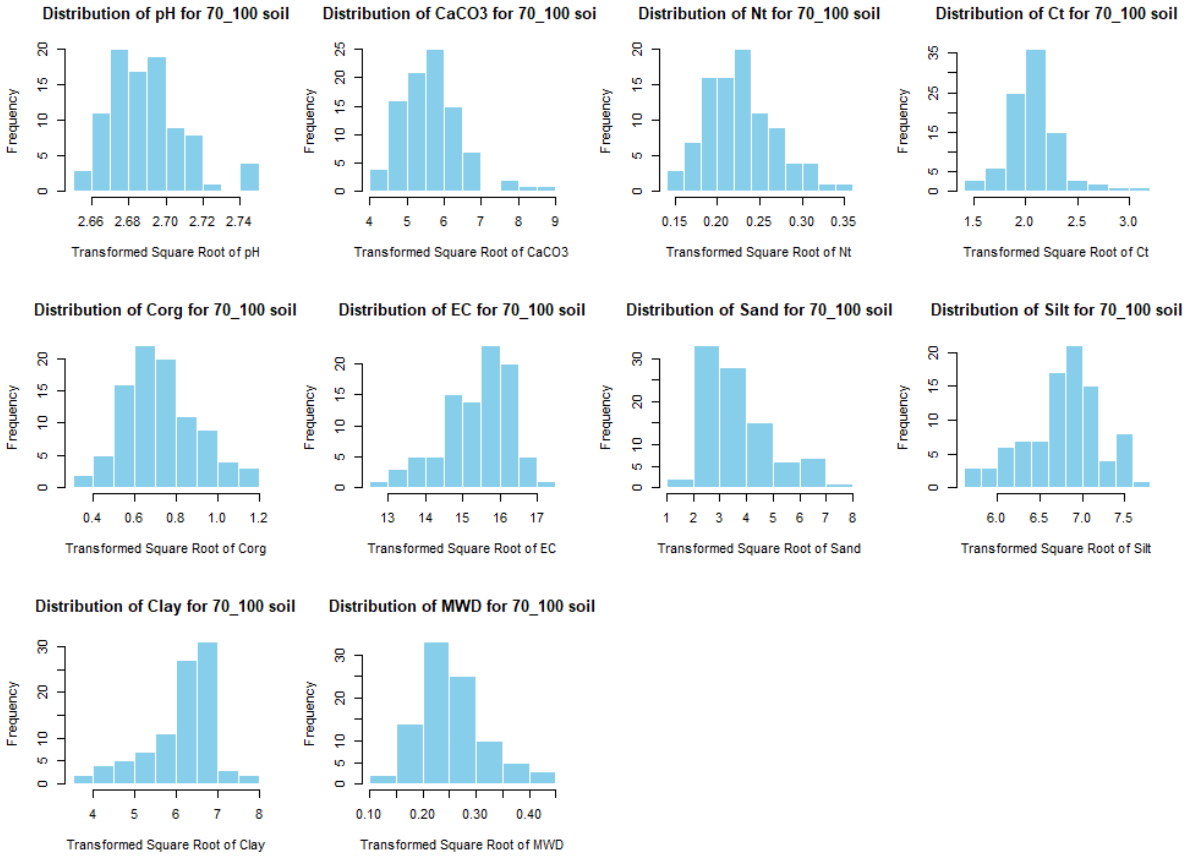

Figure S67: Histogram SQRT for 70 - 100 cm.

Prepare the covariates

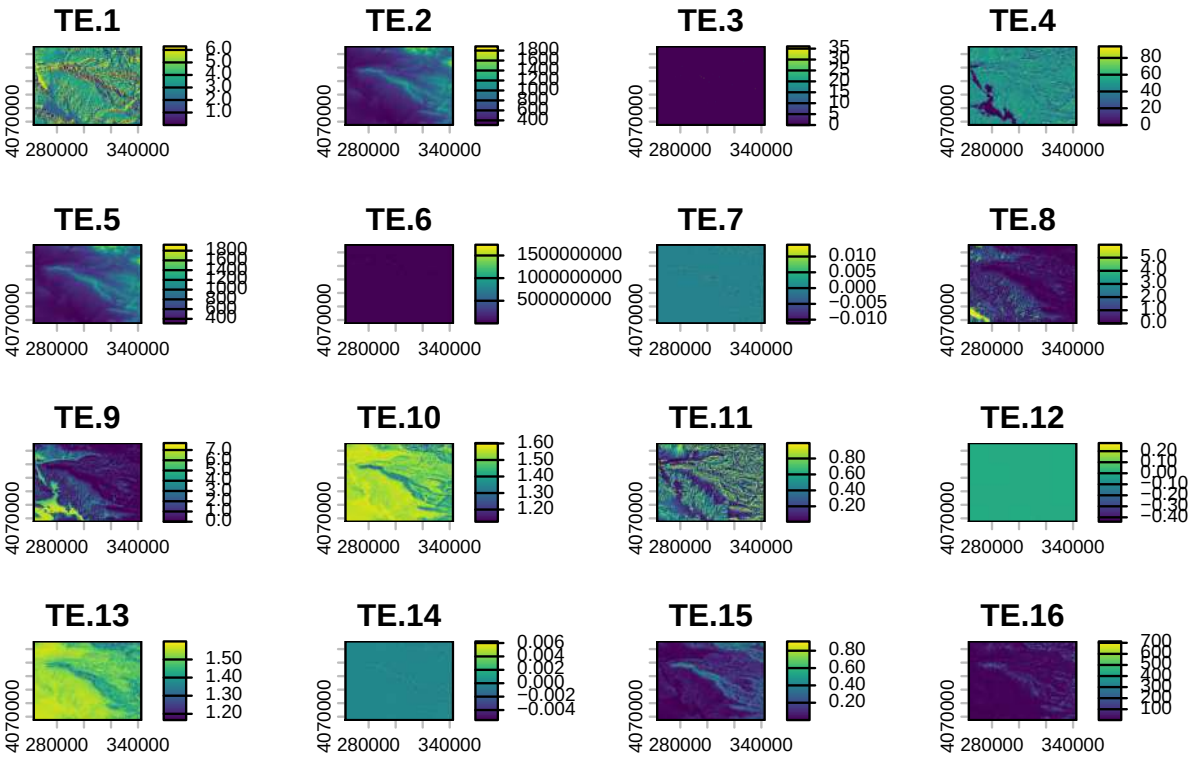

Figure S68: Covariates from DSM.

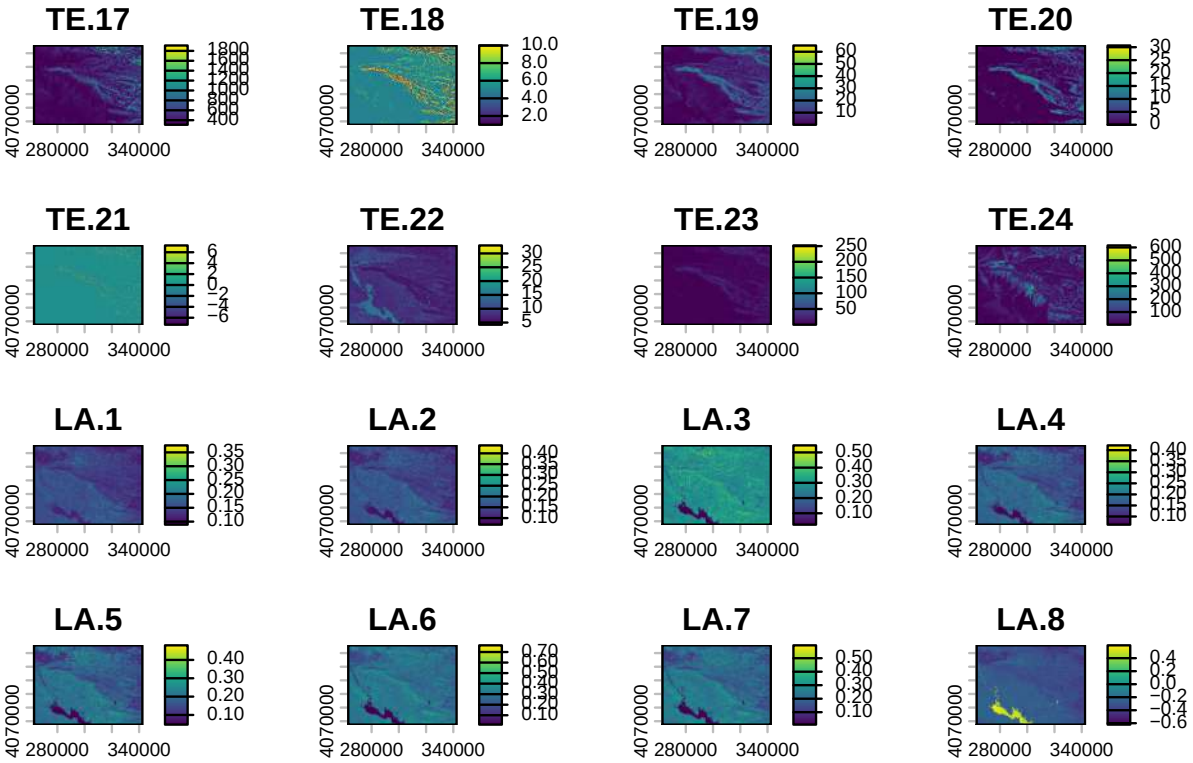

Figure S69: Covariates from DSM.

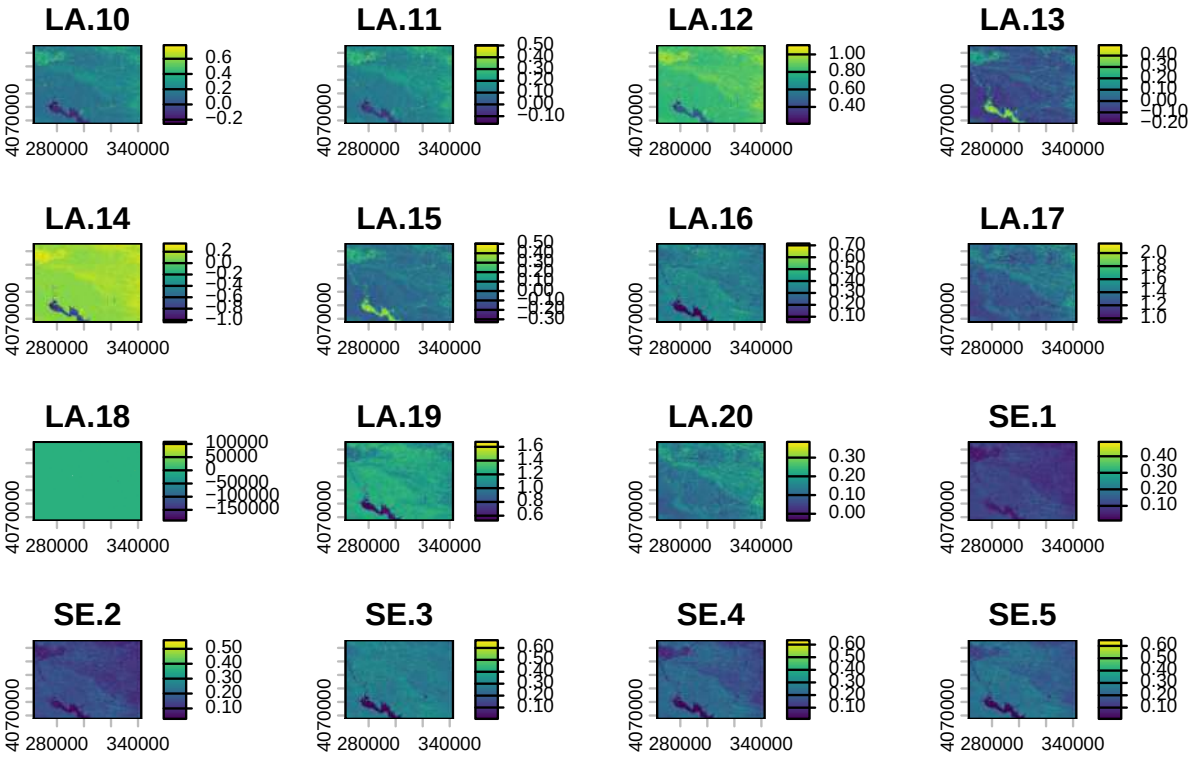

Figure S70: Covariates from DSM.

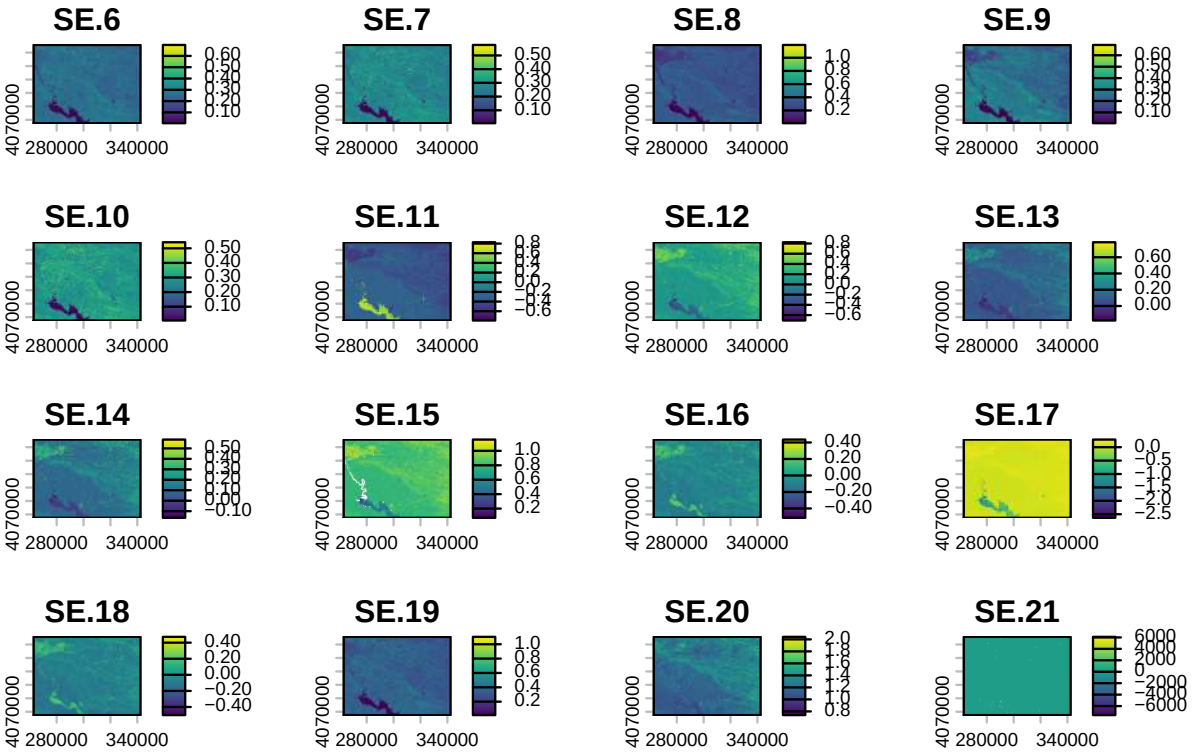

Figure S71: Covariates from DSM.

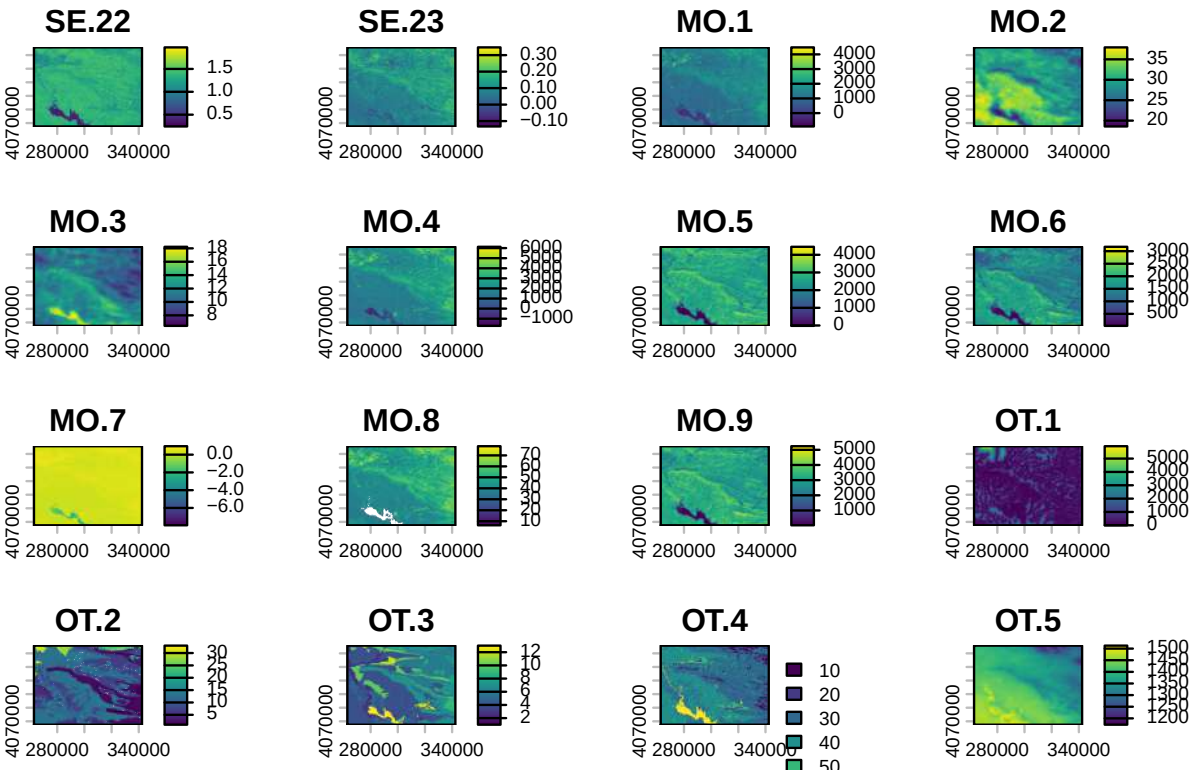

Figure S72: Covariates from DSM.

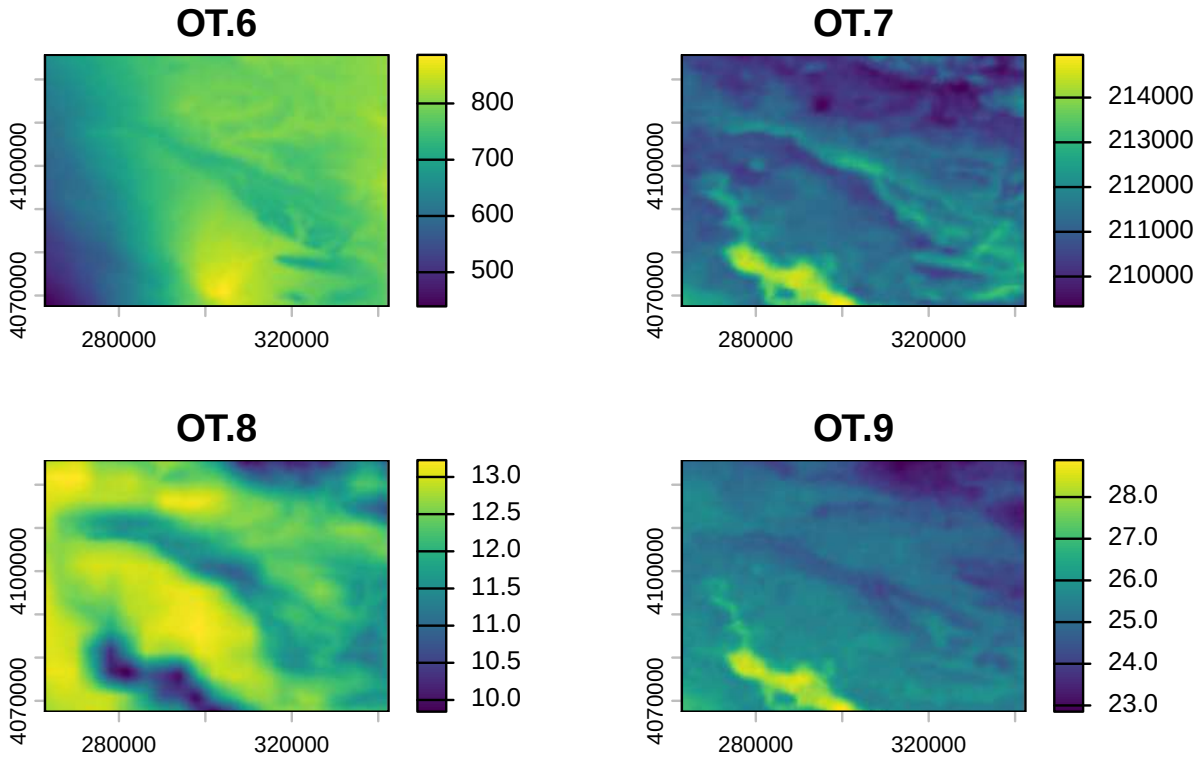

Figure S73: Covariates from DSM.

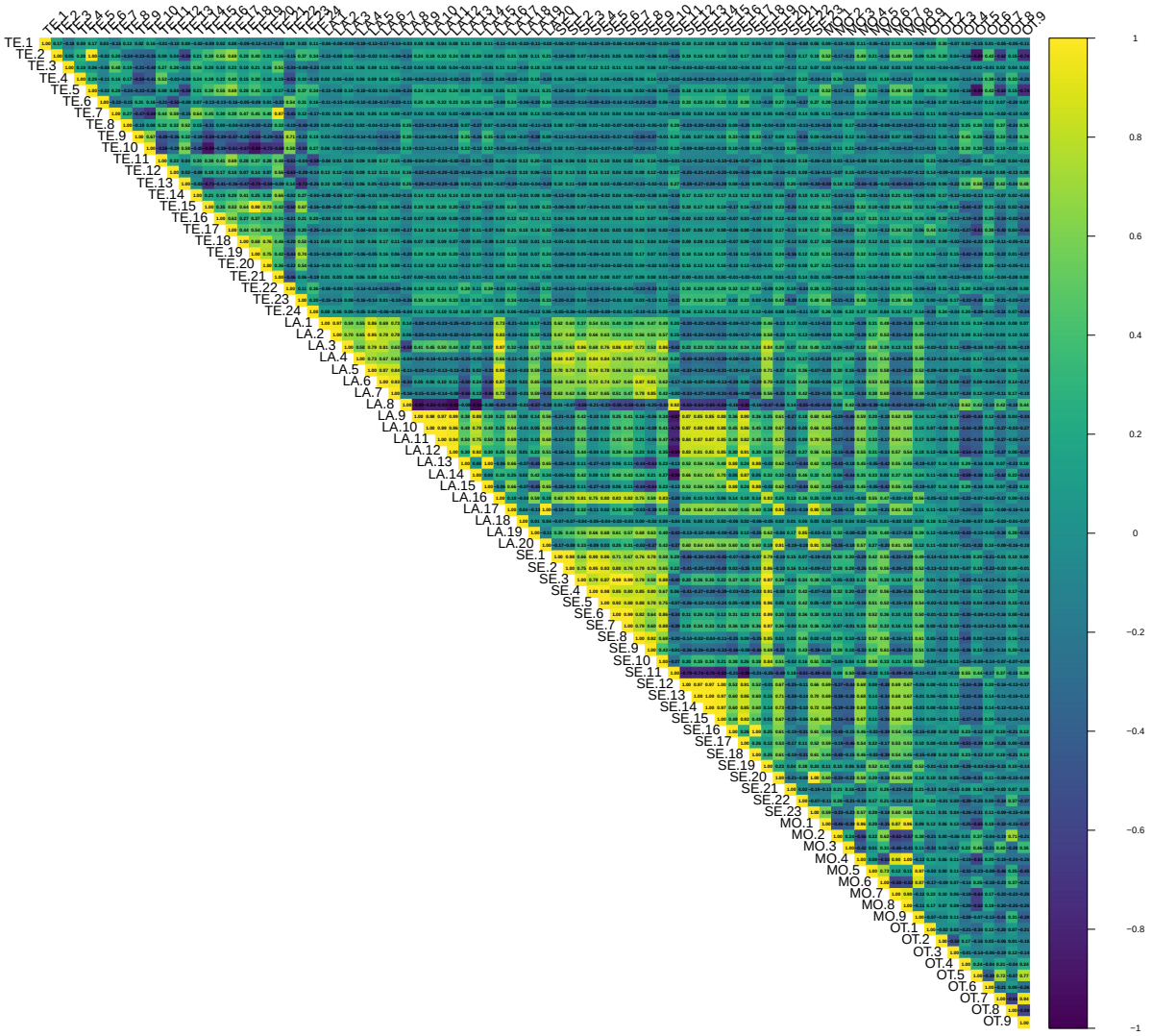

Figure S74: Auto-correlation plot 0_10.

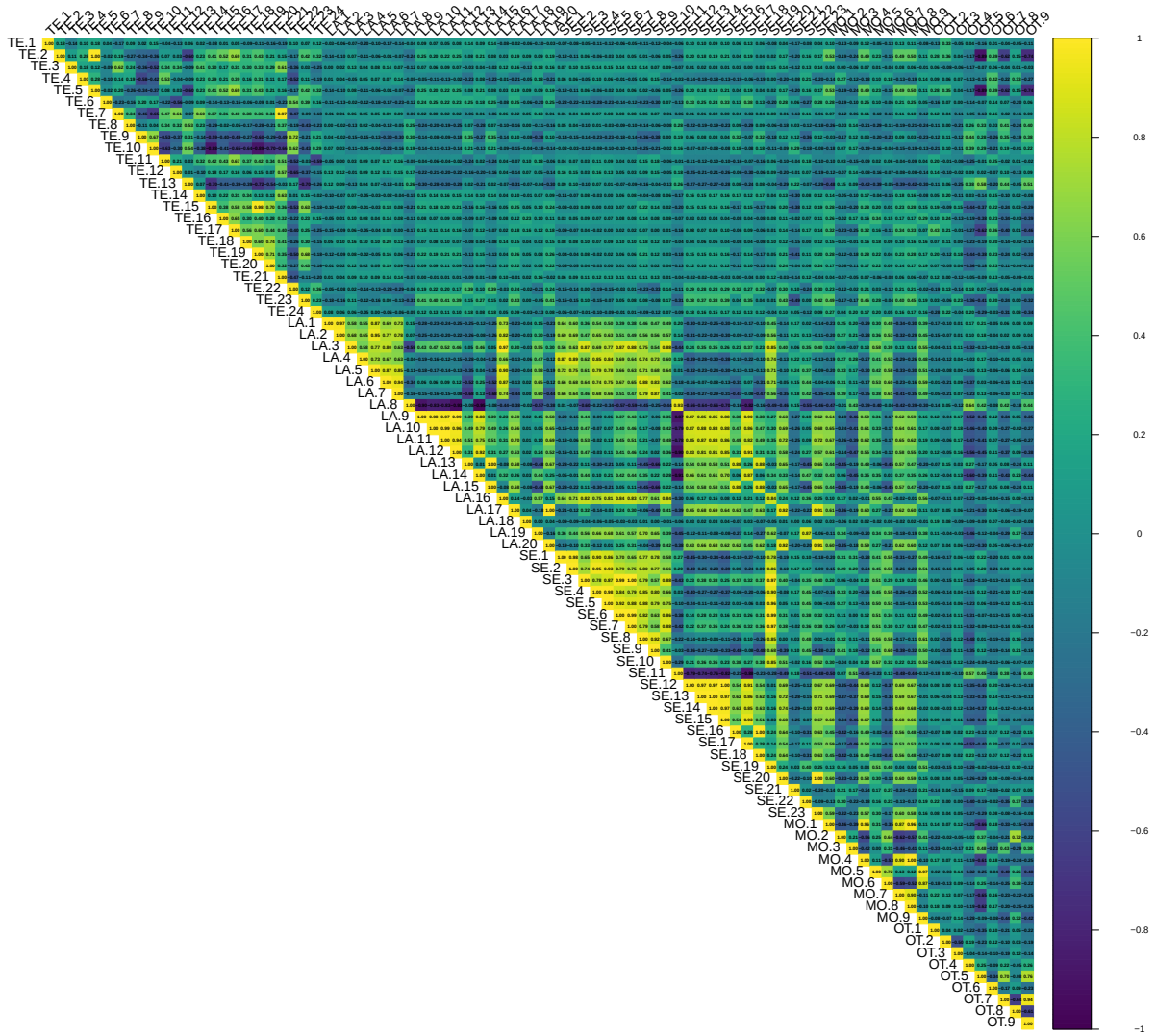

Figure S75: Auto-correlation plot 10_30.

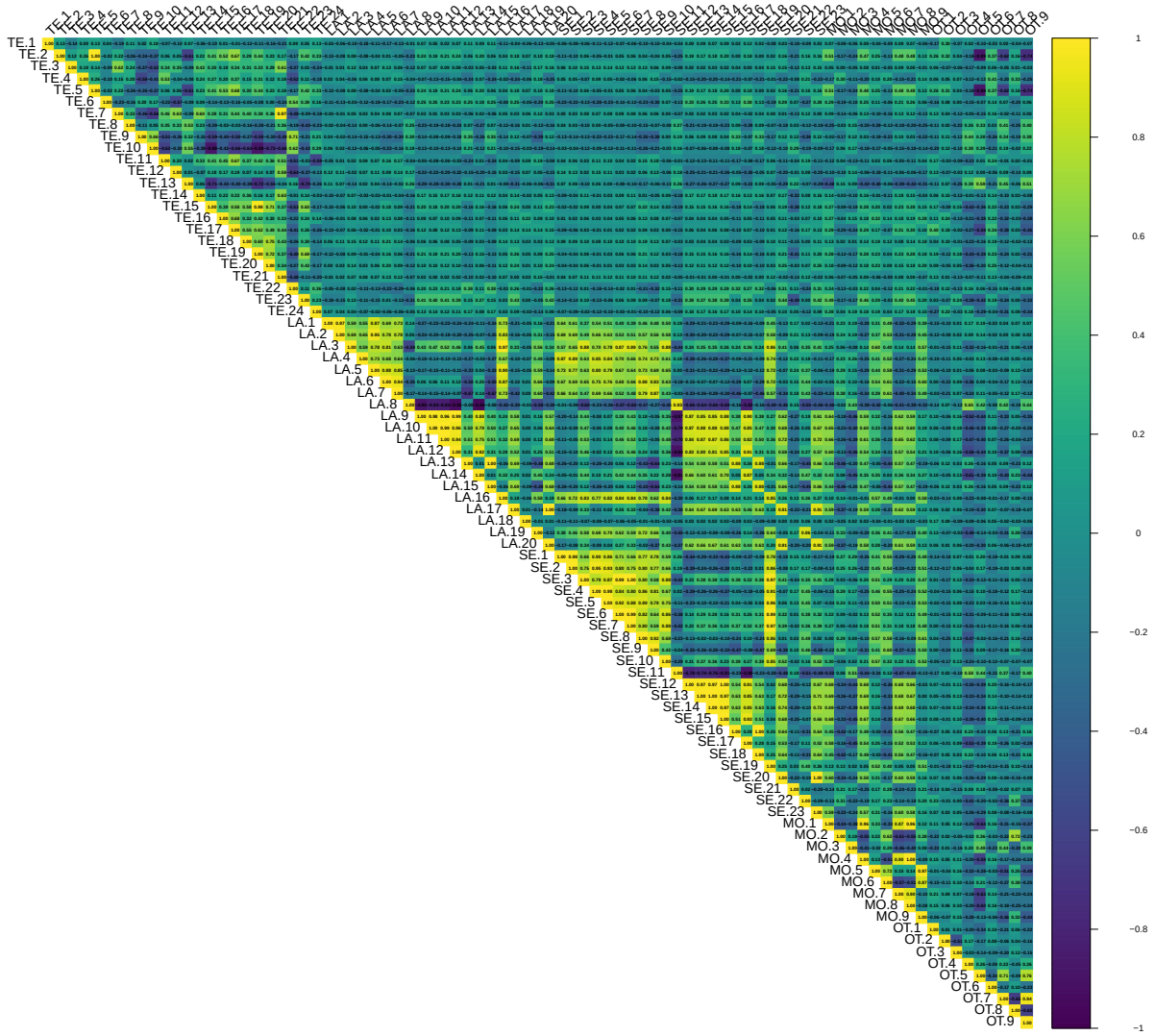

Figure S76: Auto-correlation plot 30_50.

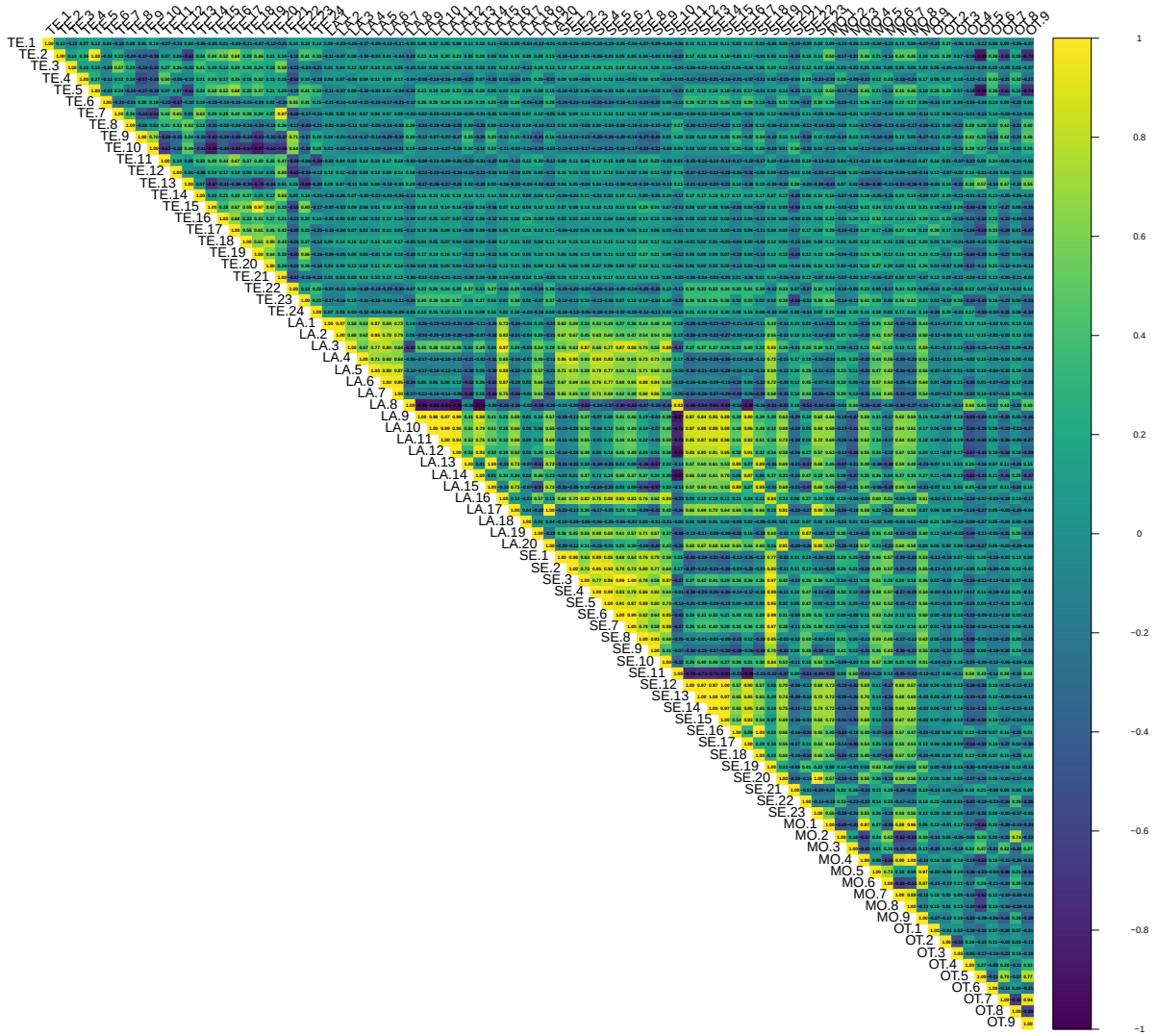

Figure S77: Auto-correlation plot 50_70.

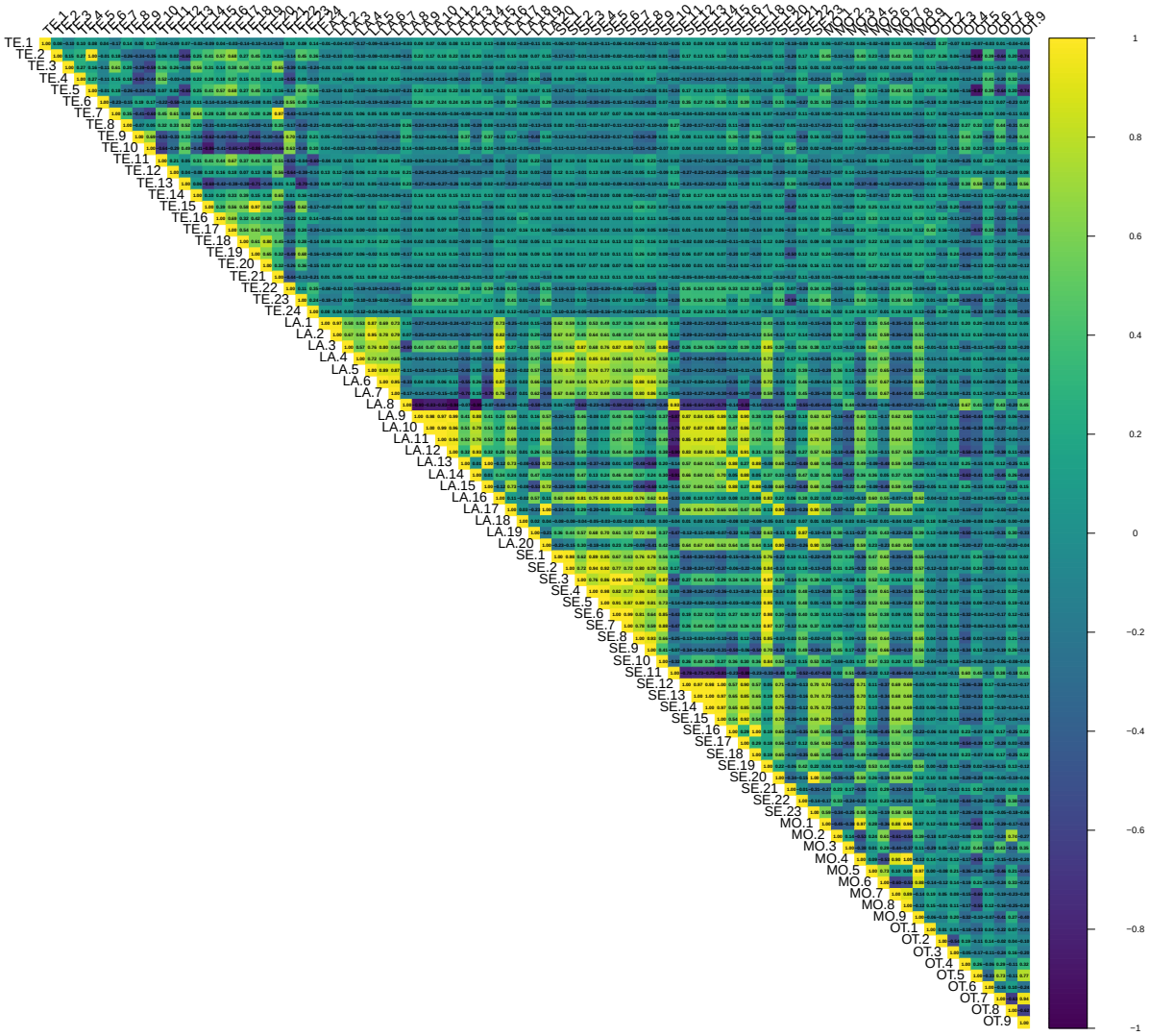

Figure S78: Auto-correlation plot 70_100.

Regarding the correlation plot the Landsat and Sentinel bands are the ones with the higher correlation followed by terrain derivatives from DEM.

ALR transformation

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Boruta and RFE selections

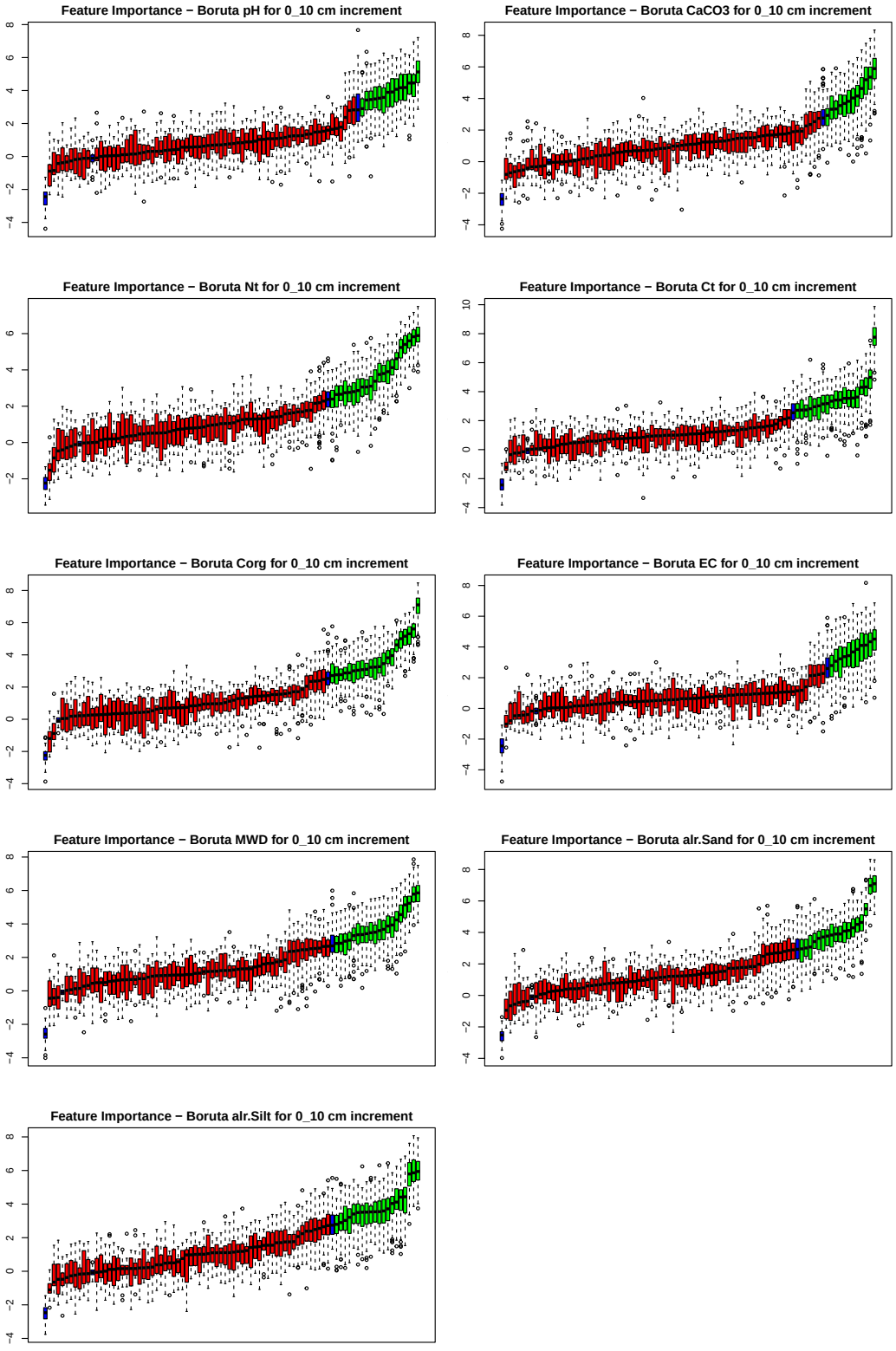

Figure S79: Boruta plot 0_10.

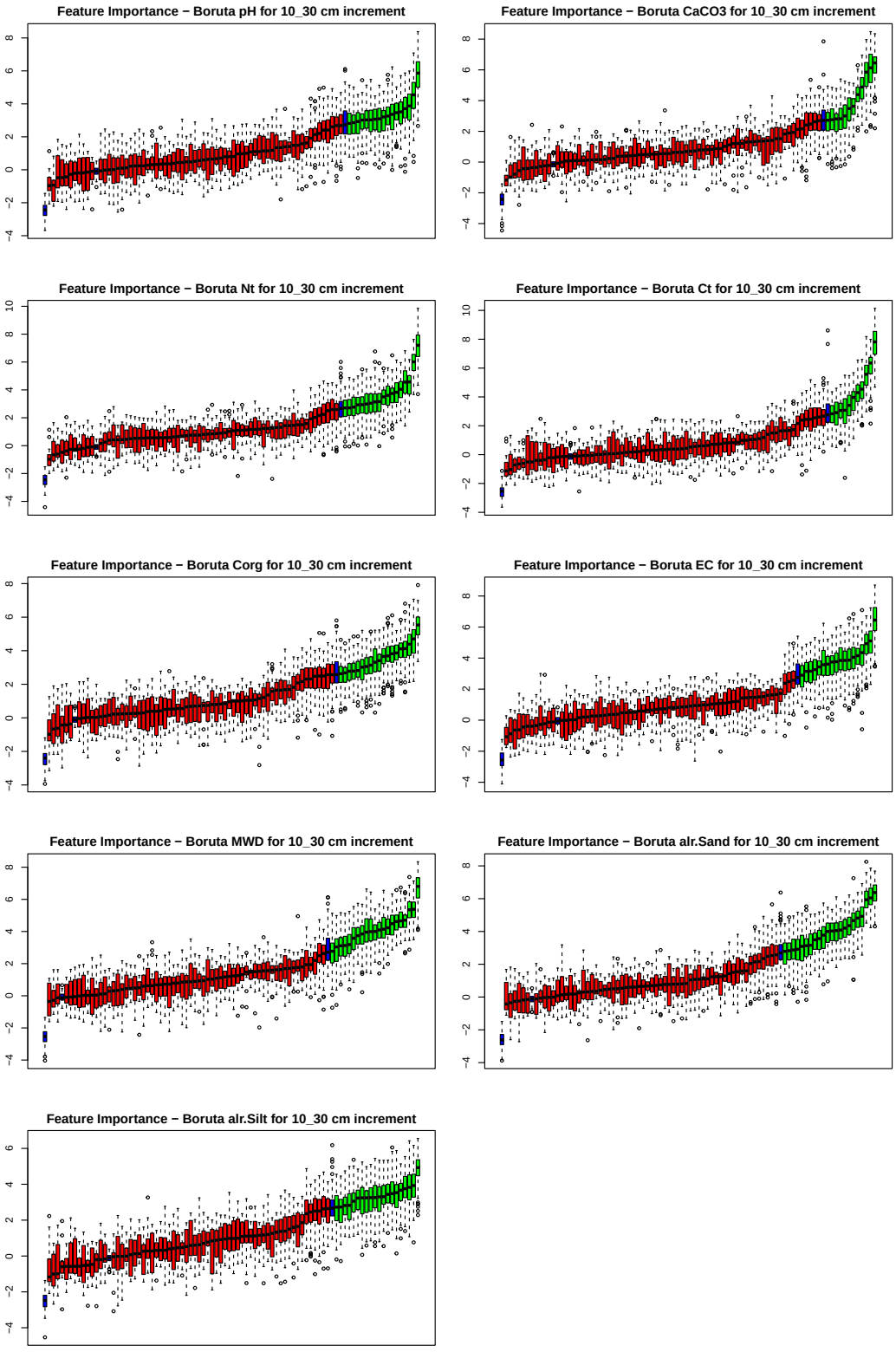

Figure S80: Boruta plot 10_30.

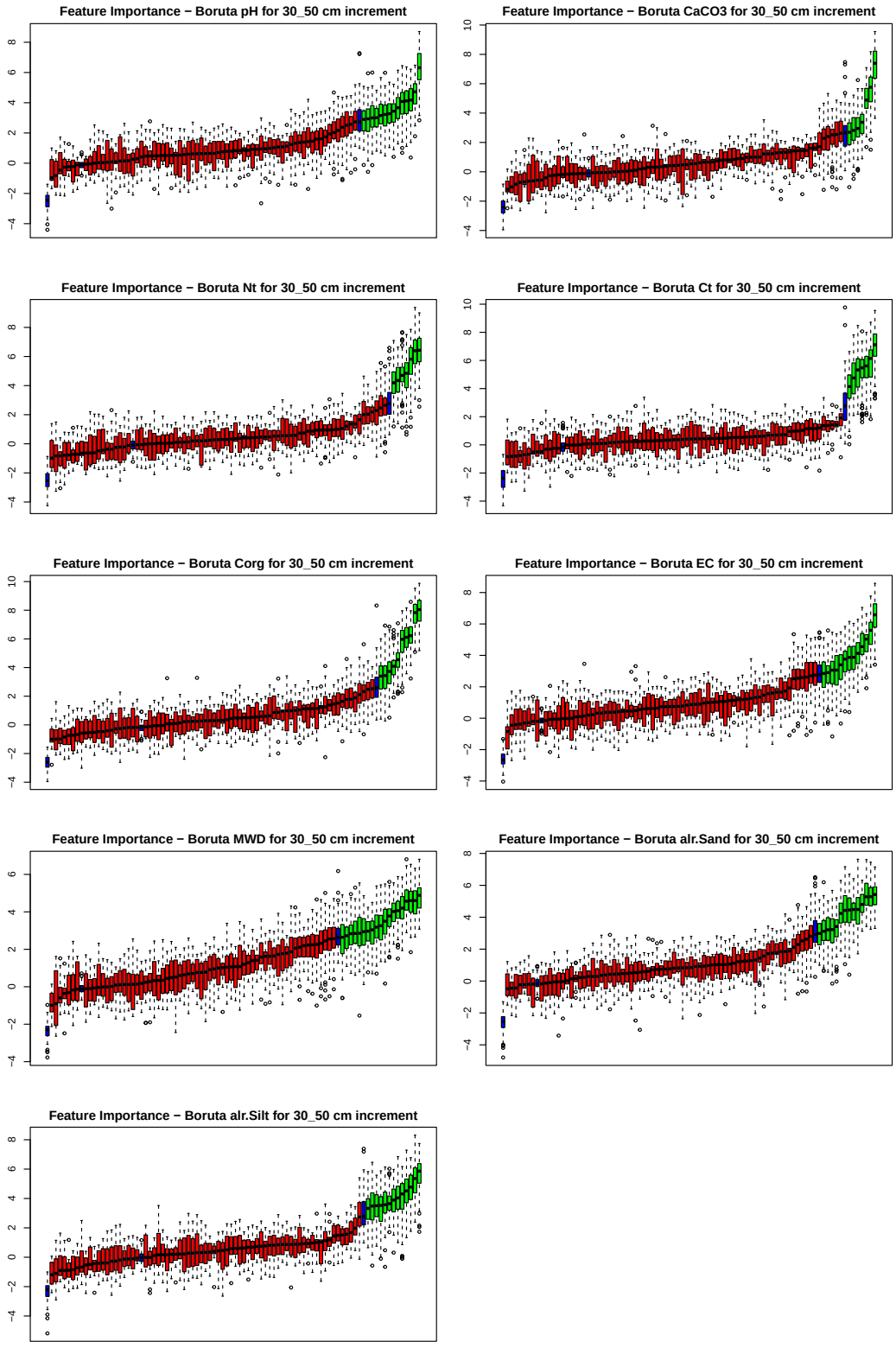

Figure S81: Boruta plot 30_50.

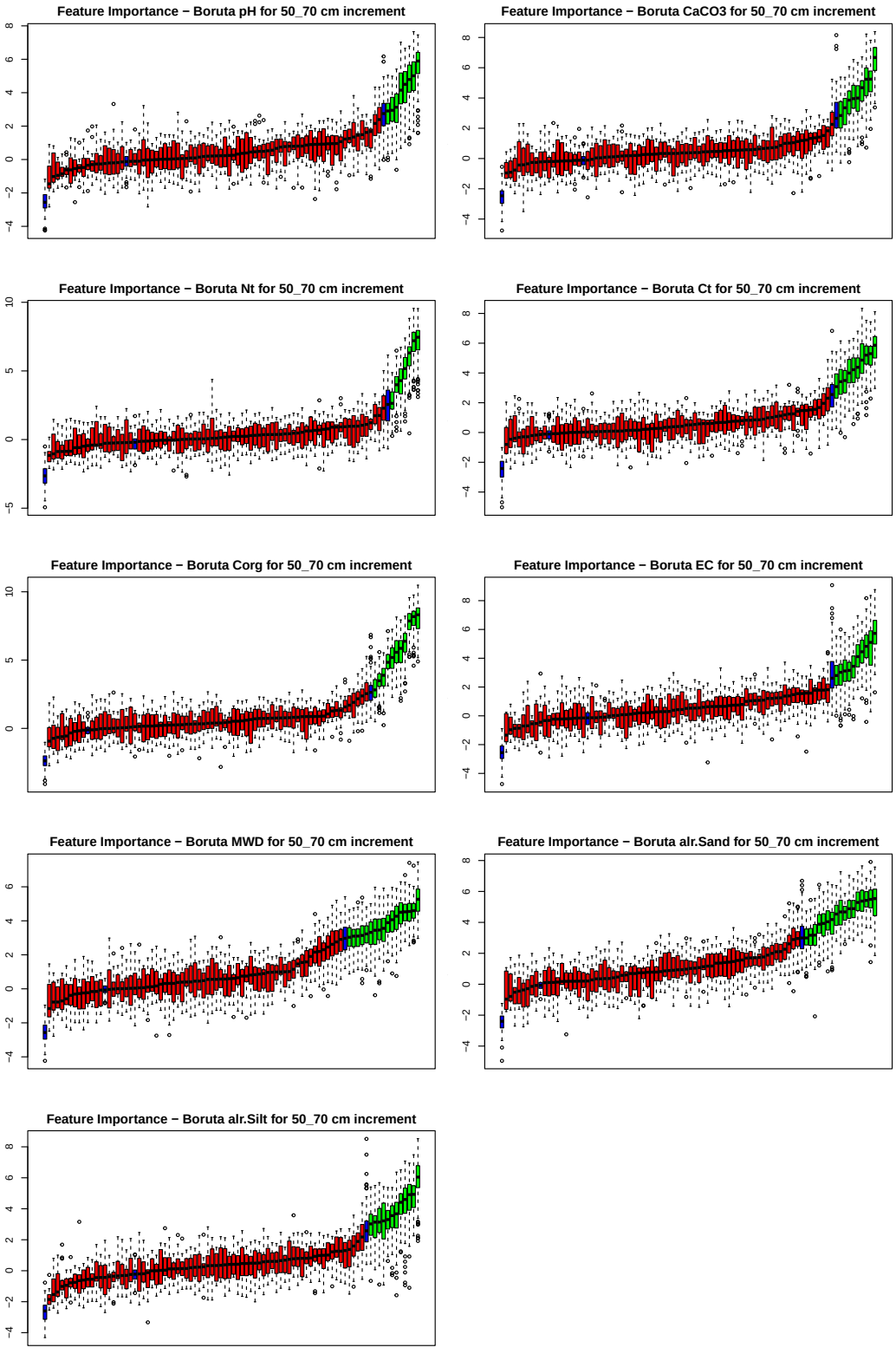

Figure S82: Boruta plot 50_70.

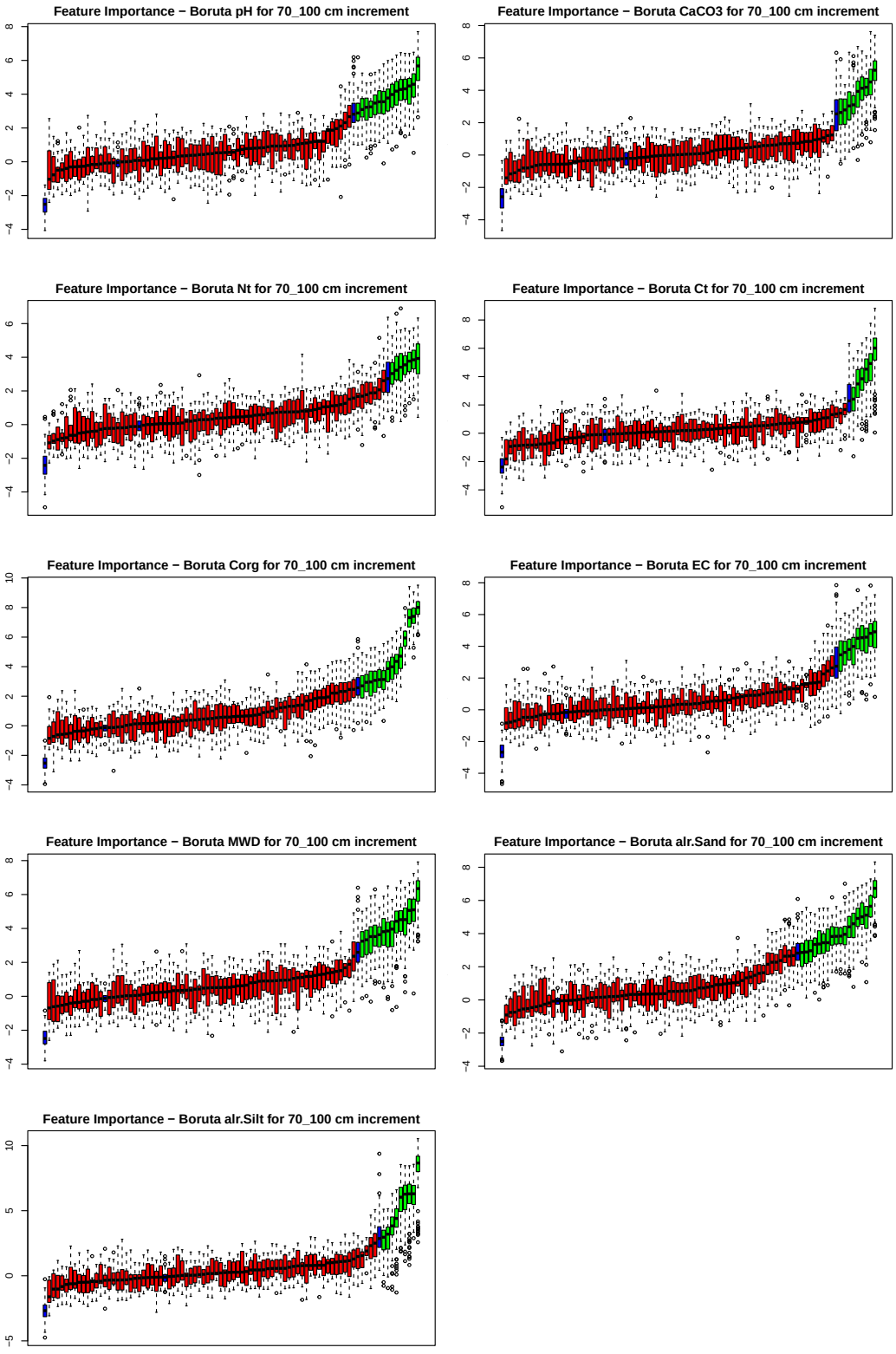

Figure S83: Boruta plot 70_100.

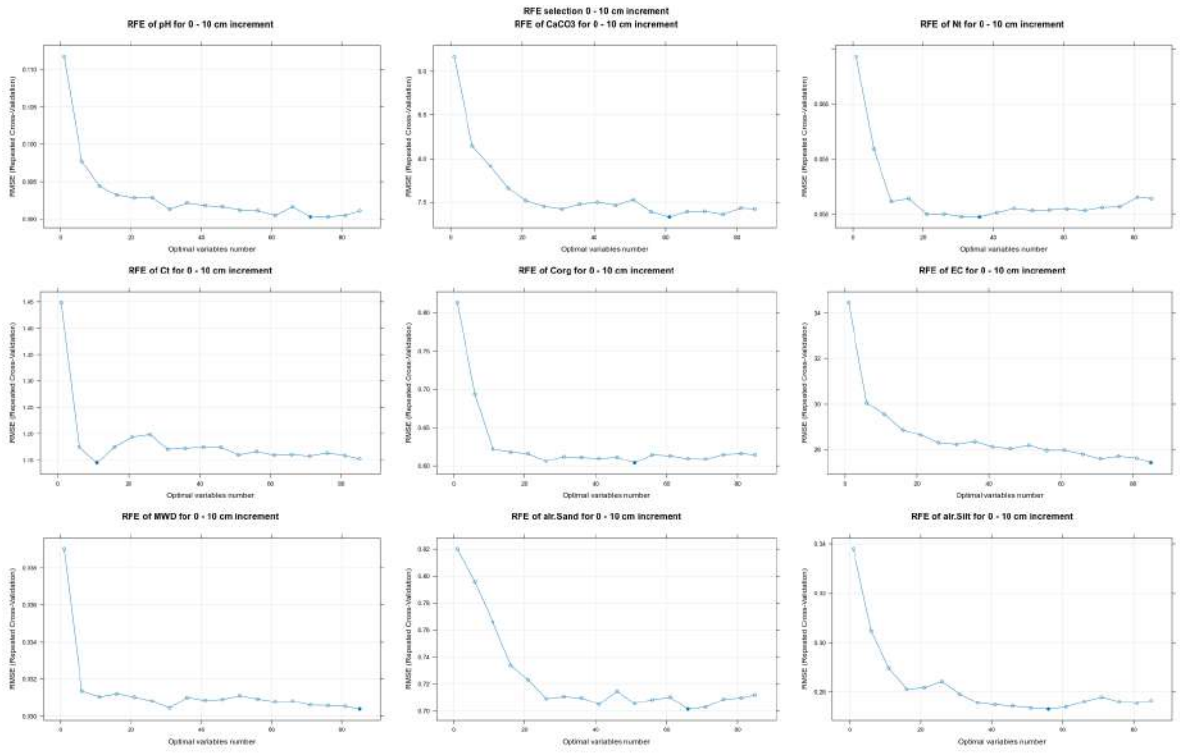

Figure S84: RFE selection for 0 - 10 cm depth.

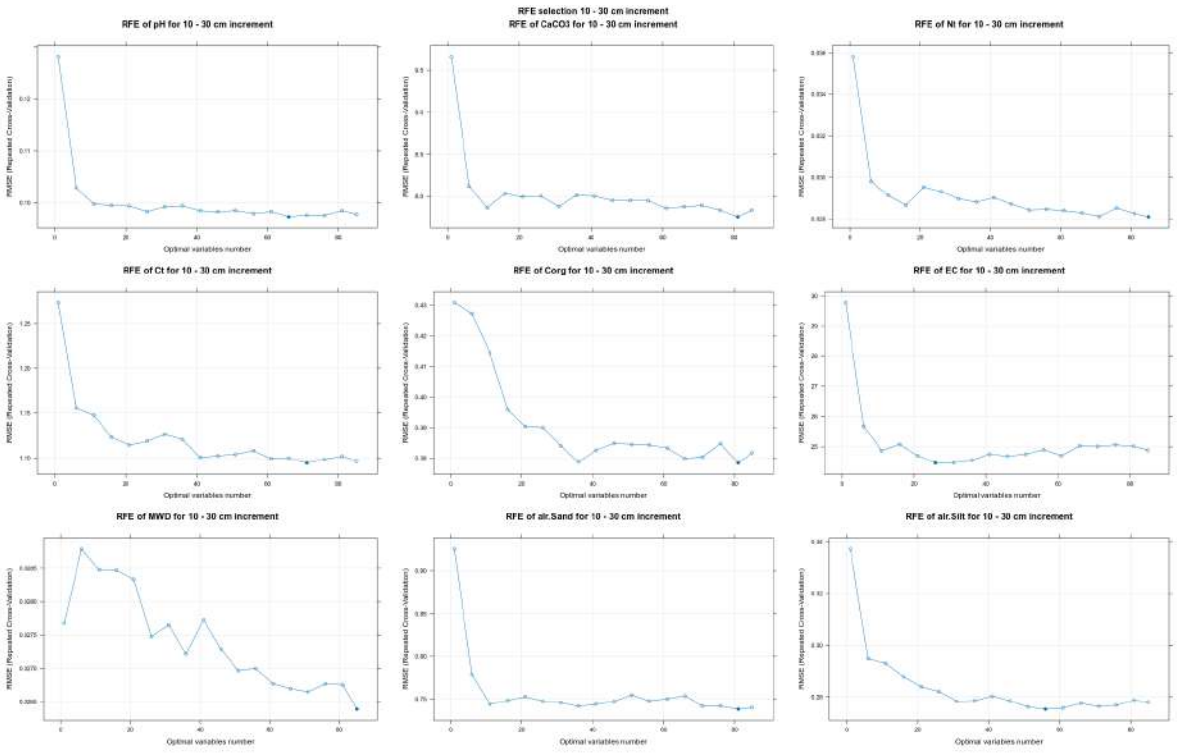

Figure S85: RFE selection for 10 - 30 cm depth.

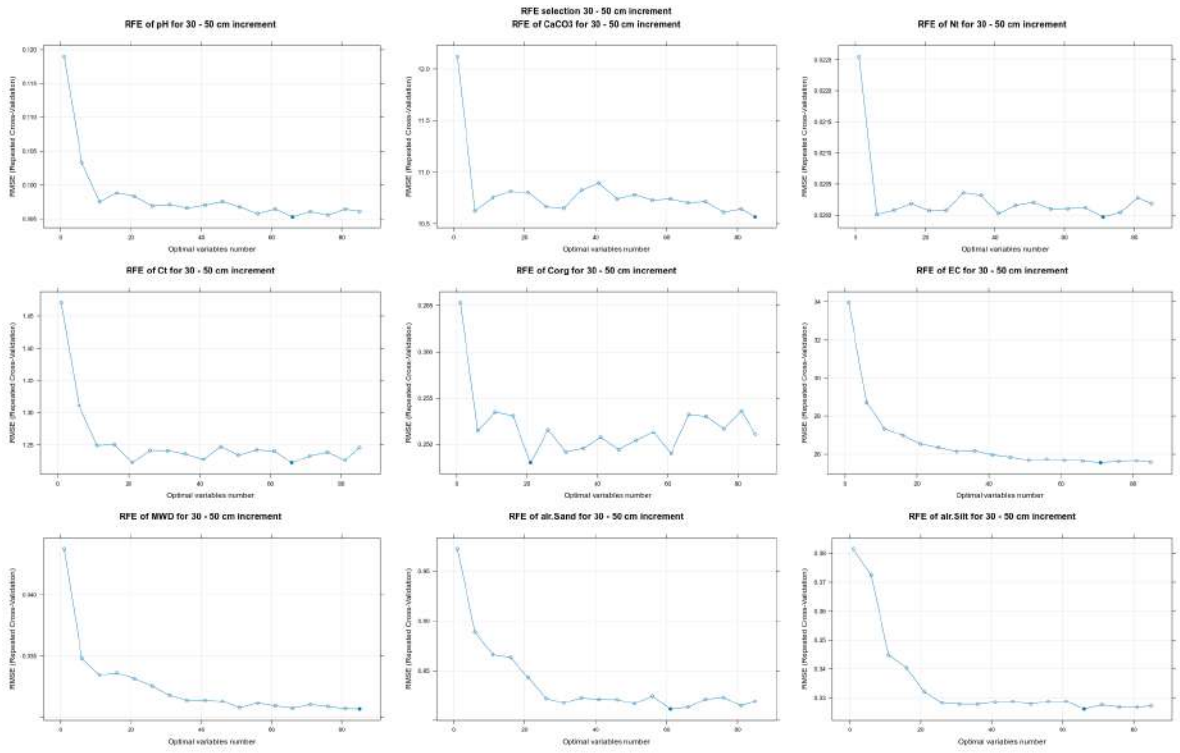

Figure S86: RFE selection for 30 - 50 cm depth.

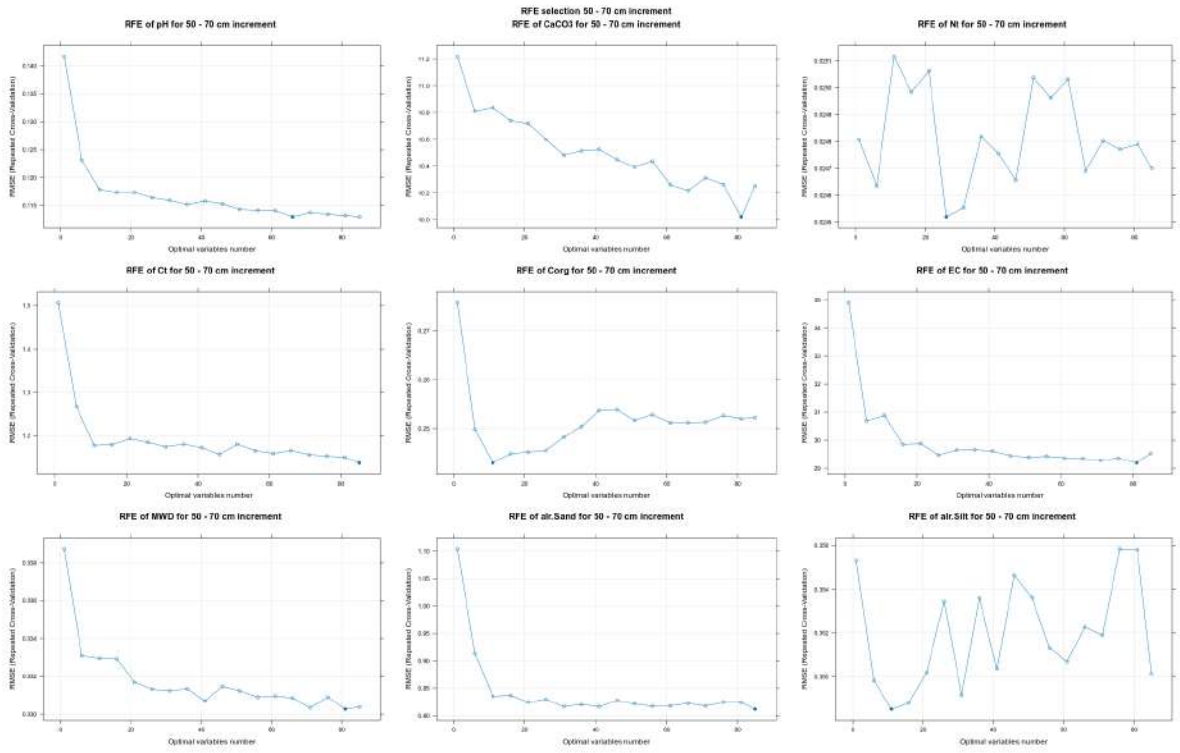

Figure S87: RFE selection for 50 - 70 cm depth.

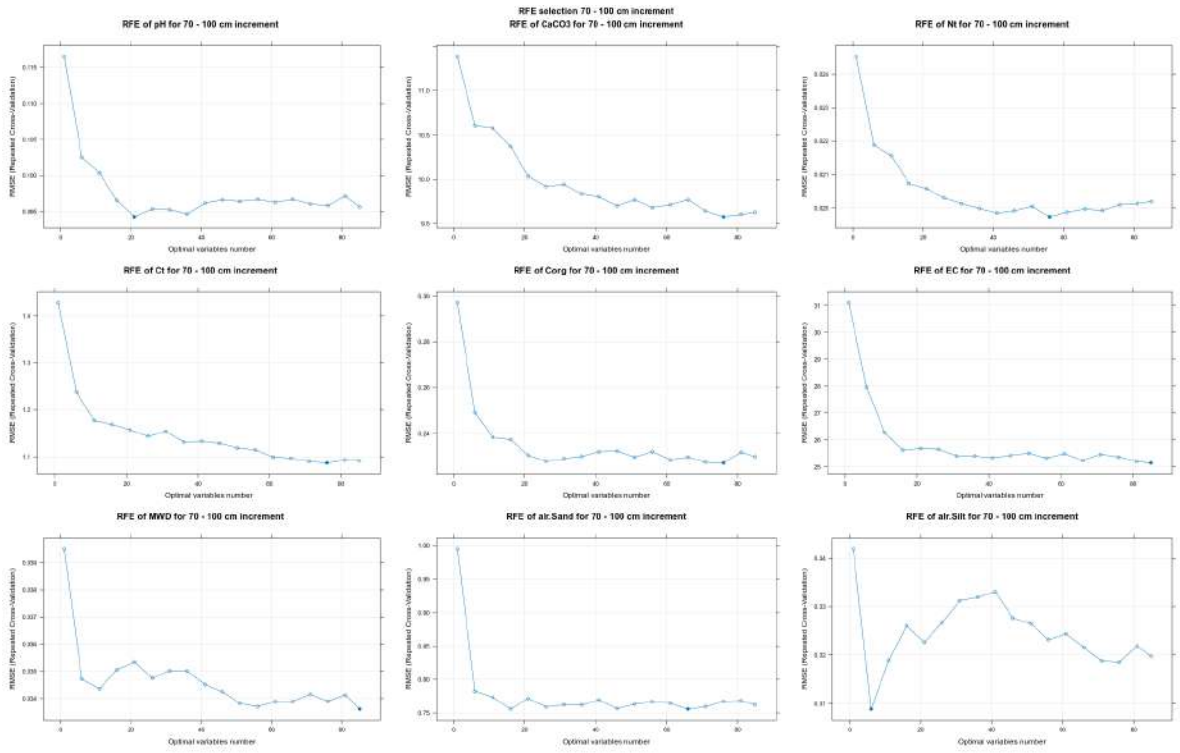

Figure S88: RFE selection for 70 - 100 cm depth.

Models developpement

Preparation of the environment

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Custom QRF model

```
library(caret)
library(quantregForest)

qrf_caret <- list(

  type = "Regression",
  library = "quantregForest",
```

```
  parameters = data.frame(
    parameter = c("mtry", "nodesize"),
    class = c("numeric", "numeric"),
    label = c("mtry", "Node size")
  ),

  grid = function(x, y, len = NULL, search = "grid") {
    expand.grid(
      mtry = unique(pmax(1, floor(seq(1, ncol(x), length.out = len)))),
      nodesize = c(5, 10, 15)
    )
  },

  fit = function(x, y, wts, param, lev, last, classProbs, ...) {
    quantregForest(
      x = x,
      y = y,
      mtry = param$mtry,
      nodesize = param$nodesize,
      keep.inbag = TRUE,
      ...
    )
  },

  predict = function(modelFit, newdata, submodels = NULL) {
    predict(modelFit, newdata, what = c(0.5, 0.05, 0.95))
  },
  prob = NULL,

  sort = function(x) x[order(x$mtry), ],

  levels = function(x) NULL
)
```

Run the models

Boruta models

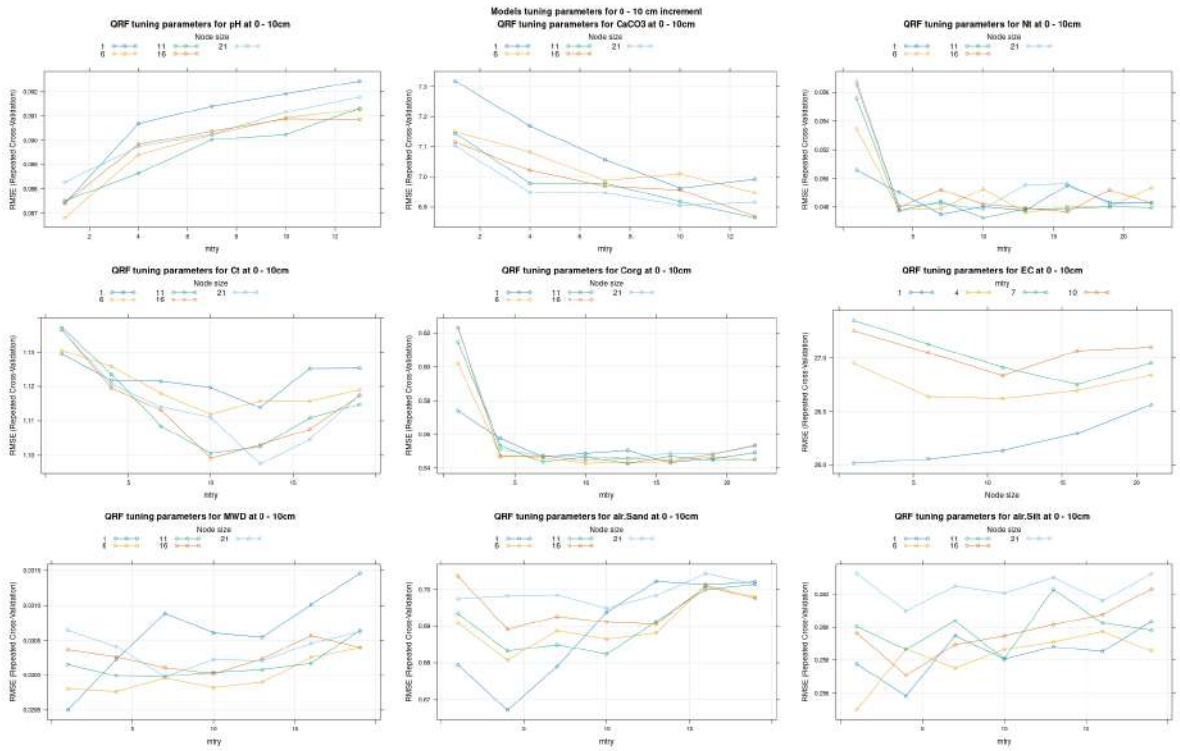

Figure S89: Hyperparameters for 0 - 10 cm.

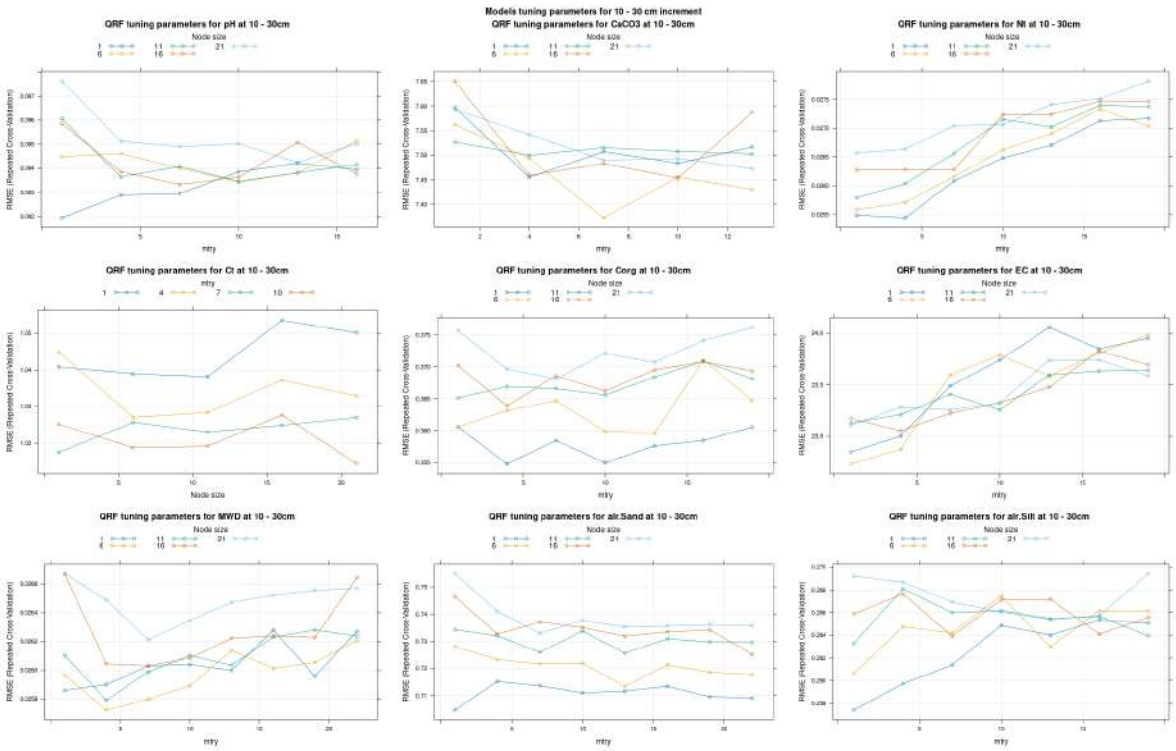

Figure S90: Hyperparameters for 10 - 30 cm.

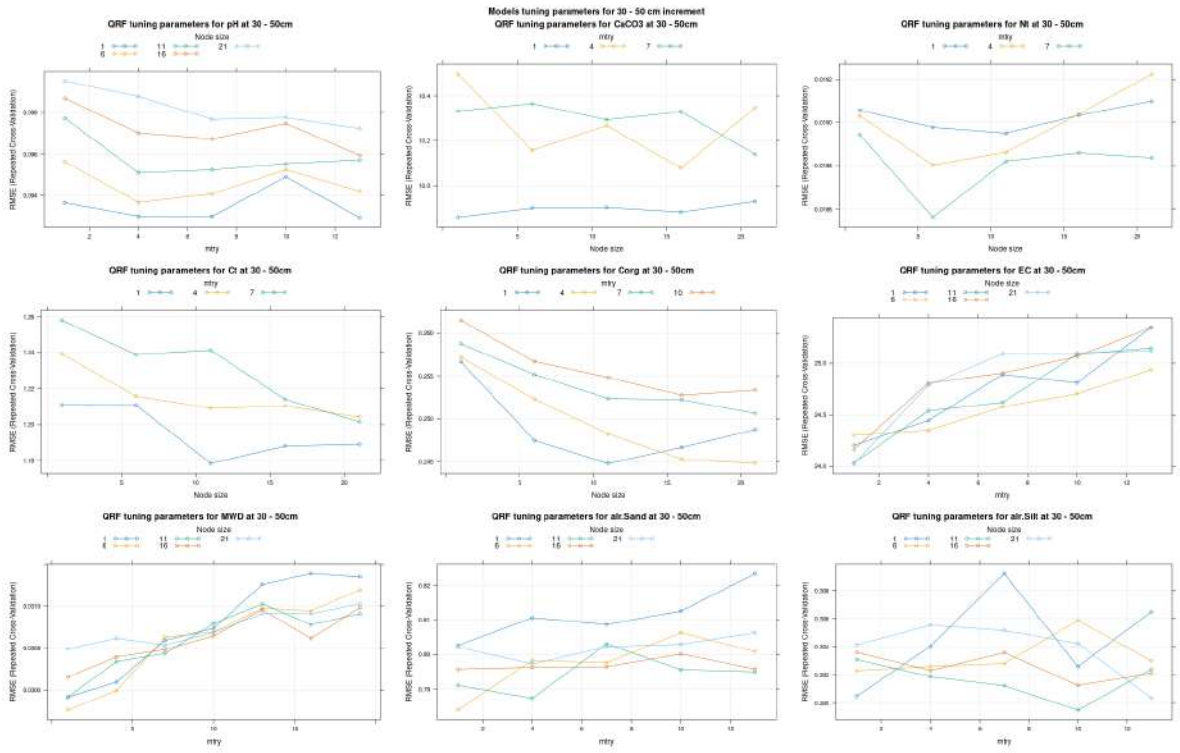

Figure S91: Hyperparameters for 30 - 50 cm.

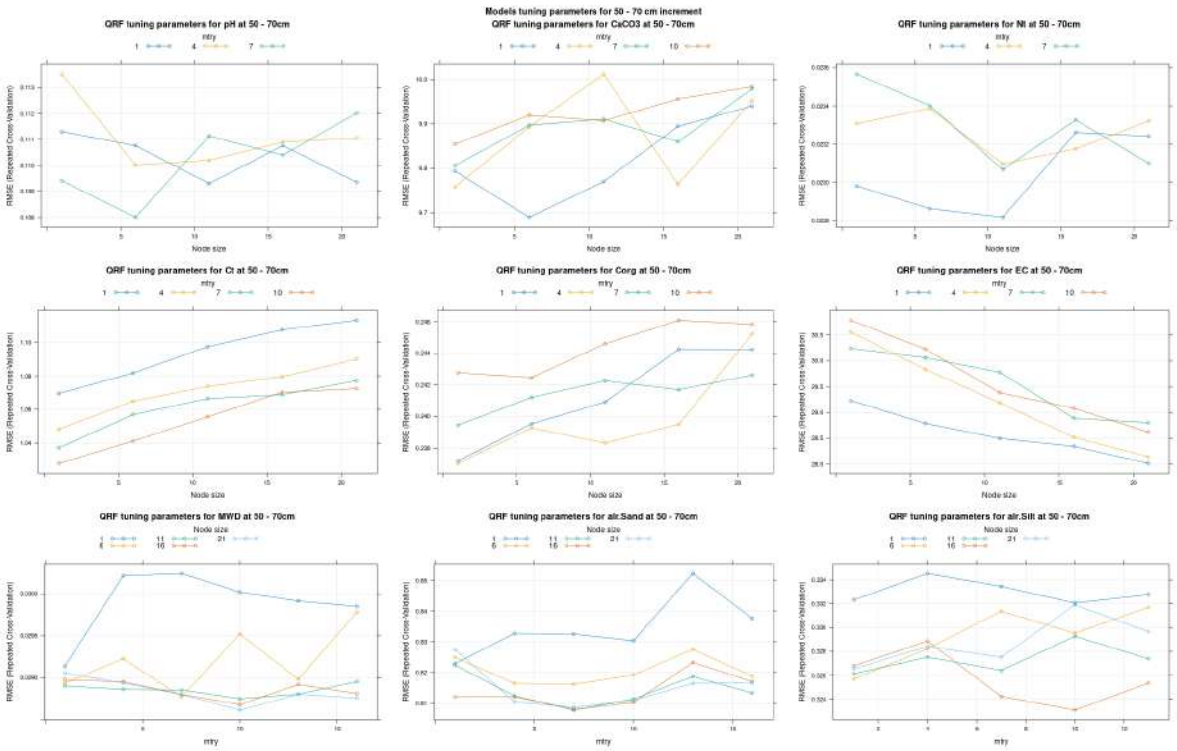

Figure S92: Hyperparameters for 50 - 70 cm.

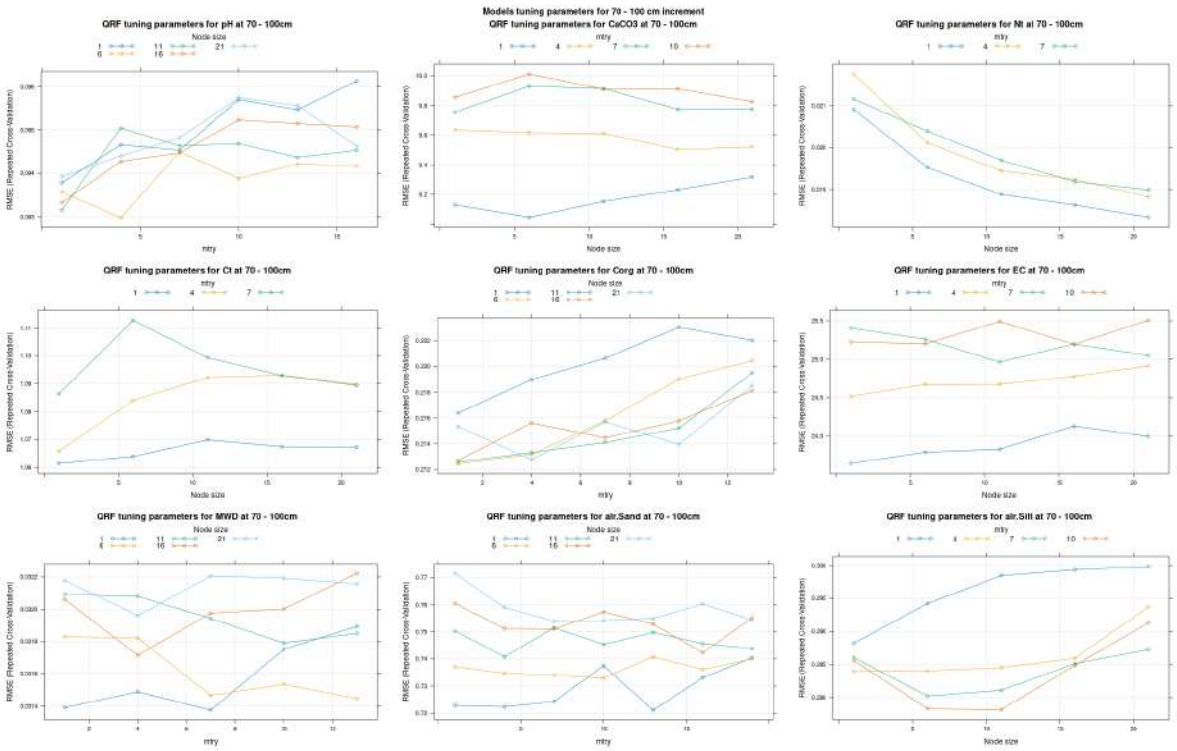

Figure S93: Hyperparameters for 70 - 100 cm.

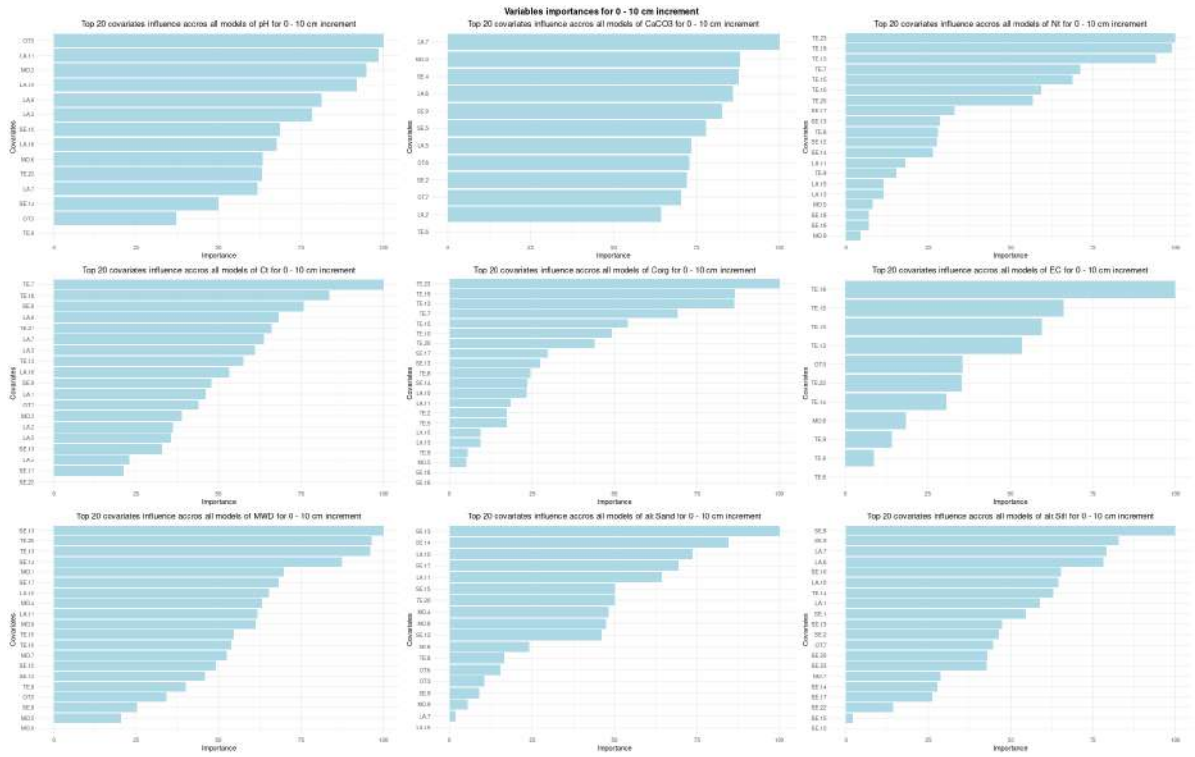

Figure S94: Variables importances for 0 - 10 cm.

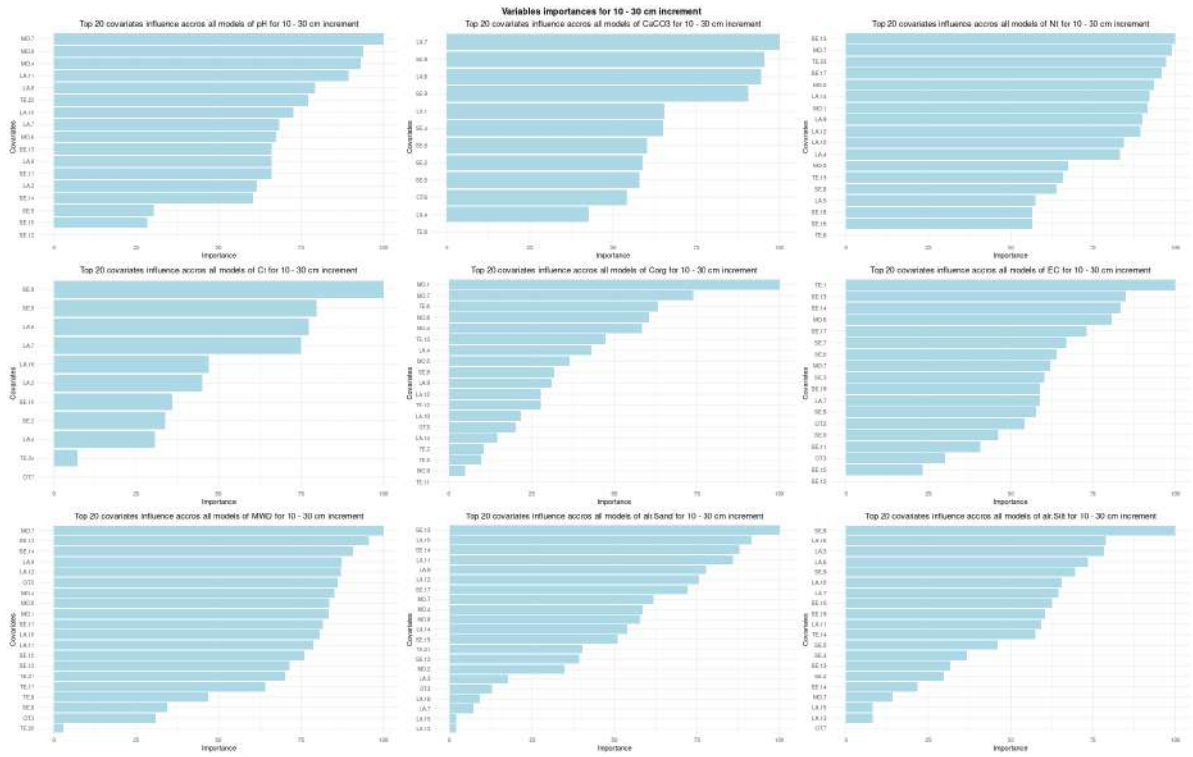

Figure S95: Variables importances for 10 - 30 cm.

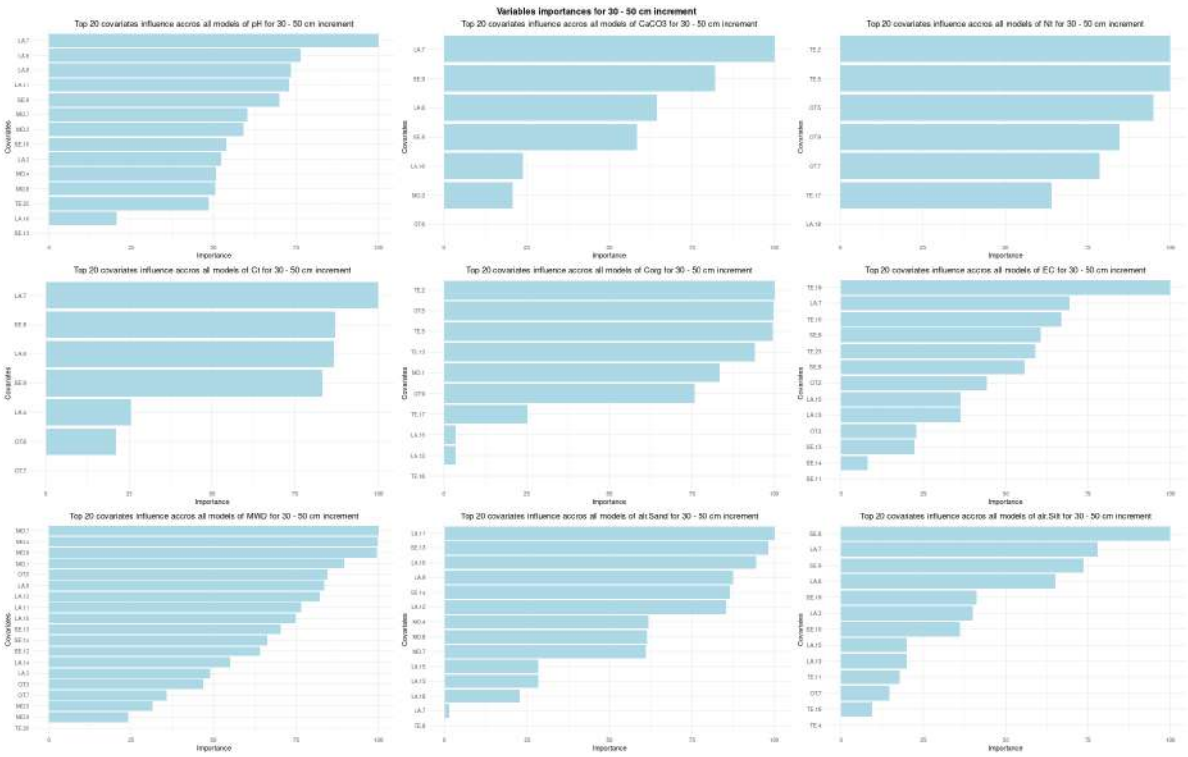

Figure S96: Variables importances for 30 - 50 cm.

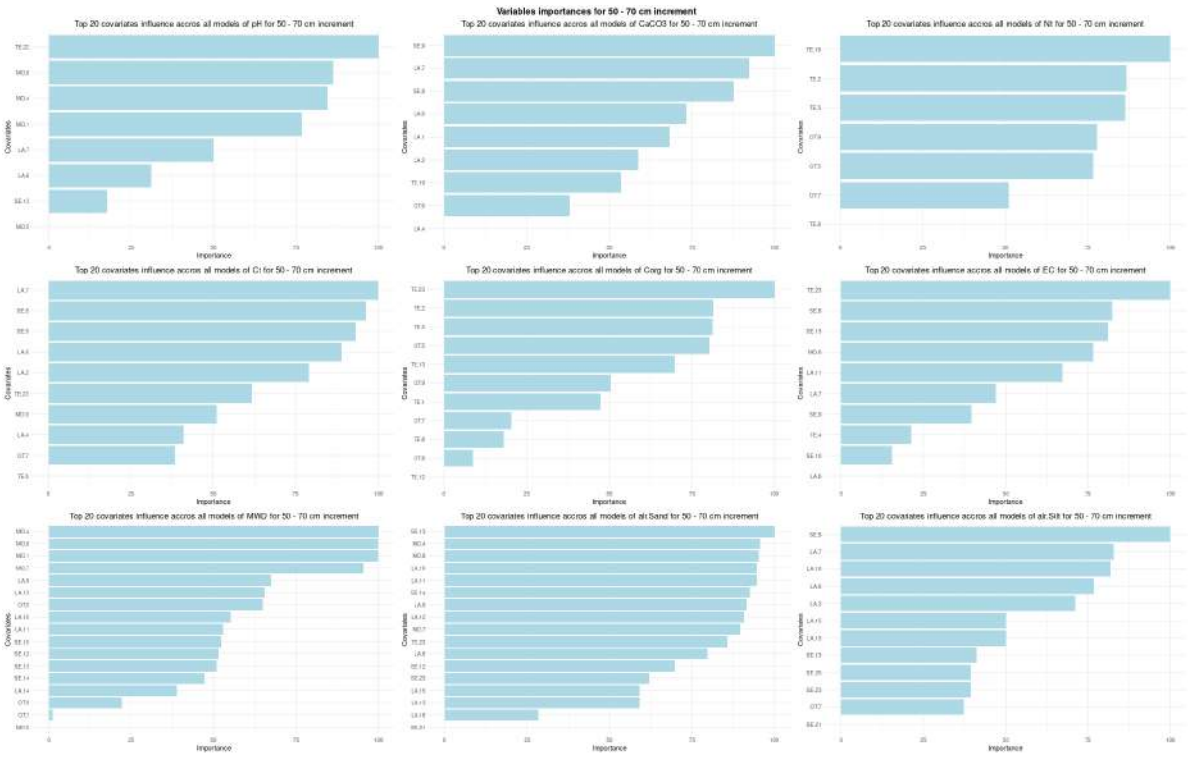

Figure S97: Variables importances for 50 - 70 cm.

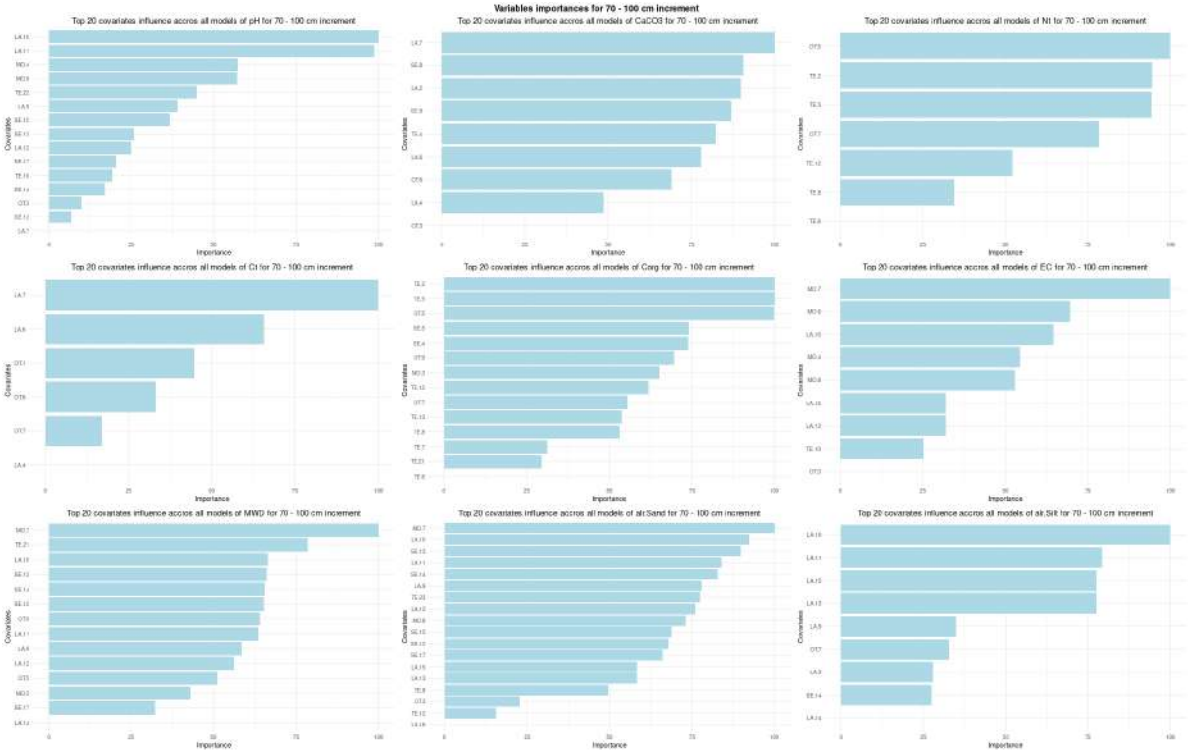

Figure S98: Variables importances for 70 - 100 cm.

RFE and Boruta split models

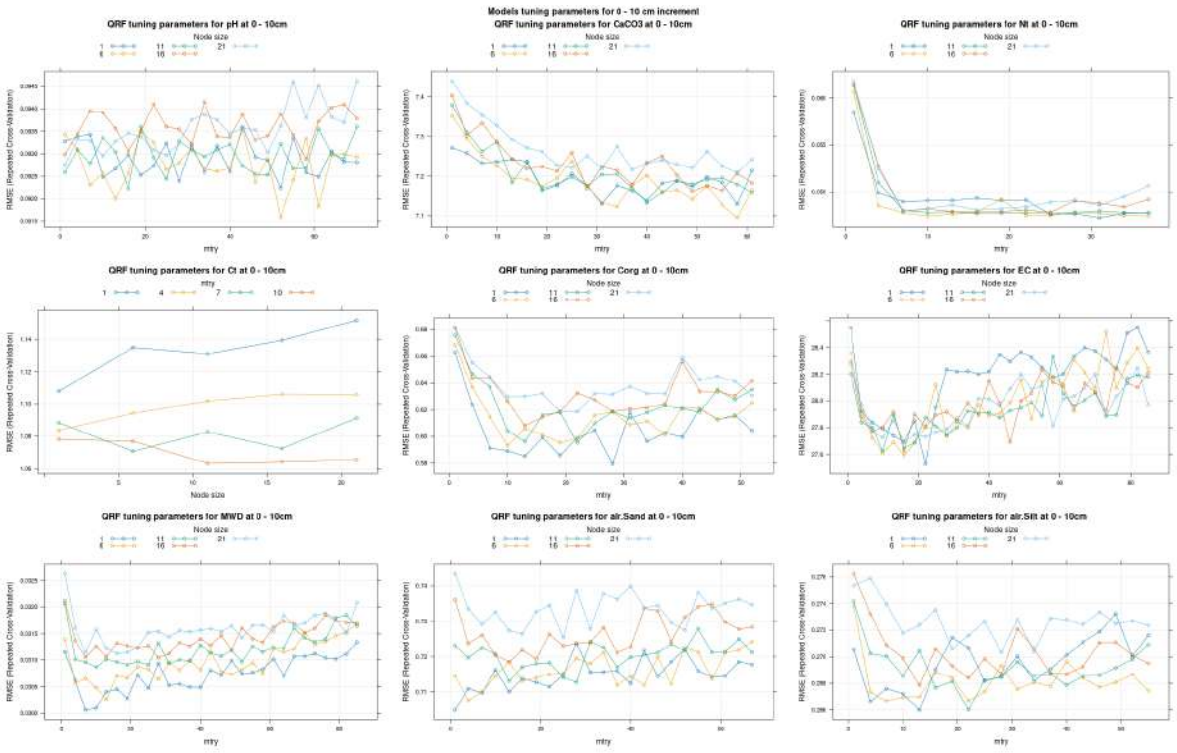

Figure S99: RFE hyperparameters for 0 - 10 cm.

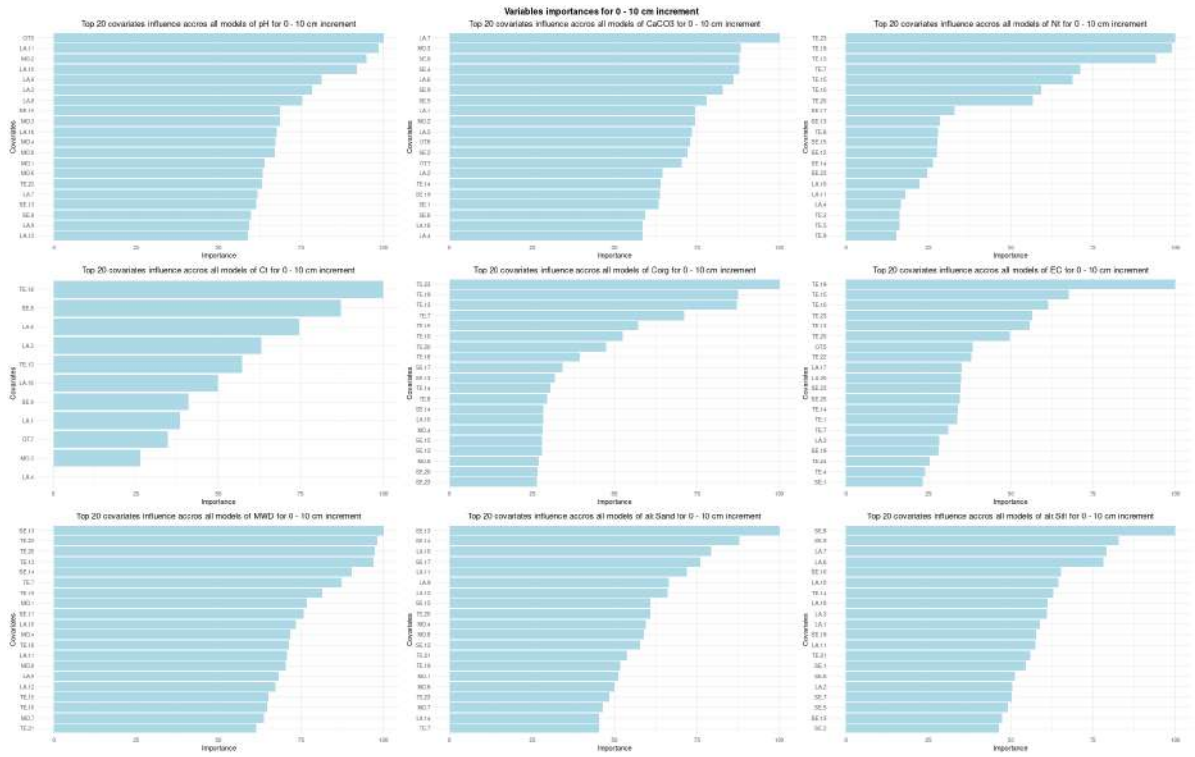

Figure S100: RFE variables importances for 0 - 10 cm.

The RFE selection was only performed on the 0 - 10 cm interval.

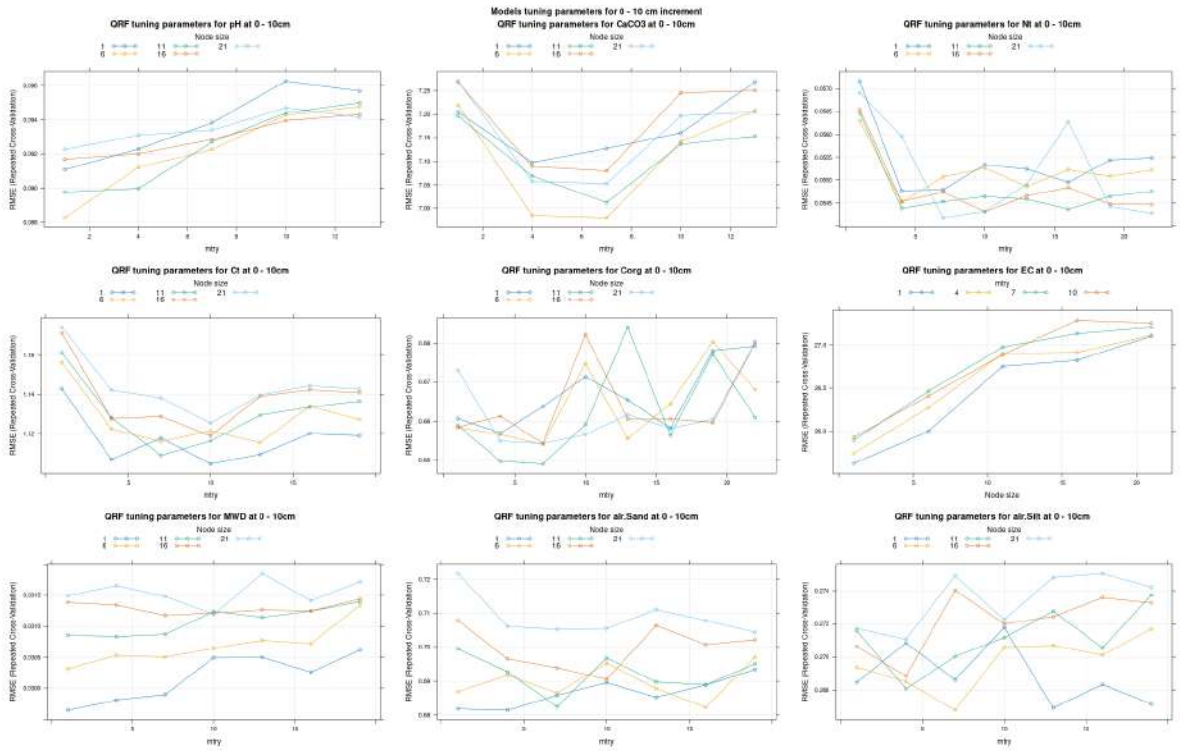

Figure S101: Split Boruta hyperparameters for 0 - 10 cm.

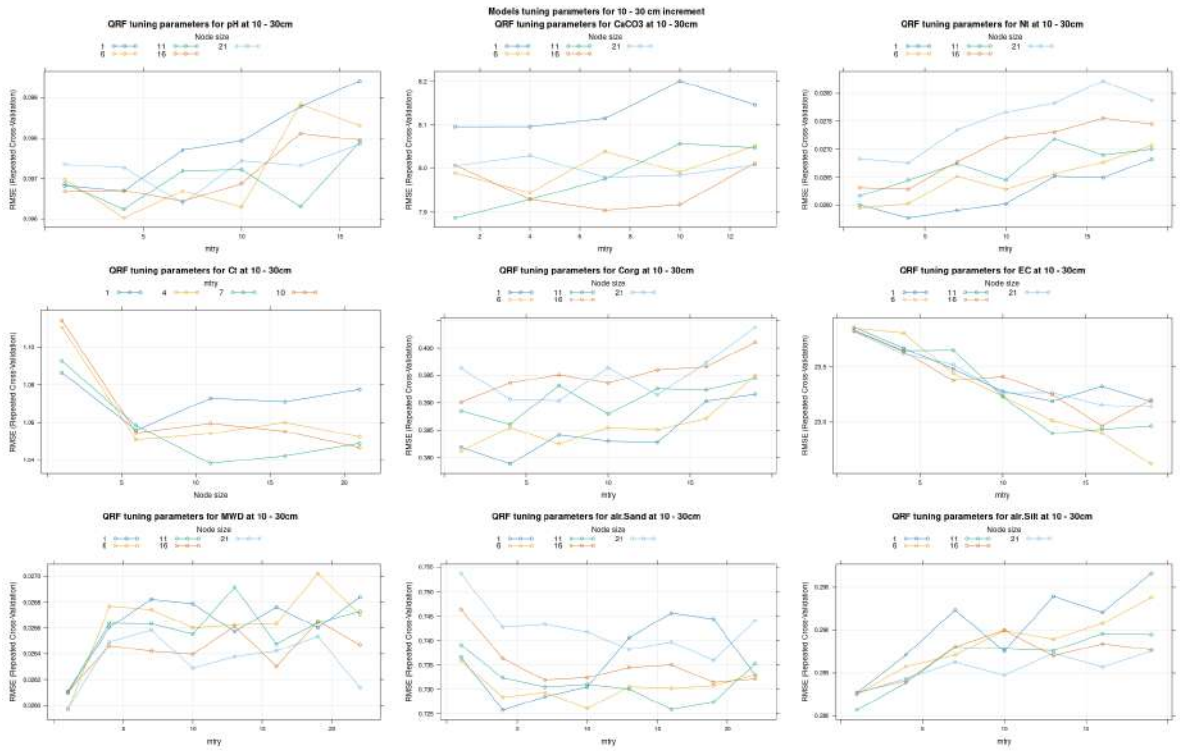

Figure S102: Split Boruta hyperparameters for 10 - 30 cm.

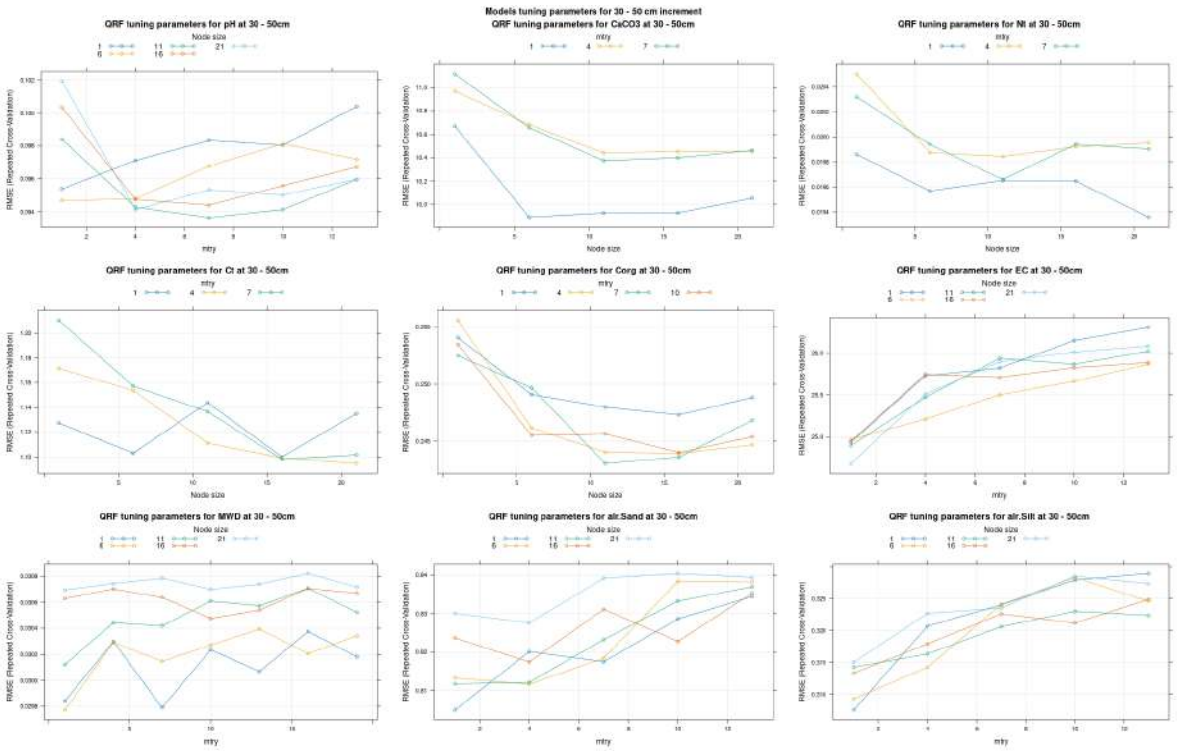

Figure S103: Split Boruta hyperparameters for 30 - 50 cm.

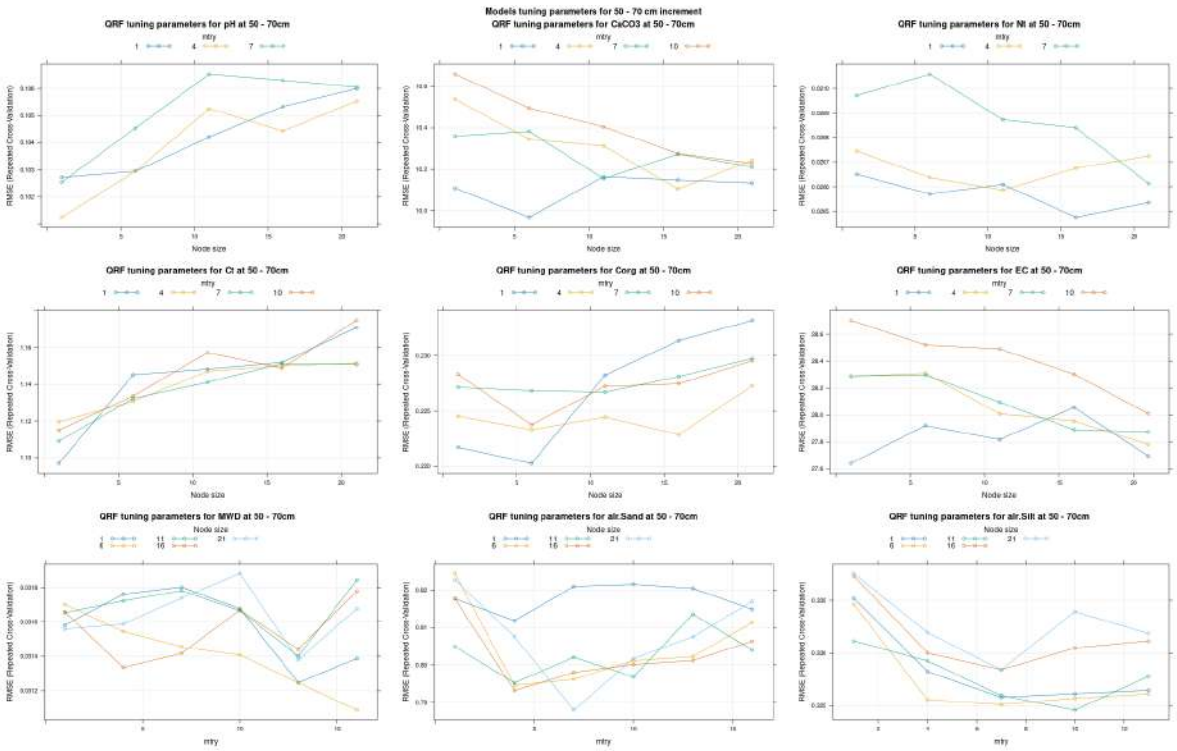

Figure S104: Split Boruta hyperparameters for 50 - 70 cm.

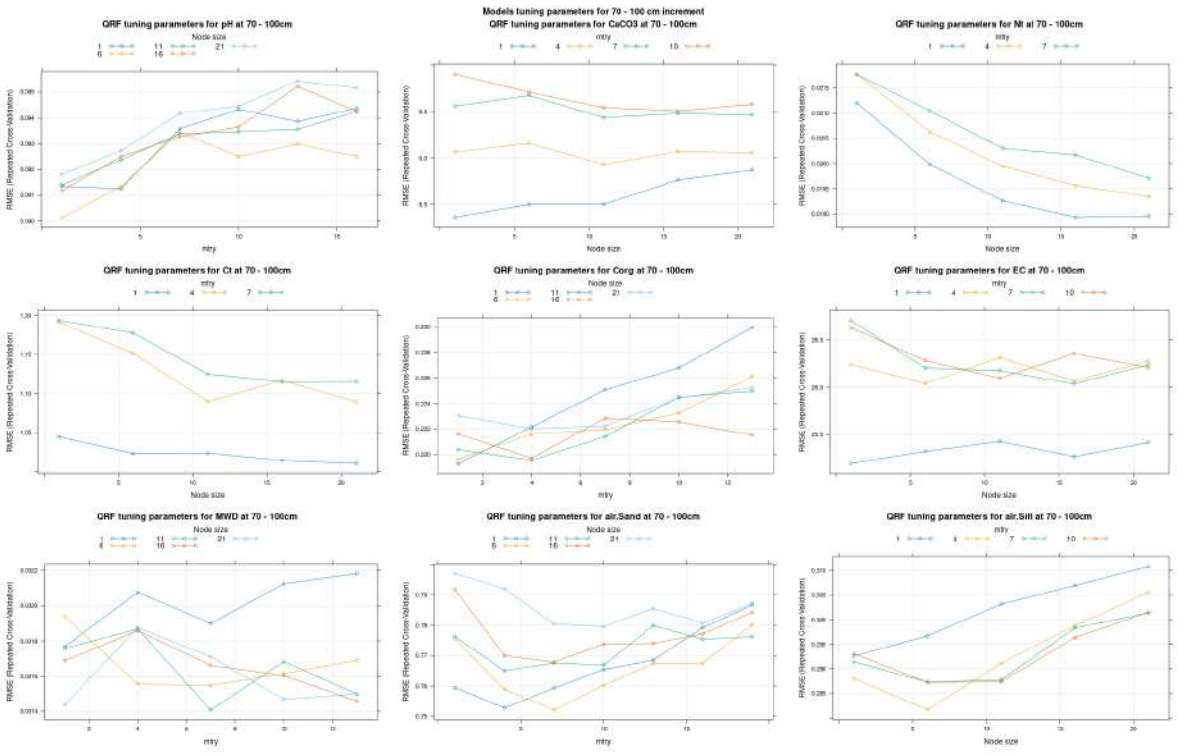

Figure S105: Split Boruta hyperparameters for 70 - 100 cm.

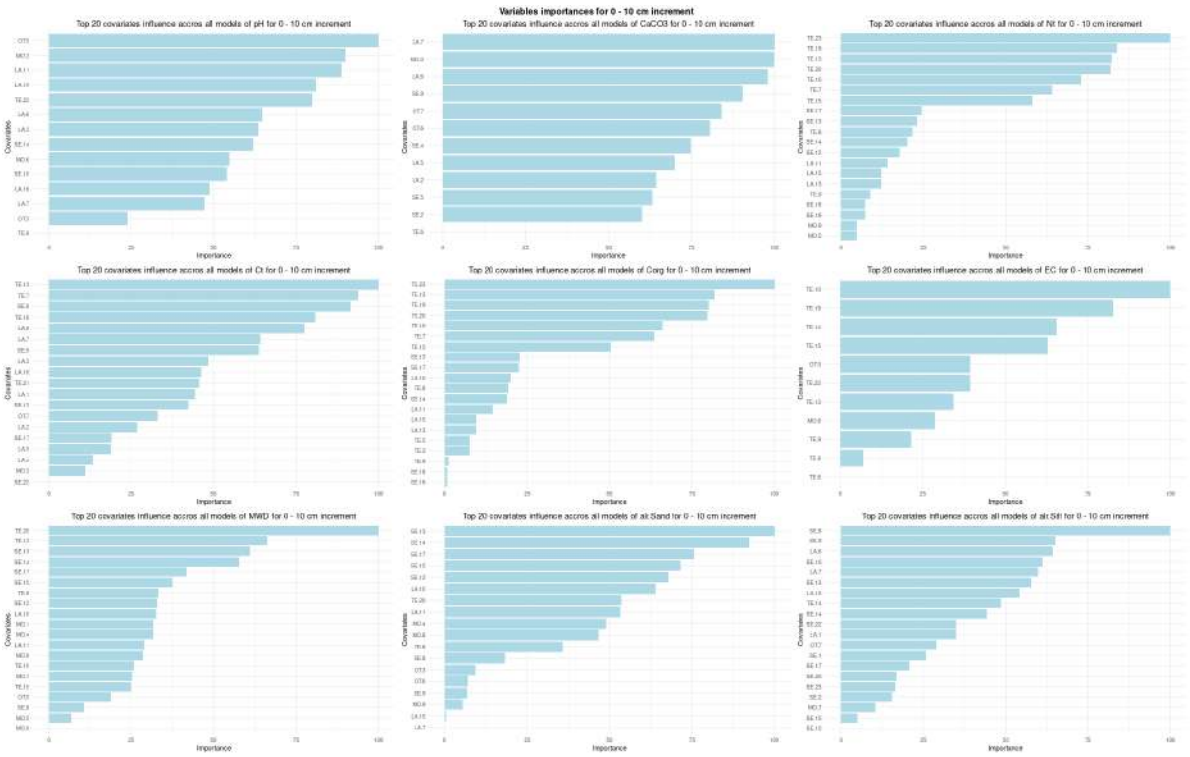

Figure S106: Split variables importances for 0 - 10 cm.

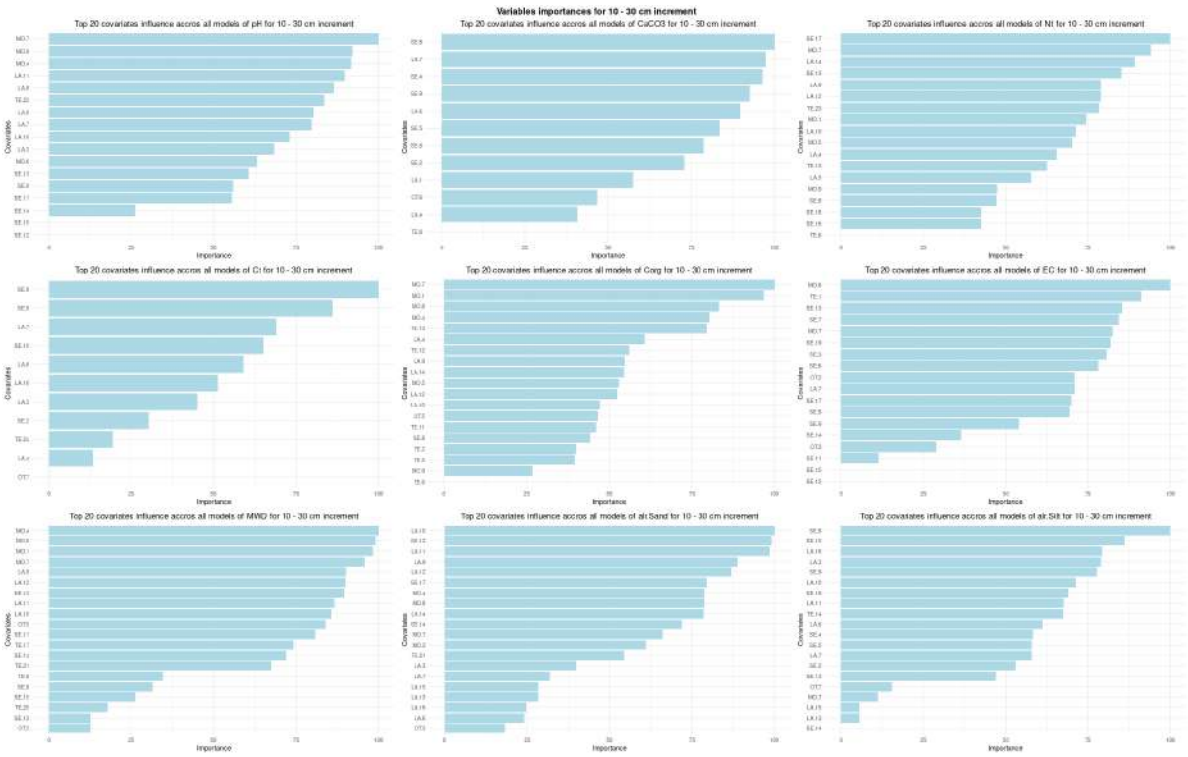

Figure S107: Split variables importances for 10 - 30 cm.

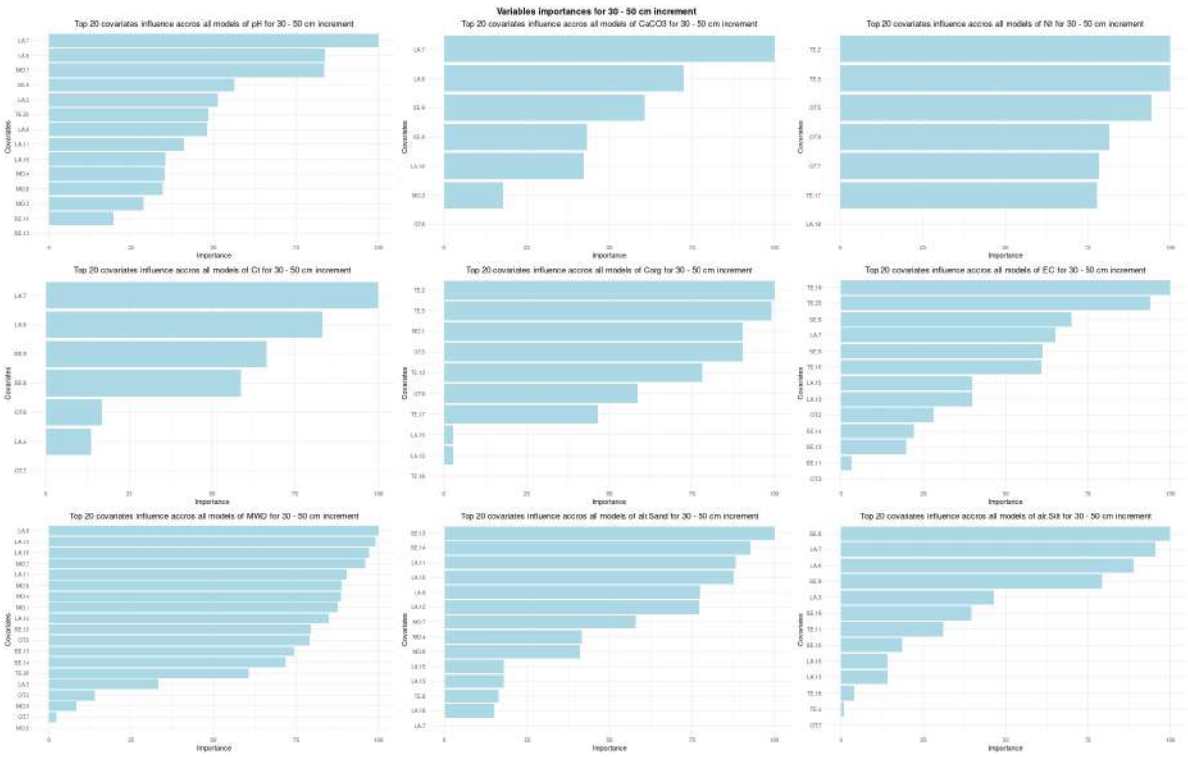

Figure S108: Split variables importances for 30 - 50 cm.

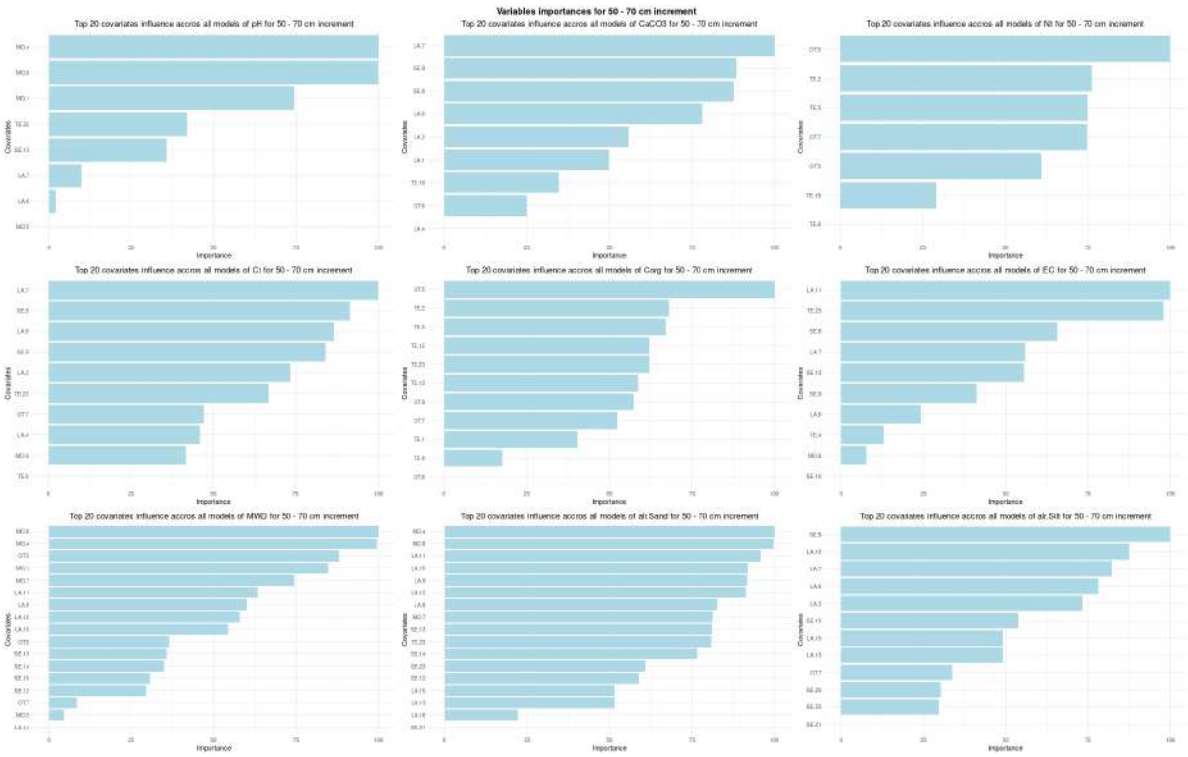

Figure S109: Split variables importances for 50 - 70 cm.

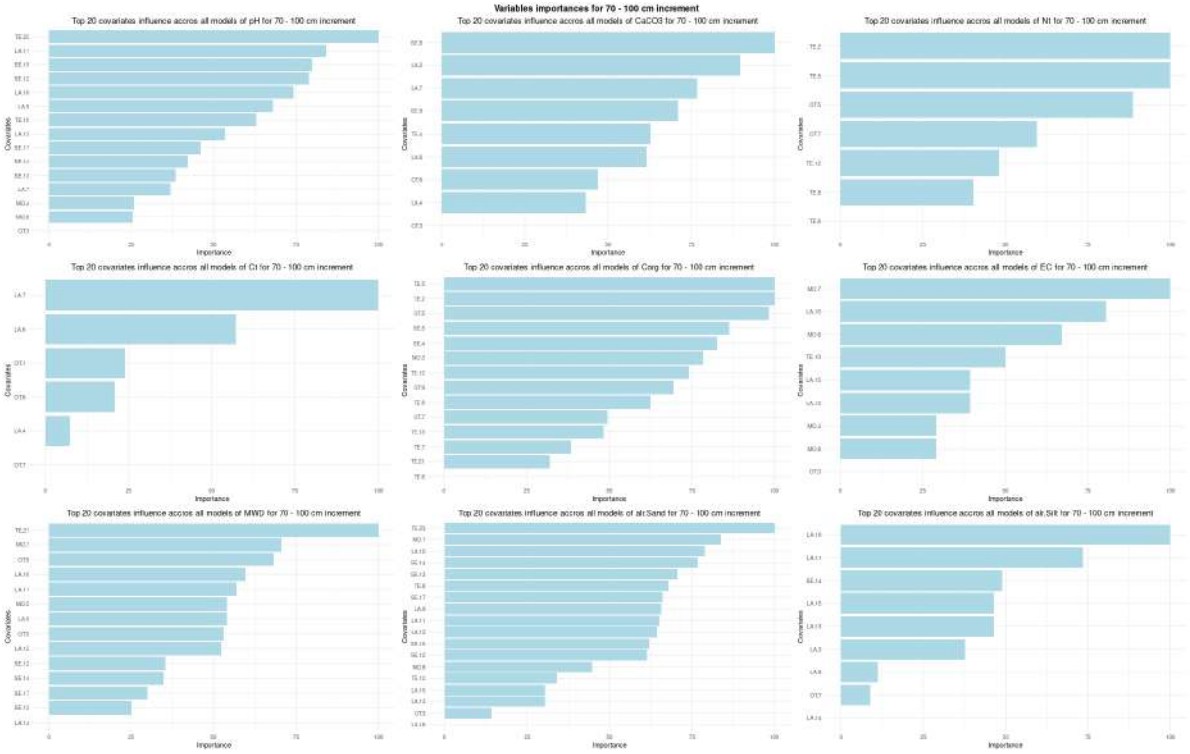

Figure S110: Split variables importances for 70 - 100 cm.

Model evaluation

We used four metrics to evaluate the model:

- ME = Bias with the mean error.
- $RMSE$ = Root mean square error.
- R^2 = Coefficient of determination, also called rsquared.
- $PICP$ = Prediction interval coverage probability set at 90% interval

$$ME = \frac{1}{n}\sum_{i=n}^{n}(\hat{y}_i - y_i)$$

Where: n is the number of observations, y_i the observed value for i, and $\hat{y}_i$ the predicted value for i.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}}$$

Where: n is the number of observations, y_i the observed value for i, and $\hat{y}_i$ the predicted value for i.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where: y_i the observed value for i, $\hat{y}_i$ the predicted value for i, and $\bar{y}$ is the mean of observed values.

$$PICP = \frac{1}{v}\ count\ jj = PL_j^L \leq t_j \leq PL_j^U$$

Where: v is the number of observations, obs_i is the observed value for i, PL_i^L is the lower prediction limit for i, and PL_i^U is the upper prediction limit for i.

Preparation of the environment

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Insert table here

Insert table here

Prediction of the area

Due to the highly detailed data we had to split the prediction into different blocks. This process allows the computer to run separately each prediction.

The uncertainty of the model is based on the following equation from Yan et al. (2018).

$$Uncertainty = \frac{Qp\ 0.95 - Qp\ 0.05}{Qp\ 0.5}$$

Visualisation and comparison with SoilGrid product

Preparation of the environment

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Visualisation of the final products

Visualisations of the final products can also be reached on Shiny app website at https://mathias-bellat.shinyapps.io/Northern-Kurdistan-map/.

We produced also maps of all properties prediction and uncertainty.

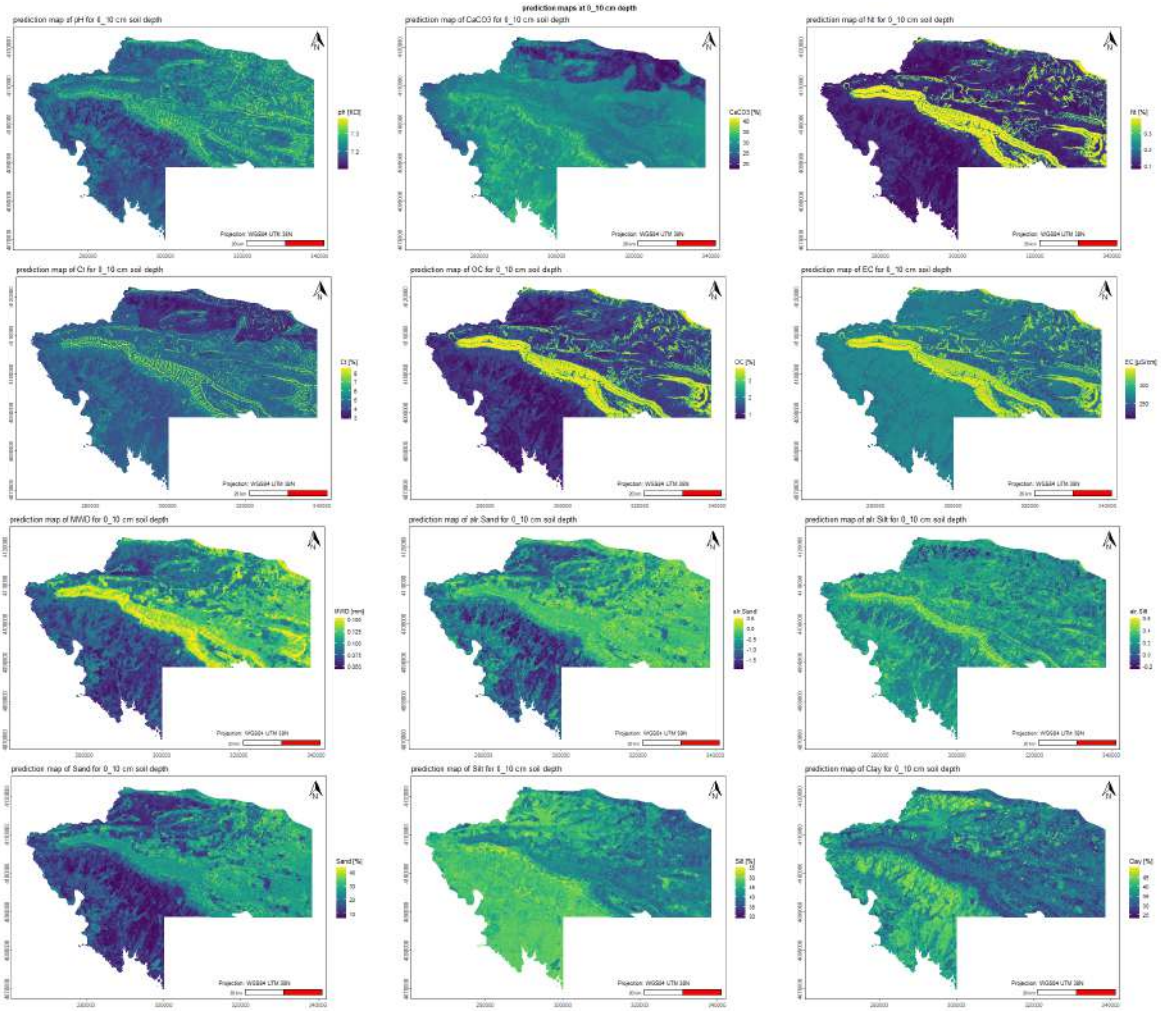

Figure S111: DSM predictions for 0 - 10 cm.

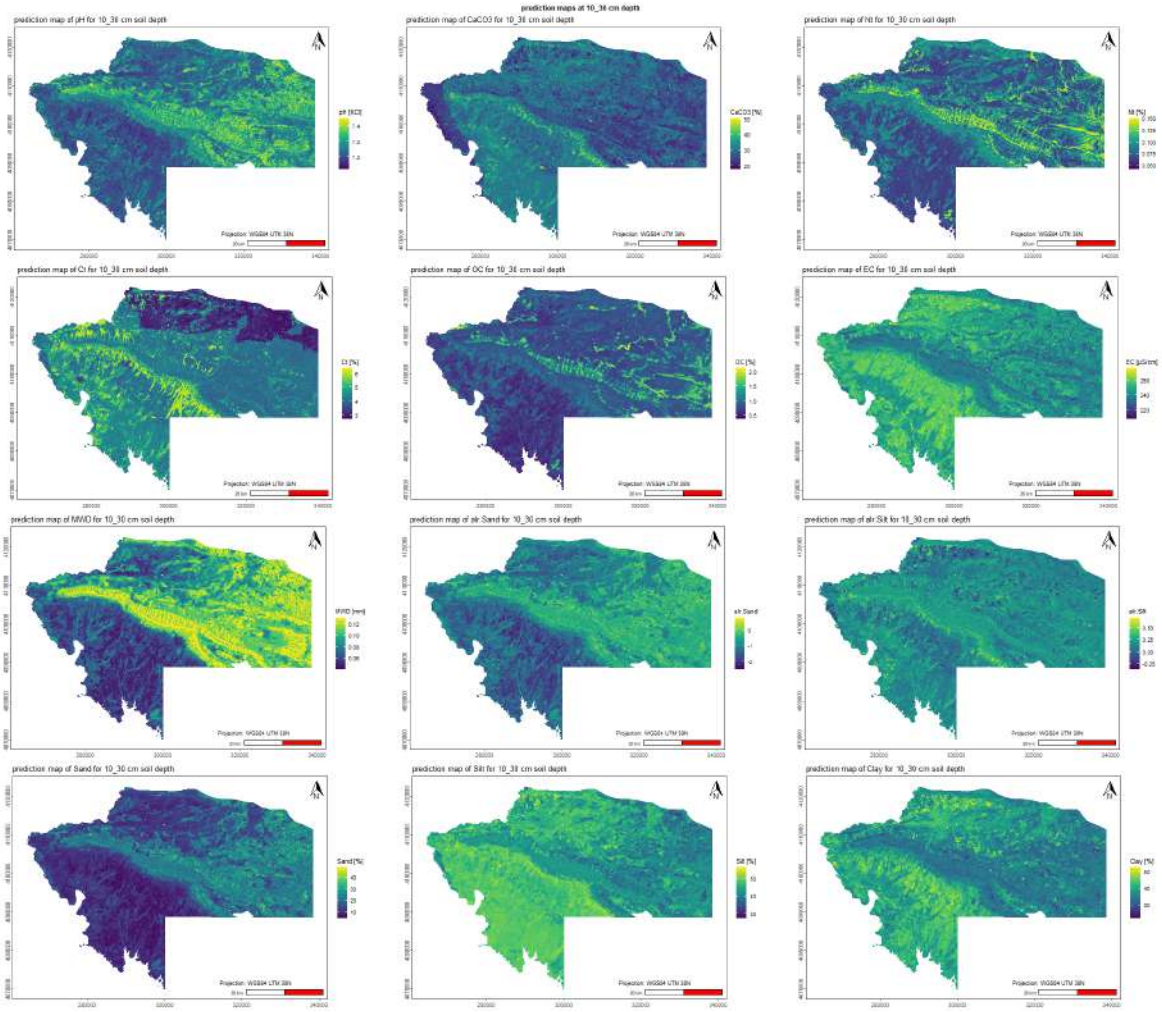

Figure S112: DSM predictions for 10 - 30 cm.

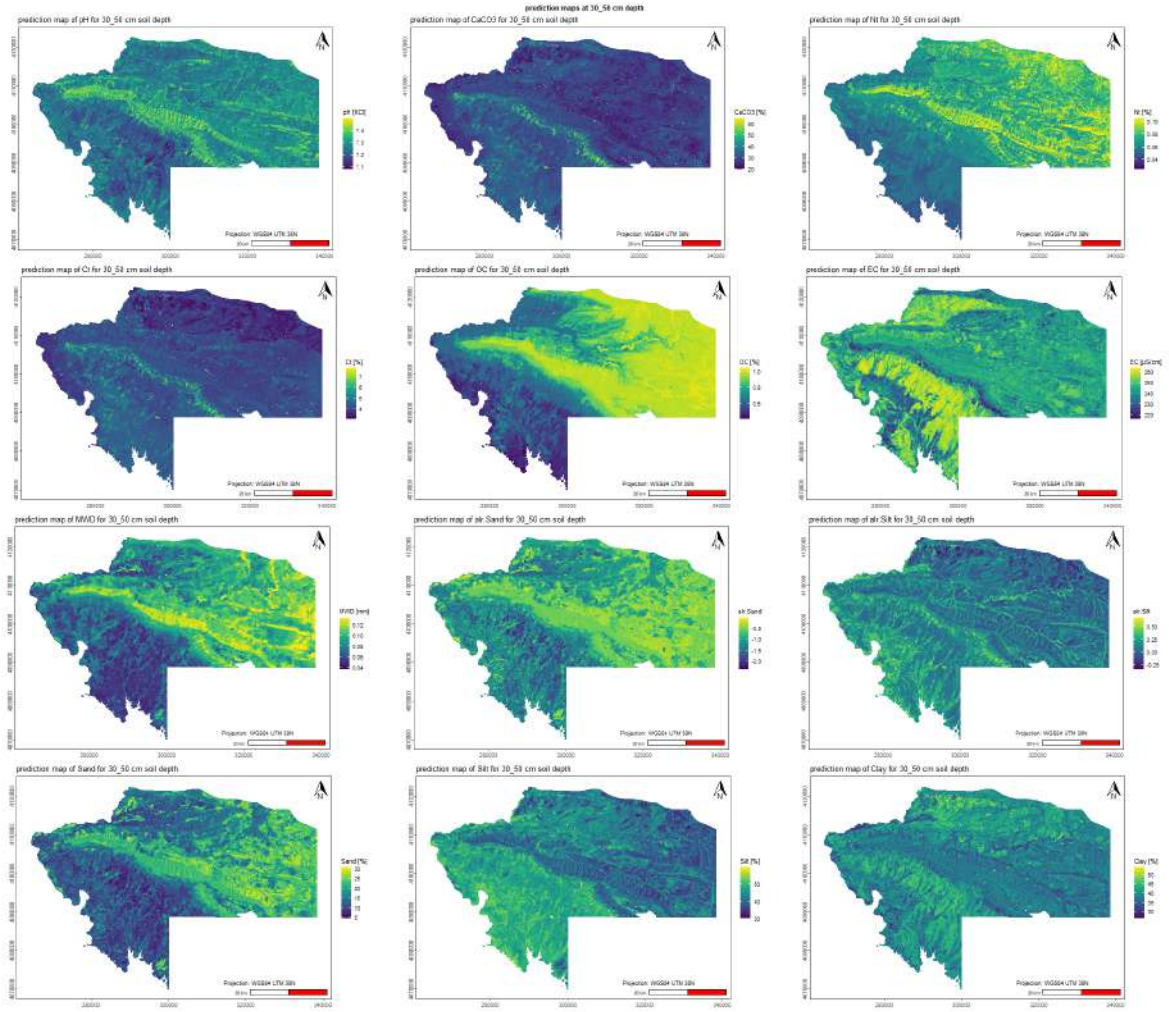

Figure S113: DSM predictions for 30 - 50 cm.

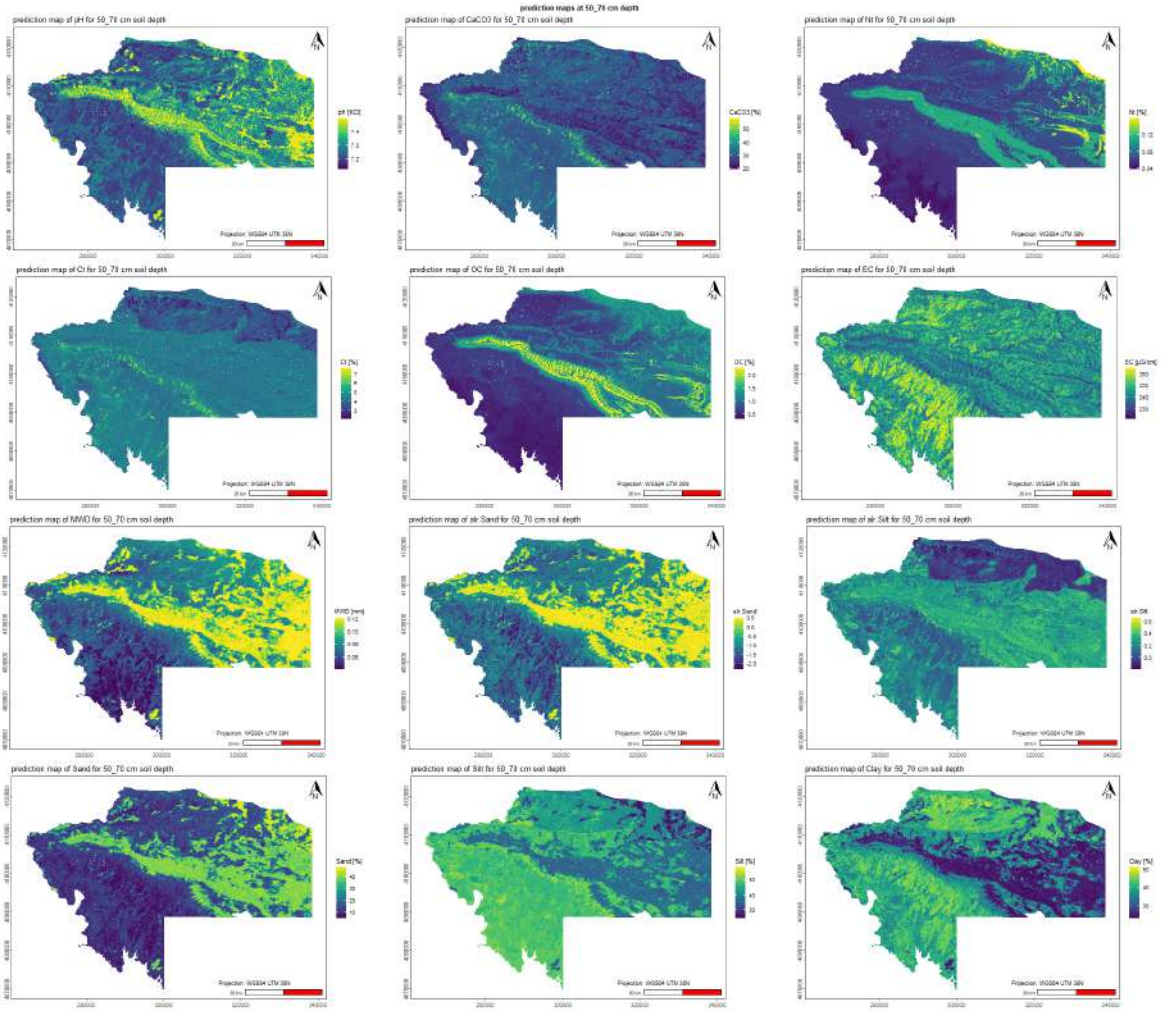

Figure S114: DSM predictions for 50 - 70 cm.

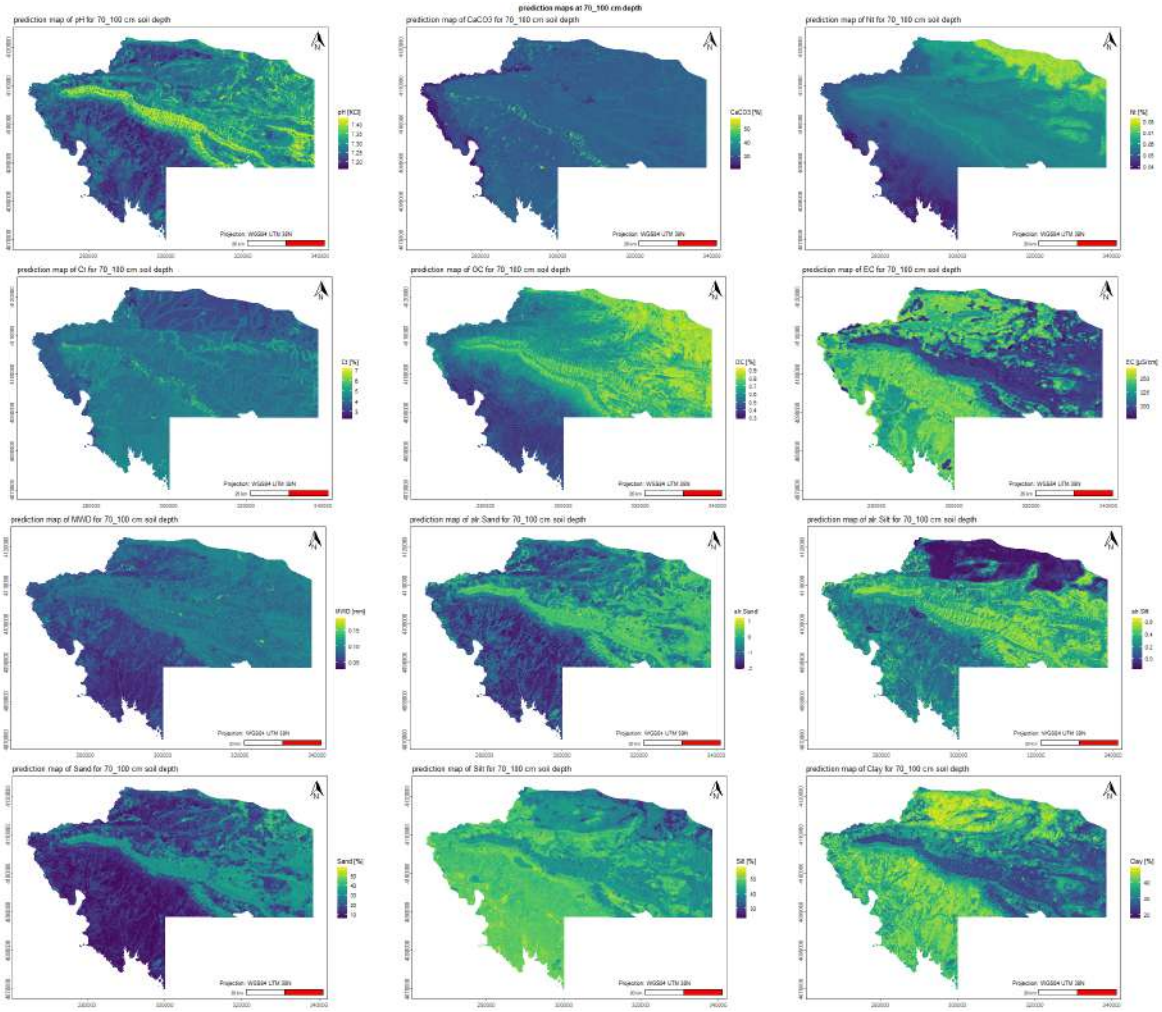

Figure S115: DSM predictions for 70 - 100 cm.

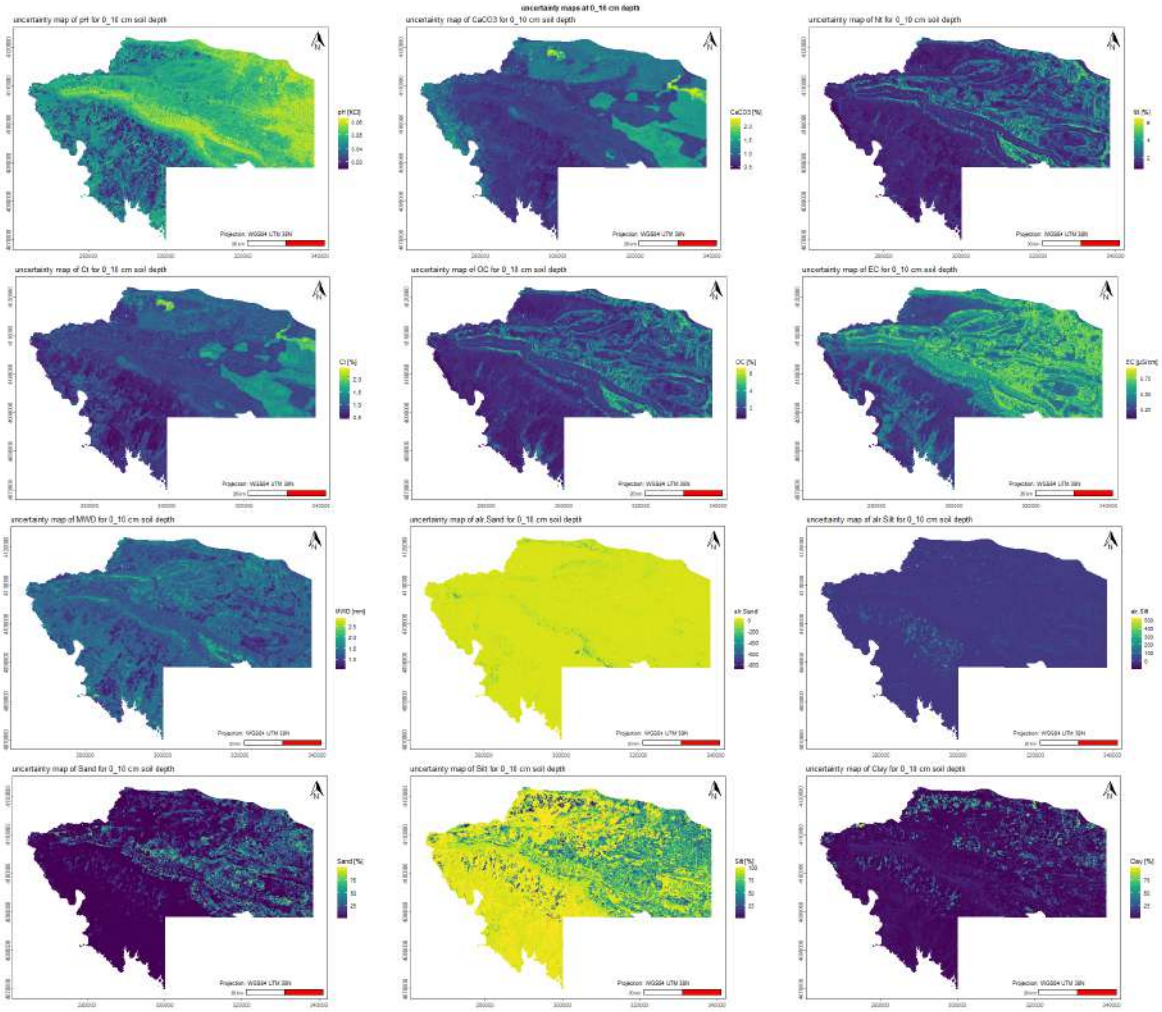

Figure S116: DSM uncertainties for 0 - 10 cm.

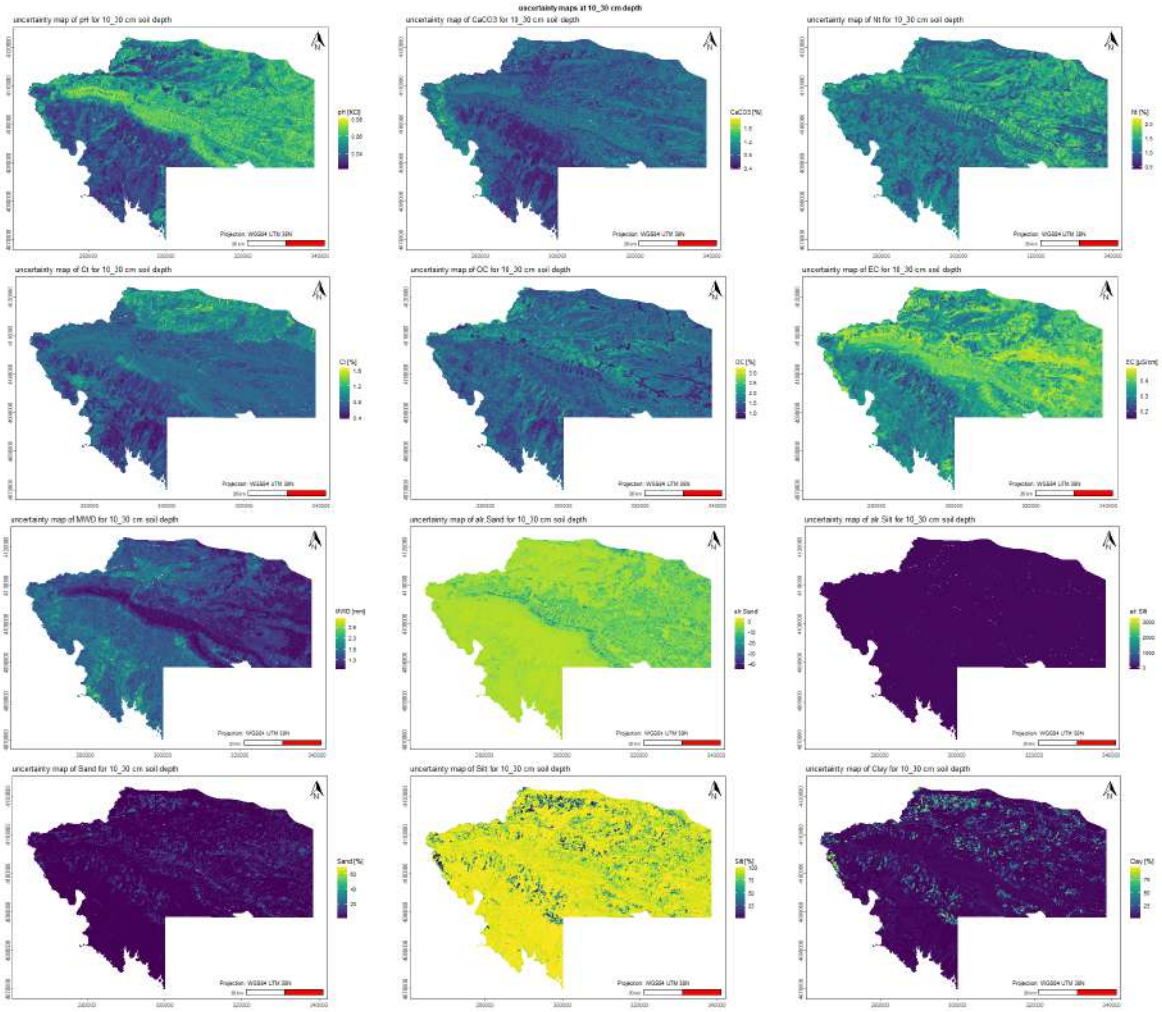

Figure S117: DSM uncertainties for 10 - 30 cm.

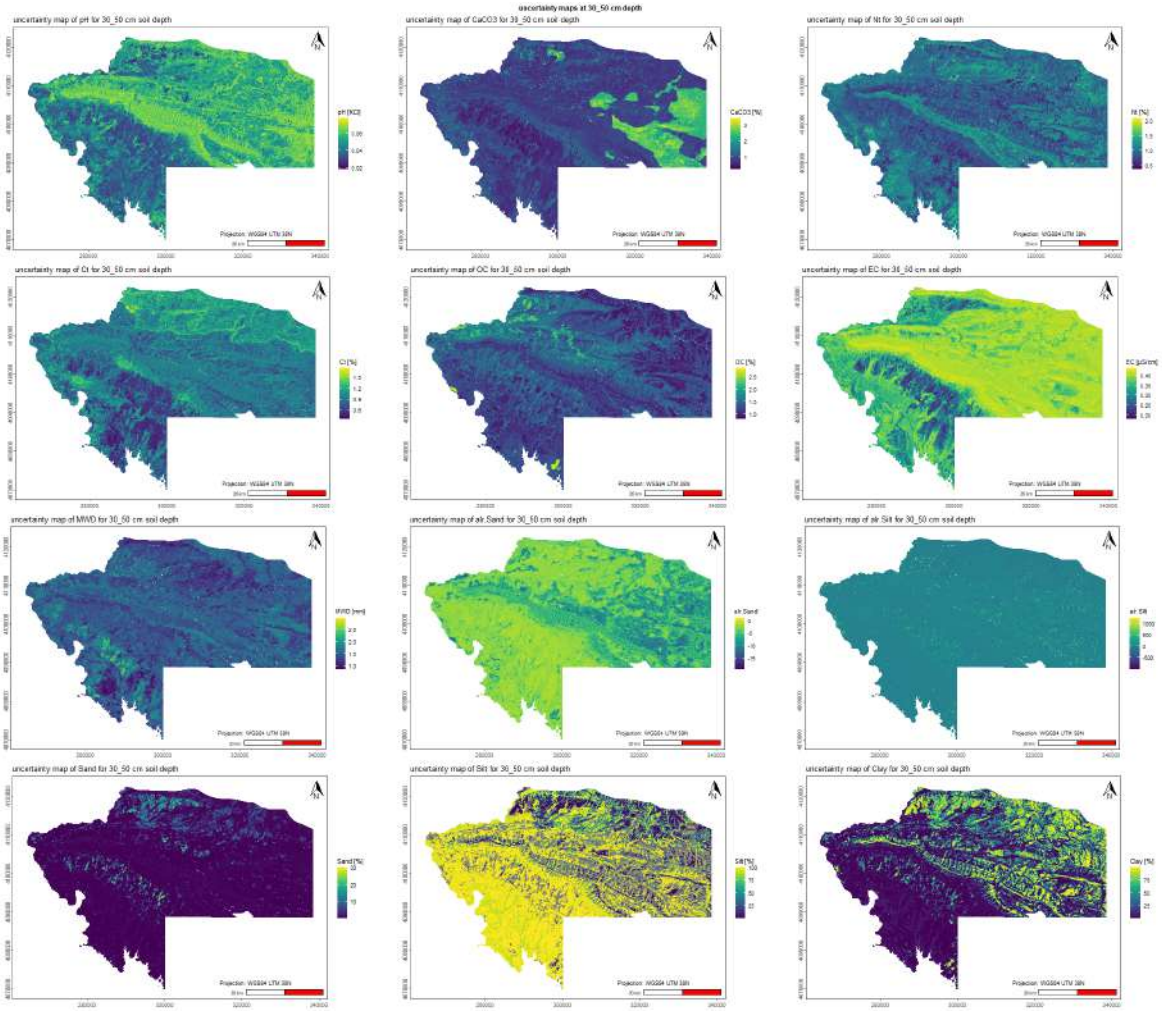

Figure S118: DSM uncertainties for 30 - 50 cm.

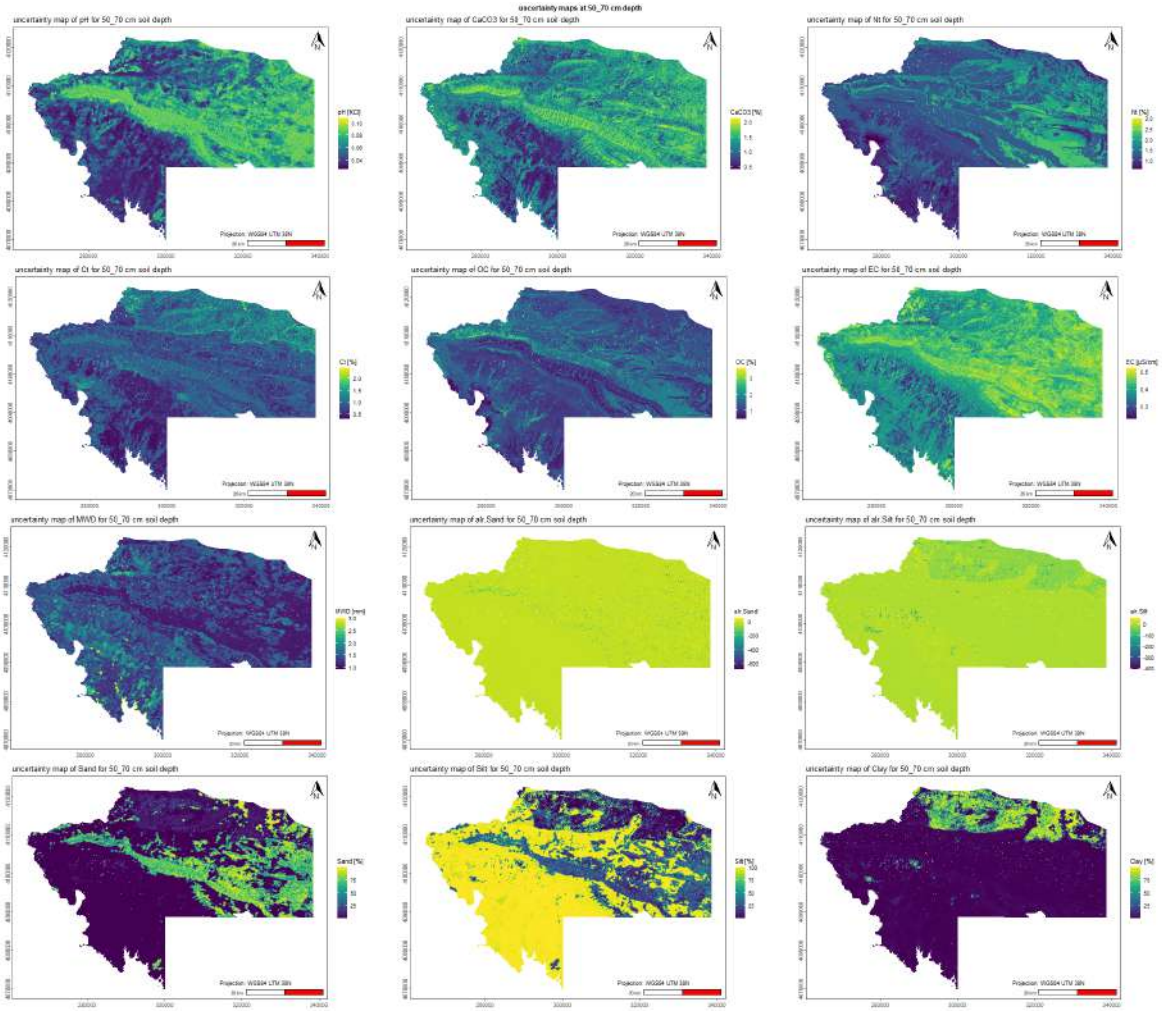

Figure S119: DSM uncertainties for 50 - 70 cm.

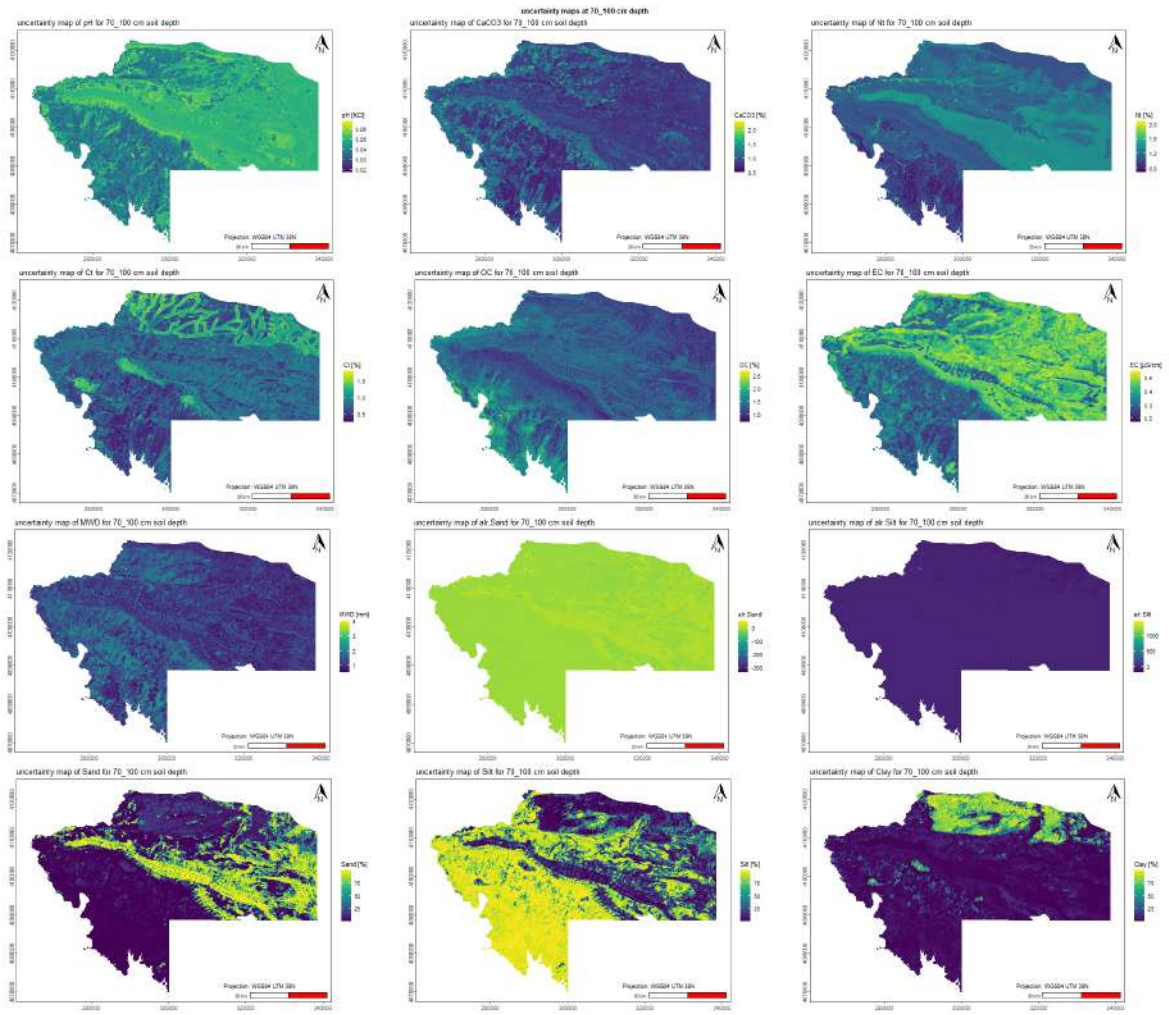

Figure S120: DSM uncertainties for 70 - 100 cm.

A better visualisation can be accessed on the only material at https://mathias-bellat.github.io/DSM-Kurdistan/.

SoilGrid evaluation and preparation

We choose to compute top, sub and lower soil to evaluate the SoilGrid 2.0 product with our predictions. Our interval were based on top-soil with 0 - 30 cm; sub-soil with 30 - 70 for our own prediction and 30 - 60 cm for the SoilGrid; lower-soil with 70 - 100 cm for our own prediction and 60 - 100 for SoilGrid 2.0 product.

We had to standardised the values of our prediction to compare them with the SoilGrid 2.0 product (Committee 2015).

Add table

Bivariate and density plot of the two predictions

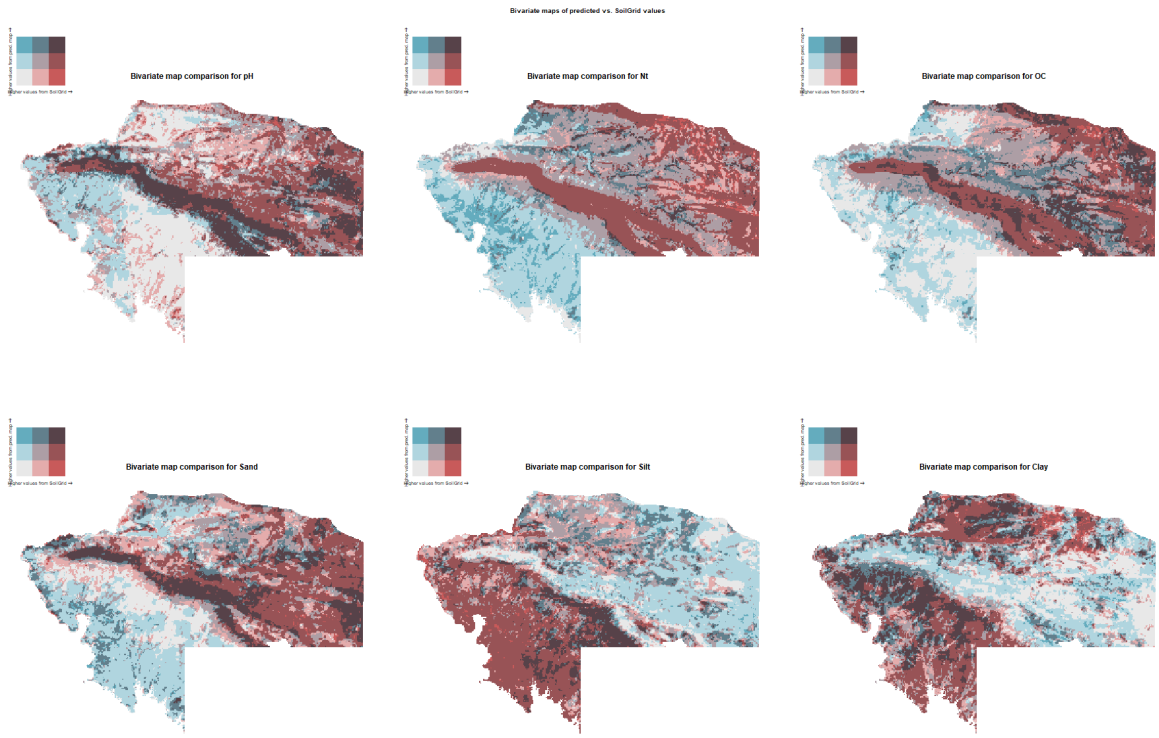

Figure S121: Bivariate map.

Soil depth mapping

Introduction

Purpose

This document is meant to present the methodology we used for soil depth mapping. Our data are based on the sampling campaigns 2022 - 2023 and additional remote sensing observation

for bare rock formations. We based our approach on Liu et al. (2022) and added small modifications.

Covariates

All the data listed below are available freely online excepted for the digitized maps (Geology).

Table S13: Data used in the production of the soil depth mapping.

Name/ID	Original resolution (m)	Type/Unit
Aspect/TE.1	30	Radians
DEM/TE.5	30	Meters
General curvature/TE.7	30	-
MrRTF/TE.8	30	-
MrVBF/TE.9	30	-
Plan curvature/TE.12	30	-
Profile curvature/TE.14	30	-
Slope/TE.16	30	Radians
TPI/TE.21	30	-
Landsat 5 B2/LA5.1	30	0.52 - 0.60 µm
Landsat 5 B3/LA5.2	30	0.63 - 0.69 µm
Landsat 5 B4/LA5.3	30	0.76 - 0.90 µm
Landsat 5 B7/LA5.4	30	2.08 - 2.35 µm
Landsat 5 NDVI/LA5.5	30	-
Landsat 5 NDWI/LA5.6	30	-
LST Apr-May/LST.1	1000	Kelvin
LST Feb-Mar/LST.2	1000	Kelvin
LST Jun-Jul/LST.3	1000	Kelvin
LST Oct-Nov/LST.4	1000	Kelvin
Geology/OT.2	1 : 250 000	Lithology
Landuse/OT.4	10	Landuse class
PET sum/OT.5	750	mm
Prec. sum/OT.6	1000	mm
SRAD sum/OT.7	1000	MJ m^{-2}
Wind sum/OT.9	1000	m s^{-1}
Temp avg/OT.10	1000	°C

Terrain

All terrain covariates were computed the same way as in the previous chapter.

Remote sensing images and indexes

The Landsat 5 images were collected *via* a Google Earth Engine script on a period covering 1990 - 2010, the median of the composite image from *Collection 2 Tier 1 TOA* was used. The land surface temperature (LST) were computed also on Google Earth Engine with a time covering from 2012 to 2016 from the Suomi National Polar-Orbiting satellite. We took the average of each period: February - Mars, April - May, June - July and October November from the *VIIRS* instrument. The javascript codes for scraping these images are available in the supplementary file inside the "8 - Soil depth/code" folder. We computed the following indexes: Normalized Difference Vegetation Index (**NDVI**) (McFeeters 1996) and Normalized Difference Water Index (**NDWI**) (Rouse et al. 1974).

$$NDVI = (NIR - Red)/(NIR + Red)$$

$$NDWI = (Green - NIR)/(Green + NIR)$$

Soil depths

The soil depths of the **122** soil profiles were measured during the field campaign and can be found found online at https://doi.org/10.1594/PANGAEA.973714. Additional **25** site with no soil (or zero value to soil depth), were added after looking at *Bing* and *Google earth* satellite image. We only selected bare rock formation for these sites.

Soil depth mapping prediction

Preparation of the environment

This script is accessible on the online dynamic material https://doi.org/10.57754/FDAT.d5h1h-4x027/ and GitHub repository https://mathias-bellat.github.io/DSM-Kurdistan/.

Preparation of the data

All raster were sampled to 30 x 30 m tiles to match the DEM. We used **bilinear** method excepted for the discrete maps (geology) where we used **ngb** resampling.

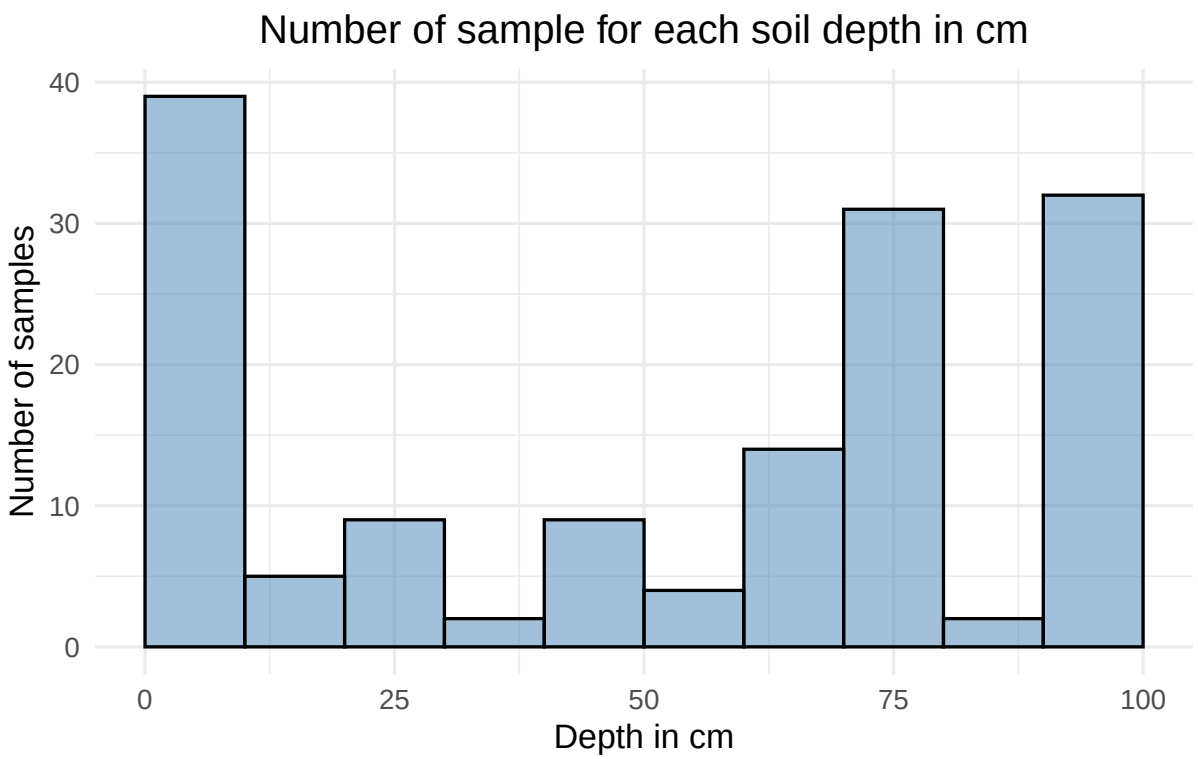

Figure S122: Soil depth values distribution.

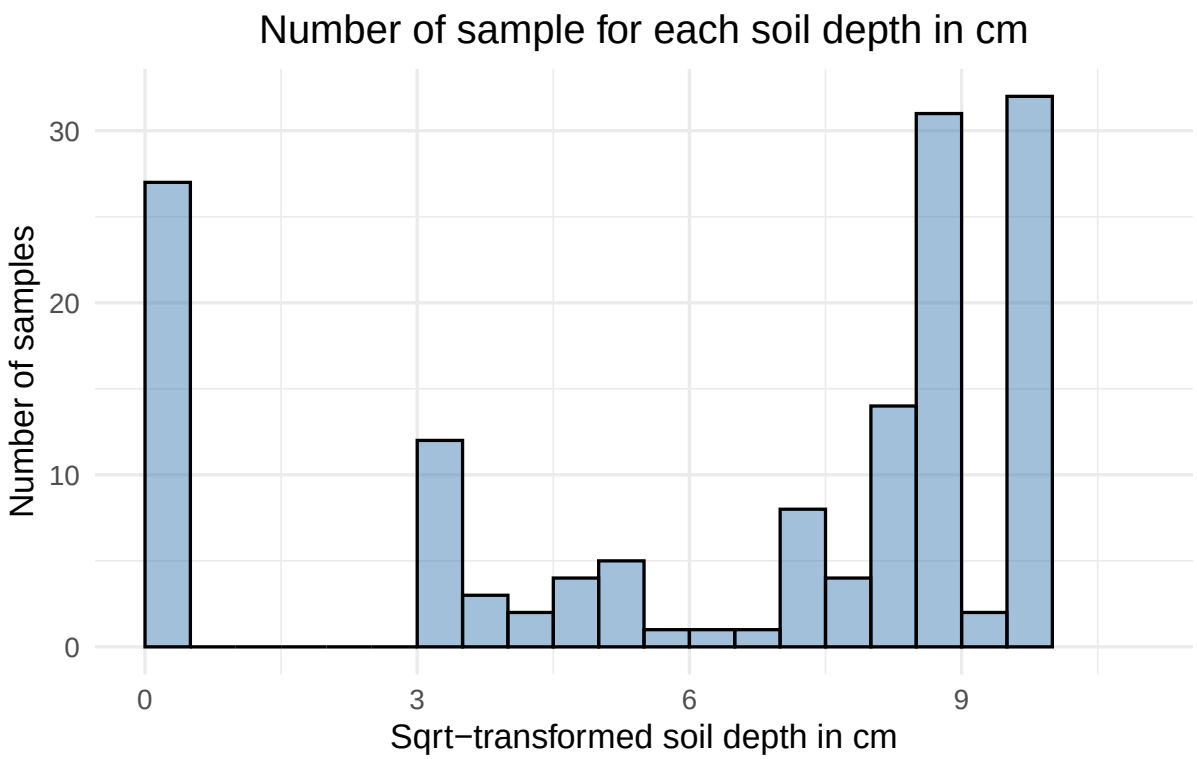

Figure S123: Soil depth values distribution sqrt.

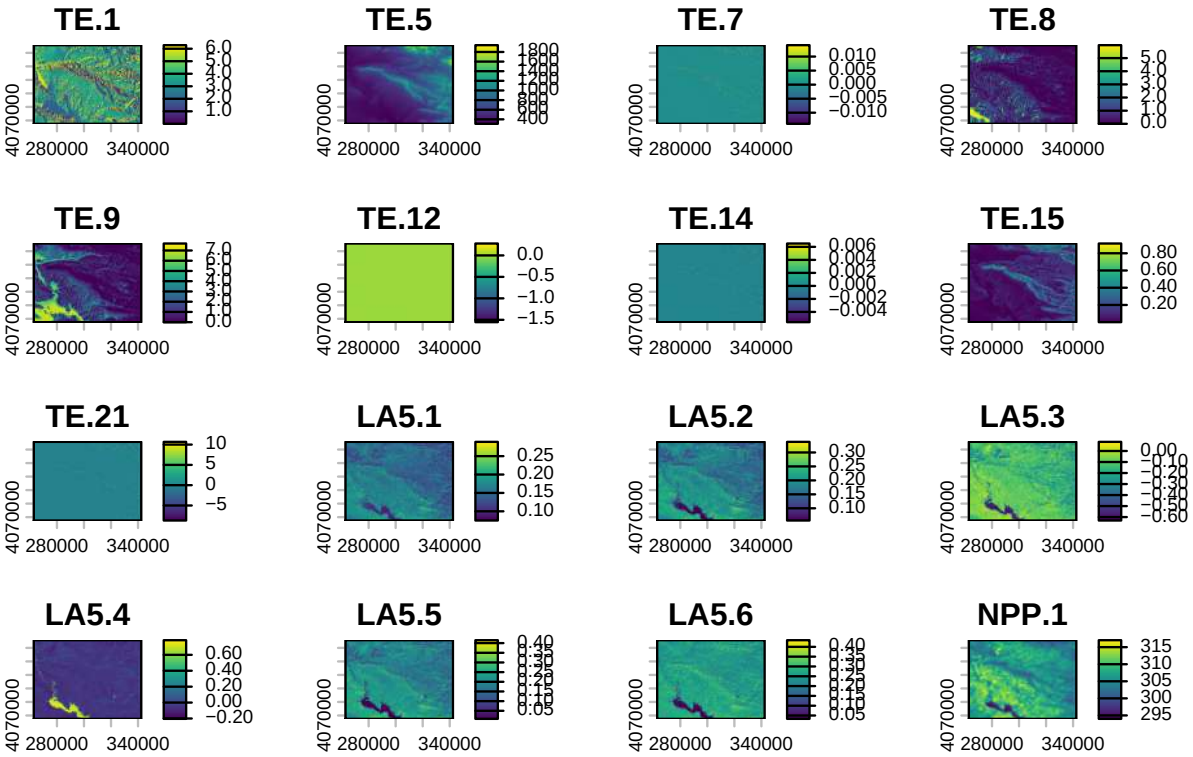

Figure S124: Covariates from soil depth model.

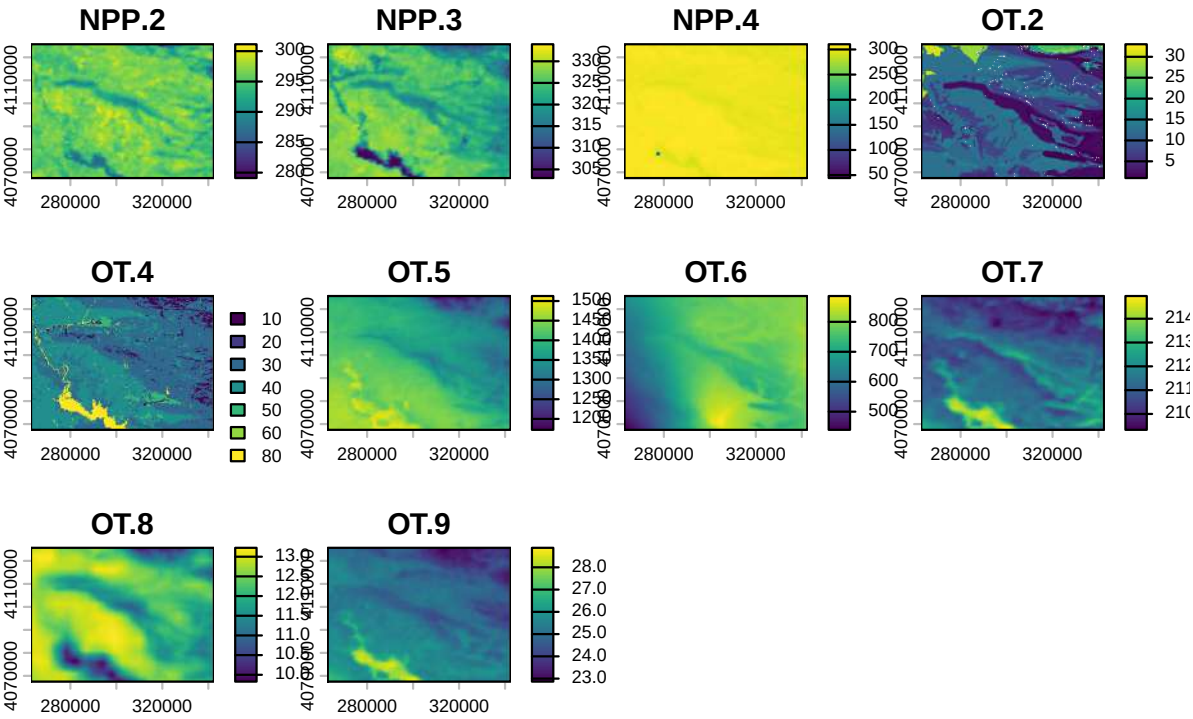

Figure S125: Covariates from soil depth model.

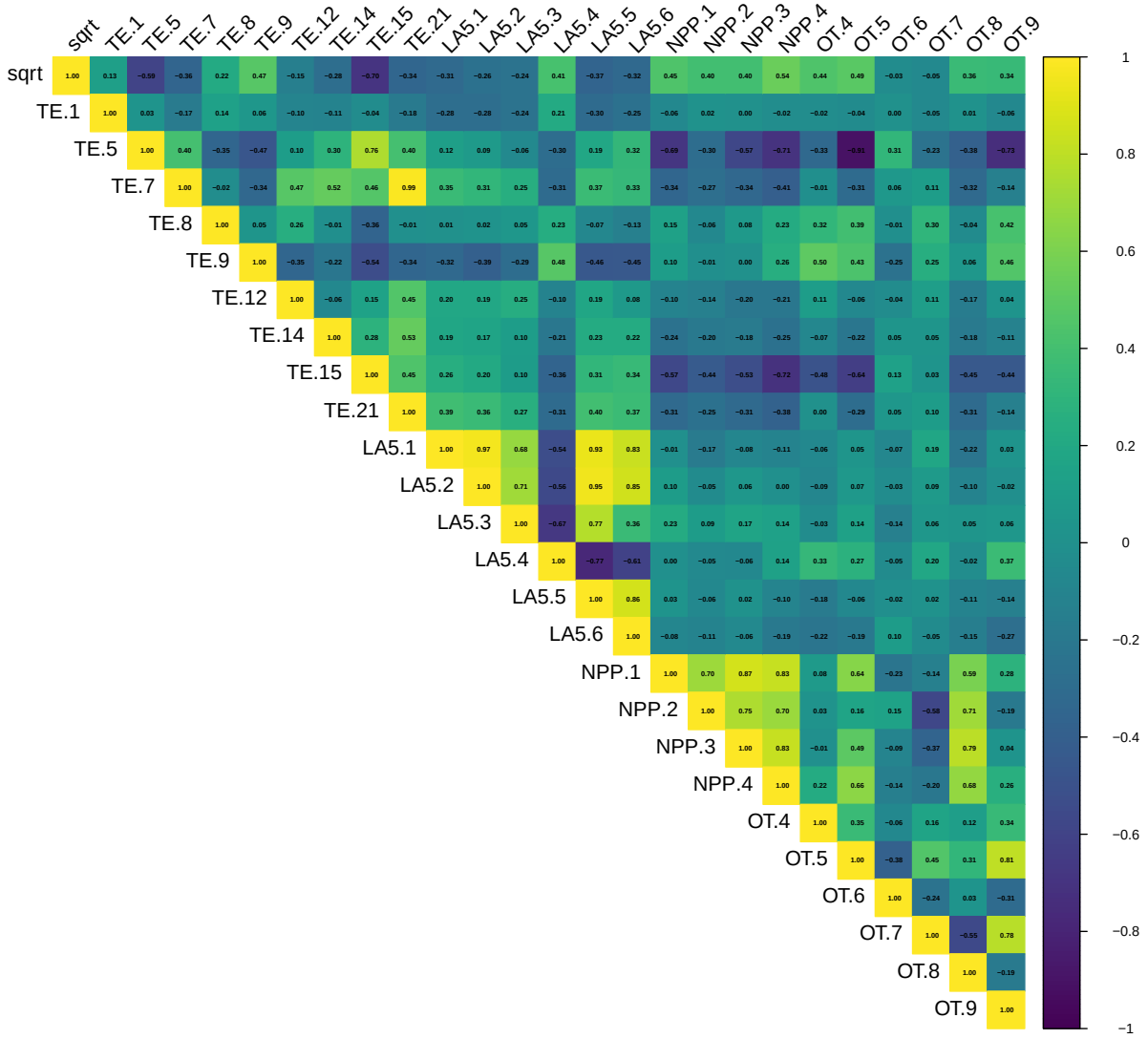

Figure S126: Auto-correlation plot soil depth covariates.

Regarding the correlation plot, the Landsat 5 bands and land surface temperature layers have correlation index over 0.75, as some transformed layers based on DEM (curvature, slope ,aspect ...) and the solar radiation with wind.

Tune the model

The QRF model used is similar to the one for DSM.

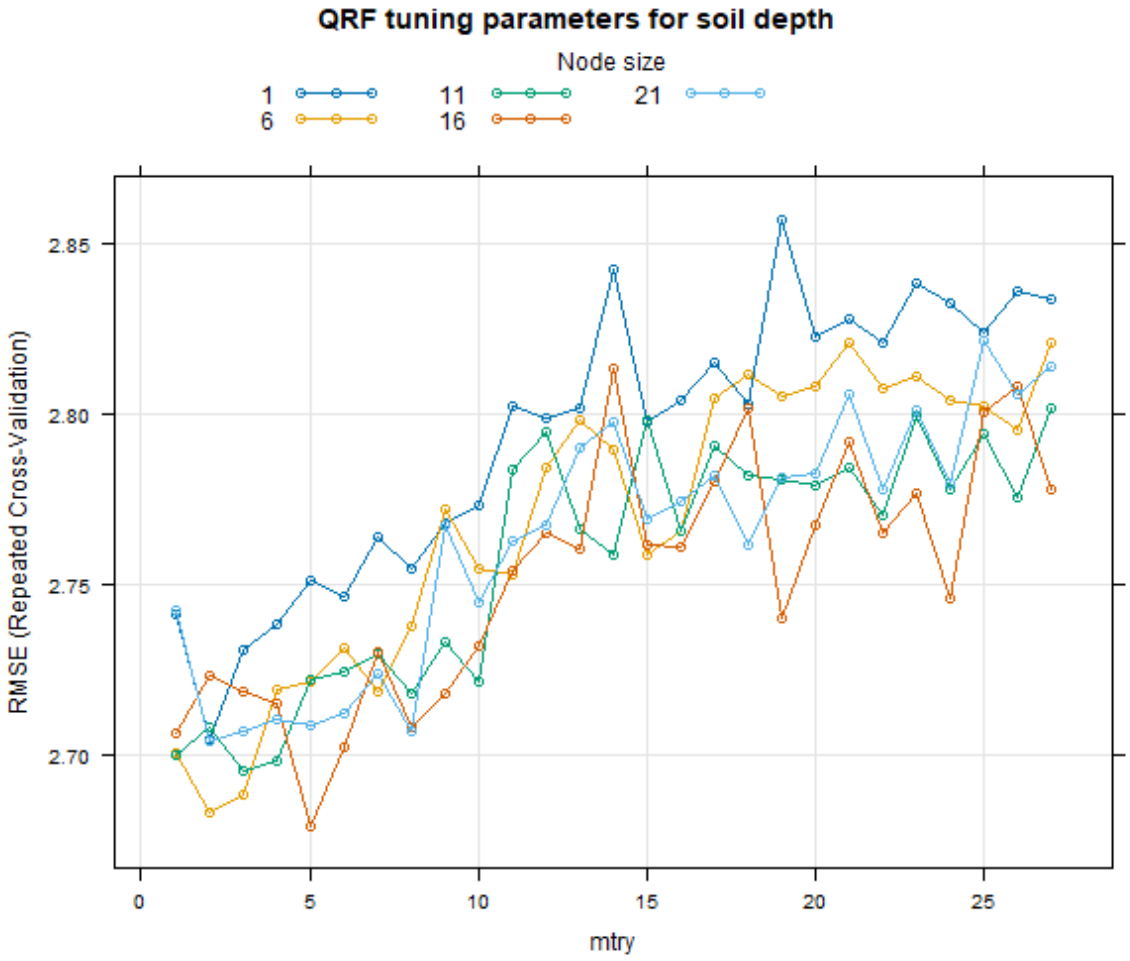

Figure S127: Hyperparameters for soil depth.

Variables imortance

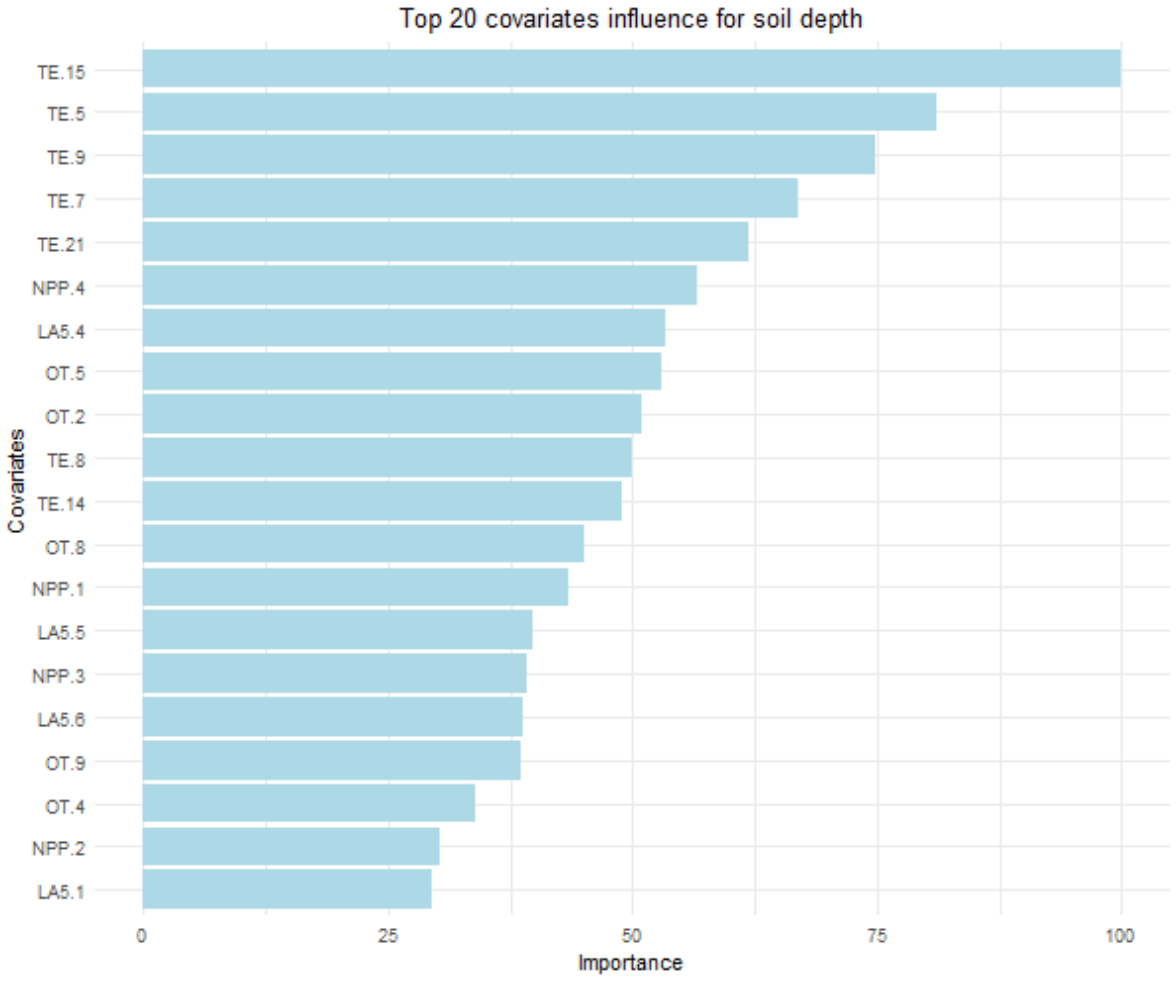

Figure S128: Variables importance for soil depth.

Model evaluation

We used four metrics to evaluate the model:

- ME = Bias with the mean error.
- $RMSE$ = Root mean square error.
- R^2 = Coefficient of determination, also called rsquared.
- $PICP$ = Prediction interval coverage probability set at 90% interval

$$ME = \frac{1}{n}\sum_{i=n}^{n}(\hat{y}_i - y_i)$$

Where: n is the number of observations, y_i the observed value for i, and $\hat{y}_i$ the predicted value for i.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}}$$

Where: n is the number of observations, y_i the observed value for i, and $\hat{y}_i$ the predicted value for i.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where: y_i the observed value for i, $\hat{y}_i$ the predicted value for i, and $\bar{y}$ is the mean of observed values.

$$PICP = \frac{1}{v}\ count\ jj = PL_j^L \leq t_j \leq PL_j^U$$

Where: v is the number of observations, obs_i is the observed value for i, PL_i^L is the lower prediction limit for i, and PL_i^U is the upper prediction limit for i.

```
   ME_mean RMSE_mean   R2_mean PICP_mean    ME_sd  RMSE_sd     R2_sd    PICP_sd
1 3.269185  30.75852 0.3908253 0.9887103 6.232449 5.033474 0.1634309 0.02571429
```

Mapping the soil depth of the area

Due to the highly detailed data we had to split the prediction into different blocks. This process allows the computer to run separately each prediction.

The uncertainty of the model is based on the following equation from Yan et al. (2018).

$$Uncertainty = \frac{Qp\ 0.95 - Qp\ 0.05}{Qp\ 0.5}$$

Visualisations and exports

Colorblind visualisations

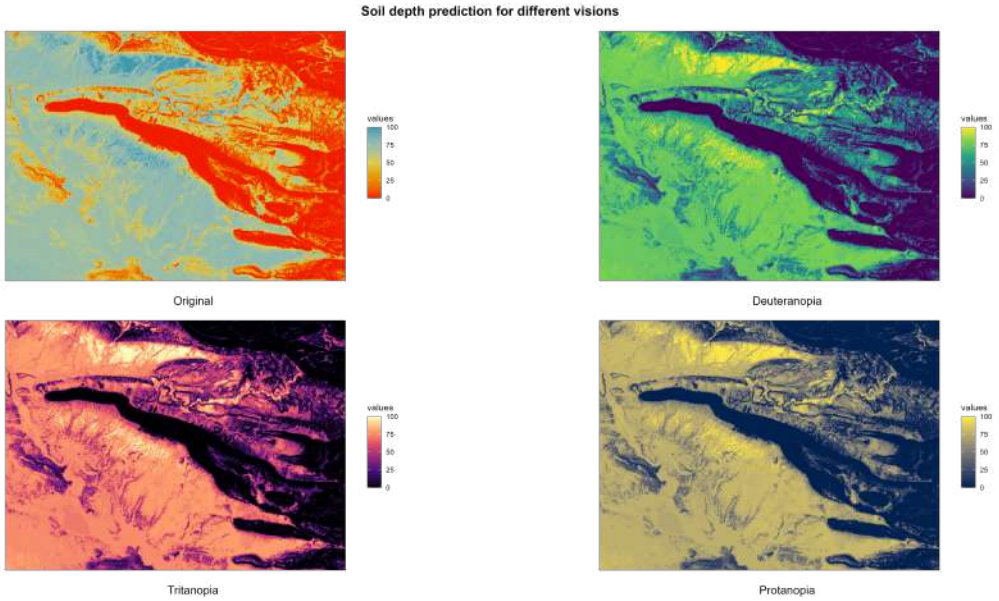

Figure S129: Color-blind soil depth prediction.

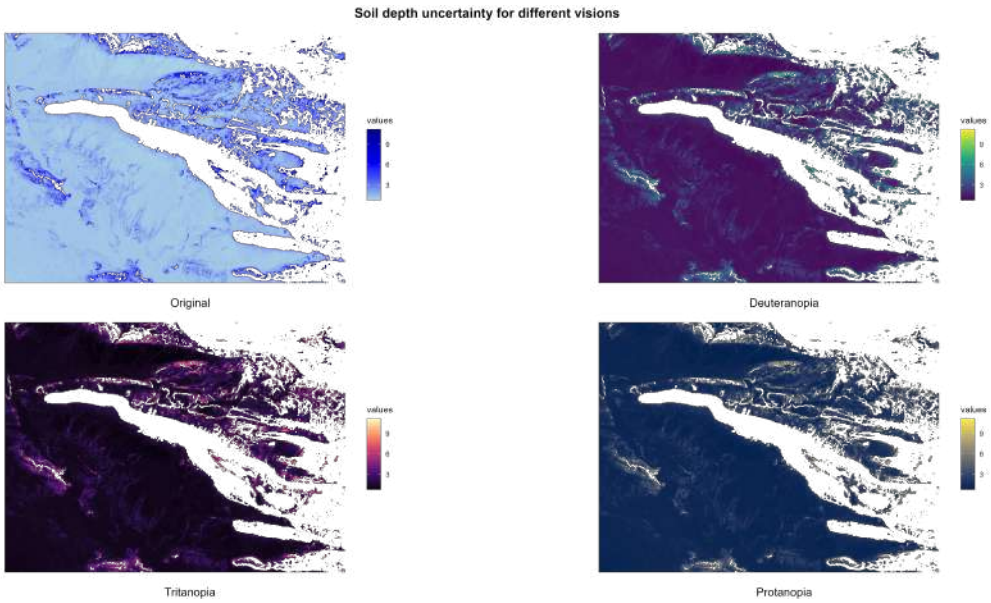

Figure S130: Color-blind soil depth uncertainty.

Export final maps

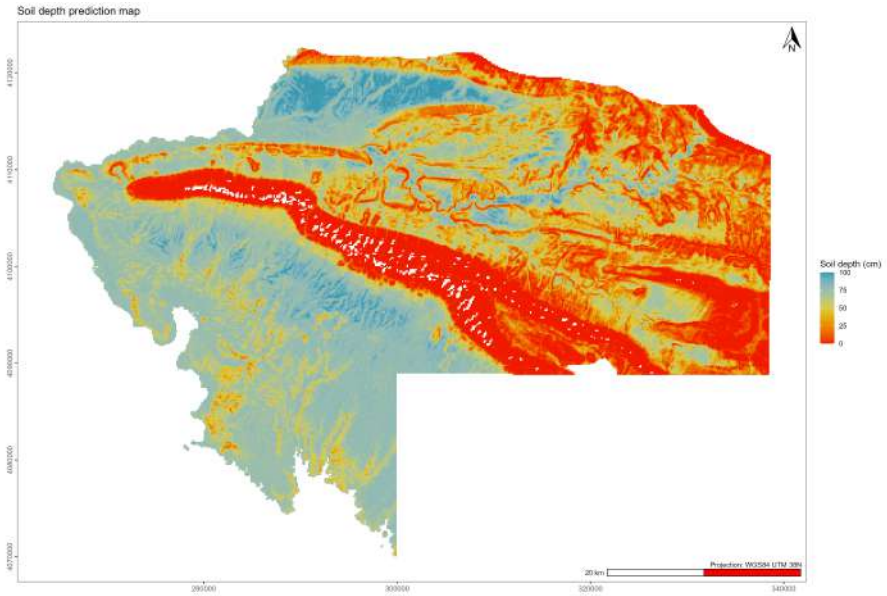

Figure S131: Map depth prediction.

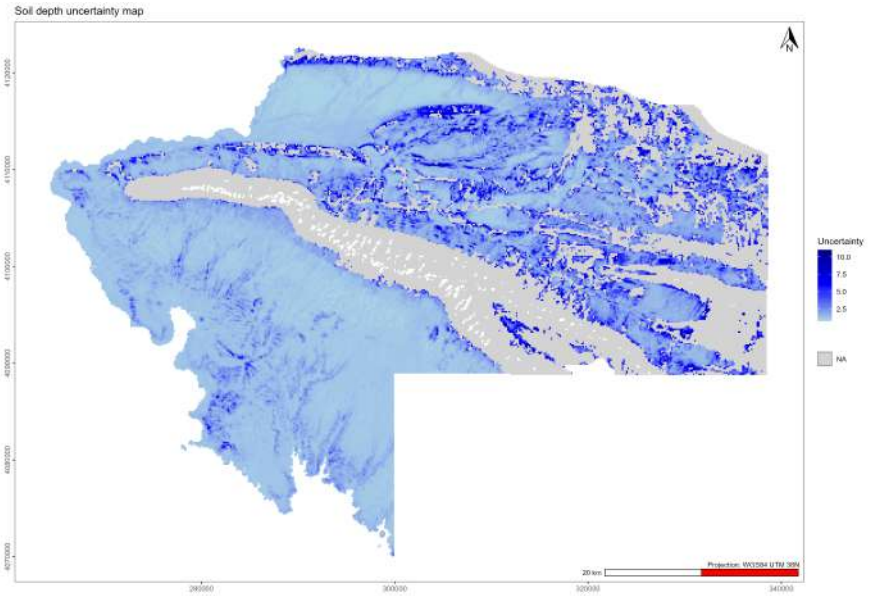

Figure S132: Map soil depth uncertainty.

Abuelgasim, Abdelgadir, and Rubab Ammad. 2019. "Mapping Soil Salinity in Arid and Semi-Arid Regions Using Landsat 8 OLI Satellite Data." *Remote Sensing Applications: Society and Environment* 13 (January): 415–25. https://doi.org/10.1016/j.rsase.2018.12.010.

Almagro, André, Thais Caregnatto Thomé, Carina Barbosa Colman, Rodrigo Bahia Pereira, José Marcato Junior, Dulce Buchala Bicca Rodrigues, and Paulo Tarso Sanches Oliveira. 2019. "Improving Cover and Management Factor (C-Factor) Estimation Using Remote Sensing Approaches for Tropical Regions." *International Soil and Water Conservation Research* 7 (4): 325–34. https://doi.org/10.1016/j.iswcr.2019.08.005.

Al-Mousawi, Hala, Saffa Fouad, and Varoujan Sissakian. 2007. "Geological Map of Zakho Quadrangle." Geological map. Baghdad: GEOSURV.

Bahrami, Amir, Majid Danesh, and Mehdi Bahrami. 2022. "Studying Sand Component of Soil Texture Using the Spectroscopic Method." *Infrared Physics & Technology* 122 (May): 104056. https://doi.org/10.1016/j.infrared.2022.104056.

Boettinger, J. L., R. D. Ramsey, J. M. Bodily, N. J. Cole, S. Kienast-Brown, S. J. Nield, A. M. Saunders, and A. K. Stum. 2008. "Landsat Spectral Data for Digital Soil Mapping." In *Digital Soil Mapping with Limited Data*, edited by Alfred E. Hartemink, Alex McBratney, and Maria de Lourdes Mendonça-Santos, 193–202. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_16.

Bousbih, Safa, Mehrez Zribi, Charlotte Pelletier, Azza Gorrab, Zohra Lili-Chabaane, Nicolas Baghdadi, Nadhira Ben Aissa, and Bernard Mougenot. 2019. "Soil Texture Estimation Using Radar and Optical Data from Sentinel-1 and Sentinel-2." *Remote Sensing* 11 (13):

1520. https://doi.org/10.3390/rs11131520.
Brus, Dick J. 2022. *Spatial Sampling with R*. 1st ed. Boca Raton: Chapman; Hall/CRC. https://doi.org/10.1201/9781003258940.
Chandrasekar, K., M. V. R. Sesha Sai, P. S. Roy, and R. S. Dwevedi. 2010. "Land Surface Water Index (LSWI) Response to Rainfall and NDVI Using the MODIS Vegetation Index Product." *International Journal of Remote Sensing* 31 (15): 3987–4005. https://doi.org/10.1080/01431160802575653.
Committee, Science. 2015. "Specifications Tiered GlobalSoilMap Products." Technical report Release 2.4.
Cossart, Etienne, Mathieu Fressard, and Brian Chaize. 2020. "Spatial Patterns of Vineyard Landscape Evolution and Their Impacts on Erosion Susceptibility: RUSLE Simulation Applied in Mercurey (Burgundy, France) Since the Mid-20th Century." *ERDKUNDE* 74 (4): 281–300. https://doi.org/10.3112/erdkunde.2020.04.04.
Deering, D. W. 1975. "Measuring Forage Production of Grazing Units from Landsat MSS Data." In *Proceedings of 10th International Symposium on Remote Sensing of Environment, 1975*, 1169–78. ERIM. https://cir.nii.ac.jp/crid/1573387449827680896.
Ertan, Nazlan. 2022. "Turkey Launches Offensive Against PKK Targets in Northern Iraq." *Al-Monitor*, April. https://www.al-monitor.com/originals/2022/04/turkey-launches-offensive-against-pkk-targets-northern-iraq.
ESA, and Airbus. 2022. "Copernicus DEM." https://doi.org/10.5270/ESA-c5d3d65.
FAO, and IIASA. 2023. *Harmonized World Soil Database Version 2.0*. FAO; International Institute for Applied Systems Analysis (IIASA); https://doi.org/10.4060/cc3823en.
Fernández-Buces, N., C. Siebe, S. Cram, and J. L. Palacio. 2006. "Mapping Soil Salinity Using a Combined Spectral Response Index for Bare Soil and Vegetation: A Case Study in the Former Lake Texcoco, Mexico." *Journal of Arid Environments* 65 (4): 644–67. https://doi.org/10.1016/j.jaridenv.2005.08.005.
Fick, Stephen E., and Robert J. Hijmans. 2017. "WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37 (12): 4302–15. https://doi.org/10.1002/joc.5086.
Forti, Luca, Alessandro Perego, Filippo Brandolini, Guido S. Mariani, Mjahid Zebari, Kathleen Nicoll, Eleonora Regattieri, et al. 2021. "Geomorphology of the Northwestern Kurdistan Region of Iraq: Landscapes of the Zagros Mountains Drained by the Tigris and Great Zab Rivers." *Journal of Maps* 17 (2): 225–36. https://doi.org/10.1080/17445647.2021.1906339.
Gao, Bo-cai. 1996. "NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space." *Remote Sensing of Environment* 58 (3): 257–66. https://doi.org/10.1016/S0034-4257(96)00067-3.
Ge, Y., Alexandre M. J.-C. Wadoux, and Y Peng. 2022. *A Primer on Soil Analysis Using Visible and Near-Infrared (Vis-NIR) and Mid-Infrared (MIR) Spectroscopy*. FAO.
Hengl, Tomislav. 2007. *A Practical Guide to Geostatistical Mapping of Environmental Variables*. Edited by European Commission. Luxembourg: Publications Office.
Henrich, V., G. Krauss, C. Götze, and C. Sandow. 2012. "IDB – Www.indexdatabase.de, Entwicklung Einer Datenbank Für Fernerkundungsindizes." Poster. Bochum. https://www.indexdatabase.de/info/credits.php.

Huete, A. R. 1988. "A Soil-Adjusted Vegetation Index (SAVI)." *Remote Sensing of Environment* 25 (3): 295–309. https://doi.org/10.1016/0034-4257(88)90106-X.

Huete, A, C Justice, and H Liu. 1994. "Development of Vegetation and Soil Indices for MODIS-EOS." *Remote Sensing of Environment* 49 (3): 224–34. https://doi.org/10.1016/0034-4257(94)90018-3.

Kennard, R. W., and L. A. Stone. 1969. "Computer Aided Design of Experiments." *Technometrics* 11 (1): 137–48. https://doi.org/10.2307/1266770.

Khan, Nasir M., Victor V. Rastoskuev, Elena V. Shalina, and Yohei Sato. 2001. "Mapping Salt-Affected Soils Using Remote Sensing Indicators-a Simple Approach with the Use of GIS IDRISI." In *22nd Asian Conference on Remote Sensing*. Singapore: Asian Association on remote Sensing. https://acrs-aars.org/proceeding/ACRS2001/Papers/AGS-05.pdf.

Kouli, Maria, Pantelis Soupios, and Filippos Vallianatos. 2009. "Soil Erosion Prediction Using the Revised Universal Soil Loss Equation (RUSLE) in a GIS Framework, Chania, Northwestern Crete, Greece." *Environmental Geology* 57 (3): 483–97. https://doi.org/10.1007/s00254-008-1318-9.

Liu, Feng, Fei Yang, Yu-guo Zhao, Gan-lin Zhang, and De-cheng Li. 2022. "Predicting Soil Depth in a Large and Complex Area Using Machine Learning and Environmental Correlations." *Journal of Integrative Agriculture* 21 (8): 2422–34. https://doi.org/10.1016/S2095-3119(21)63692-4.

Ludwig, Bernard, Isabel Greenberg, Michael Vohland, and Kerstin Michel. 2023. "Optimised Use of Data Fusion and Memory-based Learning with an Austrian Soil Library for Predictions with Infrared Data." *European Journal of Soil Science* 74 (4): e13394. https://doi.org/10.1111/ejss.13394.

McFeeters, S. K. 1996. "The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features." *International Journal of Remote Sensing* 17 (7): 1425–32. https://doi.org/10.1080/01431169608948714.

Mehri, Azade, Hazhir Karimi, Yaseen T. Mustafa, Ayad M. Fadhil Al-Quraishi, and Saman Galalizadeh. 2024. "Predicting Soil Erosion Using RUSLE Model in Duhok Governorate, Kurdistan Region of Iraq." In *Earth and Environmental Sciences Library*, 171–87. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-58315-5_9.

Morgan, R. P. C. 2005. *Soil Erosion and Conservation*. 3rd ed. Malden, MA: Blackwell Pub.

Morgan, R. P. C., D. D. V. Morgan, and H. J. Finney. 1984. "A Predictive Model for the Assessment of Soil Erosion Risk." *Journal of Agricultural Engineering Research* 30 (January): 245–53. https://doi.org/10.1016/S0021-8634(84)80025-6.

Ng, Wartini, Brendan Malone, Budiman Minasny, and Sangho Jeon. 2022. "Near and Mid Infrared Soil Spectroscopy." In *Reference Module in Earth Systems and Environmental Sciences*, B9780128229743000227. Elsevier. https://doi.org/10.1016/B978-0-12-822974-3.00022-7.

Ng, Wartini, Budiman Minasny, Brendan Malone, and Patrick Filippi. 2018. "In Search of an Optimum Sampling Algorithm for Prediction of Soil Properties from Infrared Spectra." *PeerJ* 6 (October): e5722. https://doi.org/10.7717/peerj.5722.

Nield, S. J., J. L. Boettinger, and R. D. Ramsey. 2007. "Digitally Mapping Gypsic and Natric Soil Areas Using Landsat ETM Data." *Soil Science Society of America Journal* 71 (1):

245–52. https://doi.org/10.2136/sssaj2006-0049.

Pouget, M., José Madeira Netto, É L. Floc'h, and S. Kamal. 1991. "Caractéristiques Spectrales Des Surfaces Sableuses de La Région Côtière Nord-Ouest de l'Egypte : Application Aux Données Satellitaires SPOT." In. https://www.semanticscholar.org/paper/Caract%C3%A9ristiques-spectrales-des-surfaces-sableuses-Pouget-Netto/082a77649ddd9d2b2427f96a3fb2659acb1c2883.

Ramirez-Lopez, Leonardo, Karsten Schmidt, Thorsten Behrens, Bas van Wesemael, Jose A. M. Demattê, and Thomas Scholten. 2014. "Sampling Optimal Calibration Sets in Soil Infrared Spectroscopy." *Geoderma* 226-227 (August): 140–50. https://doi.org/10.1016/j.geoderma.2014.02.002.

Renard, Kenneth G., George R. Foster, Glenn A. Weesies, and Jeffrey P. Porter. 1991. "RUSLE: Revised Universal Soil Loss Equation." *Journal of Soil and Water Conservation* 46 (1): 30–33. https://www.jswconline.org/content/46/1/30.

Rouse, J. W., R. H. Haas, D. W. Deering, J. A. Schell, and J. C. Harlan. 1974. "Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation." E75-10354. https://ntrs.nasa.gov/citations/19750020419.

Sissakian, Varoujan, Dikran Hagopian, and Eman Hasan. 1995. "Geological Map of Al-Mosul Quadrangle." Geological map. Baghdad: GEOSURV.

Stenberg, Bo, Raphael A. Viscarra Rossel, Abdul Mounem Mouazen, and Johanna Wetterlind. 2010. "Visible and Near Infrared Spectroscopy in Soil Science." In *Advances in Agronomy*, 107:163–215. Elsevier. https://doi.org/10.1016/S0065-2113(10)07005-7.

Swarnkar, Somil, Anshu Malini, Shivam Tripathi, and Rajiv Sinha. 2018. "Assessment of Uncertainties in Soil Erosion and Sediment Yield Estimates at Ungauged Basins: An Application to the Garra River Basin, India." *Hydrology and Earth System Sciences* 22 (4): 2471–85. https://doi.org/10.5194/hess-22-2471-2018.

Thapa, Pawan. 2020. "Spatial Estimation of Soil Erosion Using RUSLE Modeling: A Case Study of Dolakha District, Nepal." *Environmental Systems Research* 9 (1): 15. https://doi.org/10.1186/s40068-020-00177-2.

USGS. 2018. "Landsat Collections." 2018-3049. U.S. Geological Survey. https://doi.org/10.3133/fs20183049.

Vaysse, Kévin, and Philippe Lagacherie. 2017. "Using Quantile Regression Forest to Estimate Uncertainty of Digital Soil Mapping Products." *Geoderma* 291 (April): 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017.

Viscarra Rossel, R. A., T. Behrens, E. Ben-Dor, D. J. Brown, J. A. M. Demattê, K. D. Shepherd, Z. Shi, et al. 2016. "A Global Spectral Library to Characterize the World's Soil." *Earth-Science Reviews* 155 (April): 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012.

Wadoux, Alexandre M. J.-C., Brendan Malone, Budiman Minasny, Mario Fajardo, and Alex B. McBratney. 2021. *Soil Spectral Inference with R Analysing Digital Soil Spectra*. S.l.: SPRINGER NATURE. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2760028.

WRB, IUSS Working Group. 2006. *Guidelines for Soil Description*. 4th ed. Rome: Food; Agriculture Organization of the United Nations.

Yan, Fapeng, Wei Shangguan, Jing Zhang, and Bifeng Hu. 2018. "Depth-to-Bedrock Map of

China at a Spatial Resolution of 100 Meters." https://doi.org/10.5194/essd-2018-103.
Zanaga, Daniele, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, et al. 2021. "ESA WorldCover 10 m 2020 V100." Zenodo. https://doi.org/10.5281/ZENODO.5571936.
Zolfaghari Nia, Masoud, Mostafa Moradi, Gholamhosein Moradi, and Ruhollah Taghizadeh-Mehrjardi. 2022. "Machine Learning Models for Prediction of Soil Properties in the Riparian Forests." *Land* 12 (1): 32. https://doi.org/10.3390/land12010032.