https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

SOIL Discussions Open Access EGU

Filling a key gap: a soil infrared library for central Africa

Laura Summerauer[1], Philipp Baumann[1], Leonardo Ramirez-Lopez[2], Matti Barthel[1], Marijn Bauters[3,4], Benjamin Bukombe[5], Mario Reichenbach[5], Pascal Boeckx[3], Elizabeth Kearsley[4], Kristof Van Oost[6], Bernard Vanlauwe[7], Dieudonné Chiragaga[7], Aimé Bisimwa Heri-Kazi[6], Pieter Moonen[8], Andrew Sila[9], Keith Shepherd[9], Basile Bazirake Mujinya[10], Eric Van Ranst[11], Geert Baert[3], Sebastian Doetterl[1,5], and Johan Six[1]

[1]Department of Environmental Systems Science, ETH Zurich, Switzerland
[2]NIR Data Analytics, BUCHI, Labortechnik AG, Flawil, Switzerland
[3]Department of Green Chemistry and Technology, Ghent University, Ghent, Belgium
[4]Department of Environment, Ghent University, Ghent, Belgium
[5]Institute of Geography, University of Augsburg, Germany
[6]Earth and Life Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[7]International Institute of Tropical Agriculture, Nairobi, Kenya and Bukavu, Democratic Republic of Congo
[8]Department of Earth and Environmental Sciences, KU Leuven, Leuven, Belgium
[9]World Agroforestry Centre, Nairobi, Kenya
[10]Department of General Agricultural Sciences, University of Lubumbashi, Lubumbashi, Democratic Republic of Congo
[11]Department of Geology, Ghent University, Gent, Belgium

Correspondence: Laura Summerauer, Universitaetstrasse 16, 8092 Zurich, Switzerland (laura.summerauer@usys.ethz.ch)

Abstract. Information on soil properties is crucial for soil preservation, improving food security, and the provision of ecosystem services. Especially, for the African continent, spatially explicit information on soils and their ability to sustain these services is still scarce. To address data gaps, infrared spectroscopy has gained great success as a cost-effective solution to quantify soil properties in recent decades. Here, we present a mid-infrared soil spectral library (SSL) for central Africa (CSSL) that can predict key soil properties allowing for future soil estimates with a minimal need for expensive and time-consuming wet chemistry. Currently, our CSSL contains over 1,800 soils from ten distinct geo-climatic regions throughout the Congo Basin and wider African Great Lakes region. We selected six hold-out core regions from our SSL, augmented them with the continental AfSIS SSL, which does not cover central African soils. We present three levels of geographical extrapolation, deploying Memory-based learning (MBL) to accurately predict carbon (TC) and nitrogen (TN) contents in the selected regions. The Root Mean Square Error of the predictions ($RMSE_{pred}$) values were between 0.38–0.86 % and 0.04–0.17 % for TC and TN, respectively, when using the AfSIS SSL only to predict the six regions. Prediction accuracy could be improved for four out of six regions when adding central African soils to the AfSIS SSL. This reduction of extrapolation resulted in $RMSE_{pred}$ ranges of 0.41–0.89 % for TC and 0.03–0.12 % for TN. In general, MBL leveraged spectral similarity and thereby predicted the soils in each of the six regions accurately; the effect of avoiding geographical extrapolation and forcing regional samples in the local neighborhood (MBL-spiking) was small. We conclude that our CSSL adds valuable soil diversity that can improve predictions for the regions compared to using the continental scale AfSIS SSL alone; thus, analyses of other soils in central Africa will be able to profit from a more diverse spectral feature space. Given these promising results, the library comprises an important tool to facilitate economical soil analyses and predict soil properties in an understudied yet critical region of Africa.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Our SSL is openly available for application and for enlargement with more spectral and reference data to further improve soil diagnostic accuracy and cost-effectiveness.

1 Introduction

Soil health is critical to crop nutrition, agricultural production, food security, erosion prevention, and climate change mitigation via carbon (C) storage. Global climate change and soil degradation by deforestation and soil mismanagement critically threaten these ecosystem services (Birgé et al., 2016). In particular, the humid tropics are a front line for these anthropogenic impacts. For example, increasing temperatures and accelerating deforestation in the humid tropics are estimated to enhance greenhouse gas emissions (Cox et al., 2013; Don et al., 2011), but also to significantly reduce soil functions and ecosystem services such as soil fertility, water storage and filtration capabilities and erosion protection (Veldkamp et al., 2020). Despite the expected severity of these impacts, our understanding of the effects in the humid tropics are limited by sparse data and uneven distribution of low-latitude research.

Within the tropics, both the future impacts and data gaps are most severe in the Congo Basin, which contains the second largest tropical forest ecosystem on Earth and represents a considerable reservoir of soil C (FAO and ITTO, 2011). Forest loss in central Africa is mainly driven by smallholder farmers practicing shifting cultivation (Tyukavina et al., 2018; Curtis et al., 2018). Thus, the projected drastic population growth in the coming decades (Vollset et al., 2020) will likely contribute to further agricultural conversion (UNESCO World Heritage Centre, 2010).

In the wake of these current and future impacts, more spatially explicit soil information is urgently needed in many research fields. In recent decades, improvements have been made carrying out soil surveys and creating soil databases and maps for central Africa (Goyens et al., 2007), for Rwanda (Imerzoukene and Van Ranst, 2002) and for the DRC (Baert et al., 2013). Unfortunately, accessibility to such data is limited and gaps are still large in central Africa due to the high cost of specialized equipment and chemicals for analyses (Van Ranst et al., 2010).

Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectroscopy has gained attention as a cost effective and fast method for soil analyses (e.g., Nocita et al., 2015). Many soil minerals, as well as functional groups of soil organic matter, show distinct energy absorption features in the infrared (IR) region of the electromagnetic spectrum. These relationships can be empirically modelled to quantify soil properties relevant for soil quality, such as C, N and other crop nutrients (e.g., Janik et al., 1998; Soriano-Disla et al., 2014). Due to its simple handling, quick measurements, low costs, and minimal need for chemical consumables, infrared spectroscopy is an important tool for soil analyses. Especially in developing countries, where practices are often hampered by the prohibitive costs of conventional soil analyses, IR spectroscopy has great potential (Shepherd and Walsh, 2007; Ramirez-Lopez et al., 2019).

Despite the abundance of literature on the calibration of quantitative models of soil properties using both mid-infrared (MIR) and near-infrared (NIR) data, there is still a lack of simple and efficient modeling strategies that could bring soil spectral libraries (SSLs) to an operational level. Such workflows of spectral soil estimation can thereby target regional, field or plot-scale estimation of soil properties, and should drastically reduce the amount of chemical reference analyses required on new

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

(local) soils in order to be efficient. In particular, soil spectral libraries are useful for inferring soil properties when positive predictive transfer occurs; it applies when the SSL (compositionally) related to the local soil prediction set improves the predictive accuracies compared to a calibration using a limited set of samples with local reference analyses available (Padarian et al., 2019; Lin et al., 2013).

Partial Least Squares (PLS) regression is the most widely used tool to calibrate models that translate spectral data into meaningful chemical and or physical information. The method is especially useful in a non-complex context, where the relationships between spectra and response variables is essentially linear (e.g. spectral models developed for a small field where sol forming factors are relatively constant). One of the main aims of establishing large-scale SSLs is to minimize the need for future wet chemical analyses (e.g., Nocita et al., 2014; Stevens et al., 2013; Shi et al., 2014; Viscarra Rossel et al., 2016). However, these libraries often span vast geographical areas that include different soil types and climate zones, which comprise complex soil organic C forms and mineral compositions. Due to this heterogeneity, predictions rendered by traditional linear regression models (such PLS) are often unfeasible for proper soil assessments at small-scale studies due to their high levels of uncertainty. To overcome this issue, new methods have recently been proposed, including local data-driven resampling approaches (RS-LOCAL, Lobsey et al. (2017)) or memory based learning (MBL, Ramirez-Lopez et al. (2013)). For each new spectral observation, MBL searches a subset of spectrally similar samples in a reference spectral library, which are then used to fit a custom predictive model for the new observation. This method has shown promising results when applied to extremely complex spectral libraries such as the MIR library of the United States (Dangal et al., 2019) and in one developed for the European continent (Tsakiridis et al., 2019). Spiking of libraries with samples from the target site has also shown to improve prediction accuracy (Guerrero et al., 2010; Seidel et al., 2019; Barthès et al., 2020; Wetterlind and Stenberg, 2010). So far, spectral libraries have mainly been used for predictions of soil samples originating from the geographical domain. Studies have shown that subsetting large-scale libraries for new spectra by their geographical zones can result in good prediction accuracy (Shi et al., 2015; Nocita et al., 2014). These geographical restrictions could allow for extrapolation to new areas that contain similar soils.

Here, we present the first SSL for central Africa (CSSL) along with an improved memory-based learning algorithm to accurately predict soil chemical properties. This study had two primary goals: (1) to fill the critical data gap for central Africa and complement an existing continental library, and (2) to establish a workflow to accurately predict six selected core regions of the CSSL and demonstrate how new regions can be predicted in order to further enlarge the library. This effort represents an important first step towards fulfilling the need for spatially explicit and high-resolution soil data in an important yet understudied region in the humid tropics, promoting vital soil information that is critical to the future of the region.

2 Methods

2.1 Site descriptions

Soil samples were collected from past projects in the Congo Basin and along the central African Rift Valley. Table 1 gives an overview of corresponding references to the different sample sets and denotes the sampling layer used in this study. Site specific

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

characteristics, coordinate ranges, altitudes, average climate, and dominant soil types are summarized in Table 2. The sampling area covers a large geographic area of central Africa, from a latitude of 2.8 °N to -11.6 °N and a longitude of 12.9 °E to 30.4 °E. Annual precipitation ranges from about 1100 mm in Kongo Central to over 2000 mm in the tropical forest of Tshuapa. Mean annual temperature varies from 12.8 °C in the high altitudes of North Kivu to 25.5 °C in Équateur and Kongo Central (Fick and Hijmans, 2017). The study elevations range from nearly sea level in the very west to high altitudes of 2000 m.a.s.l. along the rift valley and 3000 m.a.s.l. on Nyiragongo volcano in North Kivu (Jarvis et al., 2008). Soil types are primarily Ferralsols, Acrisols, or Nitisols (Jones et al., 2013; WRB, 2006). The different regions contain multiple Köppen-Geiger climatic zones: The four regions located close to the equator (Équateur, Tshuapa, Tshopo, Kabarole) are classified as Af (tropical rainforest), while the north and west DRC is classified as Aw (tropical savannah). Eastern DRC and western Rwanda are classified as a mixture of climate zones Cfb (temperate, without dry season, warm summer), Csb (temperate, dry summer, warm summer), Aw (tropical savannah) and Cwb (temperate, dry winter, warm summer). The regions along the rift valley (South Kivu, North Kivu, Iburengerazuba, Kabarole) are partially also classified as Am (tropical monsoon). Finally, the south-east of the DRC is classified as Cwa (temperate, dry winter, hot summer) (Beck et al., 2018).

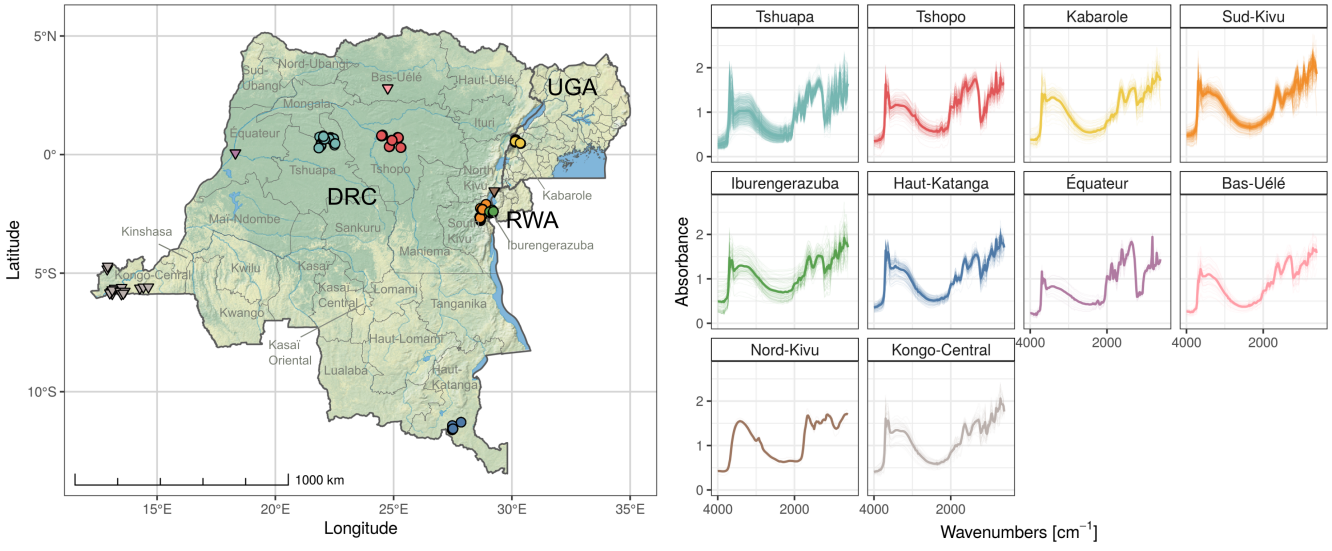

Figure 1. Locations and resampled spectra for the sampling regions (six selected core regions with a ○ symbol and the remaining four regions with a ▽ symbol). All samples are included in the archive of the spectral library for central Africa. For the Democratic Republic of Congo (DRC) and Rwanda (RWA), the regions correspond to provinces, for Uganda (UGA), the sampling region corresponds to a district (left). The average spectra of each region are shown (bold line) along with the individual sample spectra (transparent lines; right).

2.2 Laboratory soil analyses

All soil samples were dried prior to analysis. Total C (TC) and total N (TN) were analyzed via dry combustion on either a LECO 628 Elemental Analyzer (LECO Corporation, USA), on an ANCA-SL Automated Nitrogen Carbon Analyzer (SerCon,

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

SOIL Discussions Open Access EGU

Table 1. Soil sample archive used for the central African soil spectral library. The references show publications including the corresponding samples. The regions are provinces of the Democratic Republic of Congo (DRC) and Rwanda (RWA) and a district of Uganda (UGA).

References	Region	Soil depth (cm)
Bauters et al. (2015, 2017)	Tshopo (DRC)	5–10
Gallarotti et al. (2021), Baumgartner et al. (2020)	Tshopo, South Kivu, Équateur (DRC)	0–15, 15–30, 0–10, 10–20, 30–40, 40–50, 50–60, 60–70, 70–80, 80–90, 90–100
Kearsley et al. (2013, 2017)	Tshopo (DRC)	0–10, 10–20, 20–30, 30–50, 50–100
Bauters et al. (2019a)	Tshopo, South Kivu (DRC)	0–5
Bauters et al. (2021), Moonen et al. (2019)	Tshopo (DRC)	0–5, 5–10, 10–20
Bauters et al. (2019b)	Tshuapa (DRC)	0–10, 10–20, 20–30, 30–50, 50–100
Minten (2017)	Tshuapa (DRC)	0–20
Summerauer (2017)	Tshuapa (DRC)	0–20, 20–50
Heri-Kazi (2020)	South Kivu (DRC)	0–20
Université catholique de Louvain	South Kivu, Haut-Katanga (DRC)	0–20, 20–30
Mujinya (2012); Mujinya et al. (2010, 2011, 2013, 2014)	Haut-Katanga (DRC)	Termite mound profiles
IITA/ICRAF	Bas-Uélé, South Kivu (DRC)	0–20, 20–40, 20–50
Baert et al. (2009); Baert (1995)	Lower Congo and Central (DRC)	0–123 soil pit
Doetterl et al. (2021b, a)	South Kivu (DRC), Iburengerazuba (RWA), Kabarole (UGA)	0–10, 30–40, 60–70, 90–100

Table 2. Number of samples, GPS coordinates, elevation, annual precipitation (AP), mean annual temperature (MAT), Koeppen-Geiger climate classifications and soil types for the sampled regions of the Democratic Republic of Congo, Rwanda and Uganda. Data were extracted for all coordinates from raster files: Climate data is sourced from Fick and Hijmans (2017), elevation from SRTM (90m resolution; Jarvis et al. (2008)), Köppen-Geiger climate classifications from Beck et al. (2018) and soil types from the *Soil Atlas of Africa* (Jones et al., 2013; WRB, 2006)

Region	n	Longitude (° E)	Latitude (° N)	Elevation (m)	MAT (°C)	AP (mm)	Köppen-Geiger	Soil types
Haut-Katanga	119	27.48–27.85	-11.61– -11.29	1197–1323	20.6	1223	Cwa	Rhodic/Haplic Ferralsols
South Kivu	369	28.64–28.91	-2.79– -2.1	1487–2310	17.6	1627	Cfb, Csb, Aw, Cwb	Umbric Ferralsols, Haplic Acrisols
Tshopo	315	24.48–25.32	0.29–0.83	380–506	24.9	1789	Af	Xanthic/Haplic Ferralsols
Tshuapa	738	21.84–22.53	0.28–0.8	385–578	24.7	2090	Af	Xanthic/Haplic Ferralsols
Iburengerazuba	107	29.05–29.22	-2.47– -2.34	1565–1939	17.6	1496	Csb, Aw, Cwb	Haplic/Umbric Acrisols
Kabarole	101	30.13–30.37	0.46–0.63	1271–1824	19.7	1360	Af, Cfb, Am	Haplic Phaeozems, Rhodic Nitisols, Albic Luvisols
Équateur	12	18.31	0.06	322	25.5	1685	Af	Eutric Ferralsols
Bas-Uélé	49	24.75	2.8	423	25.2	1641	Aw	Haplic Ferralsols
North Kivu	4	29.25–29.27	-1.55– -1.53	2276–3250	12.8	1834	Cfb	Umbric Silandic Andosols
Kongo-Central	40	12.89–14.63	-5.88– -4.71	30–470	25.5	1088	Aw	Ferralic Cambisols, Haplic Acrisols, Umbric Nitisols, Xanthic Ferralsols, Mollic Gleysols

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

UK), or on a Vario EL Cube CNS Element Analyzer (Elementar, Germany). In order to ensure data quality and facilitate the harmonisation of all TC and TN data, a subset of these samples were remeasured on the LECO (R^2 = 0.99 for TC and TN, results not shown). Additionally, soil pH, texture, and macro/micro nutrients (Al, Fe, Ca, Mg, Mn, Na, P and K) were analyzed for a subset of samples. The chemical and MIR prediction results for these soil characteristics are not presented in this manuscript but were carried out using the same methods and are available on our GitHub repository. The large majority of the soil samples originate from highly weathered and acidic soils and do not contain any carbonates. Therefore, TC contents correspond to total organic carbon contents. Only in a few samples from termite mounds in the subtropical Haut-Katanga province calcium carbonate has been detected and pH values are > 8 (Mujinya, 2012). Note, even if the proportion of samples with inorganic carbon was very low (5 %), the term TC will be used in the study.

2.3 MIR spectral libraries

Central African spectral library

All samples were finely ground using a ball mill and measured with a VERTEX70 Fourier Transform-IR (FT-IR) spectrometer with a High Throughput Screening Extension (HTS-XT) (Bruker Optics GmbH, Germany) in order to determine the MIR reflectance. A gold standard was used as a background material for all measured soils in order to normalize the sample spectra. Reflectance was transformed into absorbance (1/reflectance) prior to further processing and subsequent modeling. Two replicates per sample were filled into the cups of a 24-well plate and the surface was flattened without compression using a spatula. For each sample, 32 co-added internal measurements were averaged and corrected for CO_2 and H_2O using the OPUS spectrometer software (Bruker Optics GmbH, Ettingen, Germany).

AfSIS spectral library

We used a MIR SSL created by the World Agroforestry (ICRAF) centre. This SSL was created as part of the Africa Soil Information Service (AfSIS) in order to improve soil information and land management on the continental scale of Sub-Saharan Africa (Vågen et al., 2020). For this continental library (see Figure A1), reference values for TC and TN were obtained by using a ThermoQuest EA 1112 elemental analyzer. The MIR spectra of the samples were obtained by scanning them on a Tensor27 FT-IR spectrometer (Bruker Optics, Karlsruhe, Germany) with a high throughput screening extension. Four replicates per sample were measured and an average of 32-co-added scans were used for each sample (Sila et al., 2016).

2.4 Spectral resampling and pre-processing

All CSSL and AfSIS spectra were processed using the R packages 'simplerspec' (Baumann, 2020), 'prospectr' (Stevens and Ramirez-Lopez, 2020) and 'resemble' (Ramirez-Lopez, 2020) in the R statistical computing environment (R Core Team, 2020). Replicates of spectral measurements were mean aggregated to obtain one spectrum per sample. The spectra were then resampled to a resolution of 16 cm^{-1} and trimmed to the 4000–600 cm^{-1} spectral range.

As spectral pre-treatments have a marked impact on the performance of quantitative infrared models (Rinnan, 2014), the pre-processing procedure was specifically optimized for the MIR spectra of the central African samples. This procedure was

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

based on the PLS method, which was also known as projection to latent structures. This method has been traditionally used for regression analysis in infrared spectroscopy. However, it is also useful for projecting the spectral data onto a low-dimensional (and therefore less complex) subspace containing all the meaningful information of the original data. This projection model can be expressed as:

$$\mathrm{X} = \mathrm{T}\mathrm{P}' + \mathrm{E} \tag{1}$$

where X is the original spectral matrix of $n \times d$ dimensions, T is the PLS score matrix of $n \times l$ dimensions (where $l \leq \min(n, d)$) which contains the extracted variables, P is the matrix of loadings of $d \times p$ dimensions which captures the spectral variability across observations. E is an error term. For spectral data with high collinearity, the optimal l (or the number of PLS factors) is usually small, which means that the first few PLS factors are enough to properly represent the original variability of X. An important aspect of this type of projection is that it is obtained in such a way that the covariance between T and an external set of one or more variables is maximized. For a detailed description on PLS, see Wold et al. (2001). In PLS, P can be used on new spectral observations to project them onto the lower dimensional space:

$$\mathrm{T}_{new} = \mathrm{X}_{new}\mathrm{P}^{-1} \tag{2}$$

The spectral reconstruction error of the projection model can be then computed by back-transforming the matrix of scores to a spectral matrix and comparing it against the original spectral matrix as follows:

$$\mathrm{E}_{new} = \mathrm{X}_{new} - \mathrm{T}_{new}\mathrm{P}' \tag{3}$$

The above spectral reconstruction error concept was used to find an optimal combination of spectral pre-treatments. We defined a set of different pre-treatments $\{h_1, h_2, ..., h_z\}$ where $h_i(\mathrm{X})$ represents one pre-treatment or a sequence of pre-treatments (with unique parameter values) to be applied on the spectral data. A projection model was built with the AfSIS spectra (using TC and TN as external variables) for each combination of spectral pre-treatments:

$$h_i(\mathrm{X}_{AfSIS}) = \mathrm{T}(i)\mathrm{P}'(i) \tag{4}$$

this model was used on the CSSL pre-treated spectra and the reconstruction error (E_{CSSL}) was computed as follows:

$$\mathrm{E}_{CSSL} = h_i(\mathrm{X}_{CSSL}) - [h_i(\mathrm{X}_{CSSL})\mathrm{P}^{-1}\mathrm{T}_{CSSL}\mathrm{P}'(i)] \tag{5}$$

The final reconstruction error (re) is computed as the root mean squared of the elements in E_{CSSL}:

$$re(i) = \frac{1}{m\,d}\sum_{j=1}^{m}\sum_{k=1}^{d} e_{jk} \tag{6}$$

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

where m is the number of samples in the CSSL. To allow for comparisons across the reconstruction errors obtained for the different pre-treatments, the $re(i)$ standardized as follows:

$$sre(i) = \frac{re(i)}{max(h_i(\mathrm{X}_{CSSL})) - min(h_i(\mathrm{X}_{CSSL}))} \tag{7}$$

The pre-treatments tested included different combinations of standard normal variate, multiplicative scatter correction, spectral detrend, first and second derivatives (with different window sizes).

The aim behind our reconstruction error approach was to identify a sequence of pre-processing steps that return spectral matrices which can be properly represented by a PLS model. In this respect, we assumed that a proper representation of the spectral data by a global PLS projection model might also be appropriate for local PLS models which are at the core of the predictive methods presented in the next sections.

Minimal spectral reconstuction error was achieved with a Savitzky-Golay filter combined with a second derivative using a second order polynomial approximation with a window size of 17 cm^{-1}, resulting in a final resolution of 272 cm^{-1} (resampling resolution of 16 cm^{-1} x window size of 17 cm^{-1}) (Savitzky and Golay, 1964), and a subsequent multiplicative scatter correction; this optimized pre-treatment was used for MBL.

2.5 Modeling scenarios

Here we describe the method we used to assess the performance of MBL for predicting TC and TN for six distinct regions at different scenarios of regional soil extrapolation. Figure 2 gives an overview of the modeling strategies. Three specific modeling strategies were tested on the selected regional sets which we call prediction sets (see subsubsection 2.5.2). With the regional analysis we demonstrate how predictions of soil properties within new sites from distinct regions—which are compositionally less variable than the available SSLs—might perform and profit from knowledge present in the AfSIS SSL. It also demonstrates the added value of our new CSSL in addition to the AfSIS SSL alone. The aim of the modeling scenarios were twofold: 1) to minimize the costs and time for traditional methods by optimizing the transfer of stored spectral information to the new region of interest 2) test different levels of geographical extrapolations to demonstrate how accurate predictions are for new regions, when no local samples area available.

2.5.1 Modeling and prediction data

We used two main data sources and subsets as follows.

1. The AfSIS data set (A): Continental large-scale SSL including 1902 soil samples with MIR spectra and corresponding reference data, originating from Sub-Saharan Africa (Figure A1).

2. The central African data set: From our CSSL, six regions were identified which contained at least 80 samples. These six regions were Haut-Katanga, South Kivu, Tshopo, and Tshuapa provinces of the DRC, Iburengerazuba, also known as the Western Province of Rwanda, and Kabarole, a district in western Uganda (Figure 1). To test how well these regions

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

could be predicted by the SSLs, they were defined as hold-out core regions. These six regions comprise together 1578 soil samples with MIR spectra and corresponding reference data. The sets were defined as following:

(a) Central African set ($\mathcal{C}$): A set formed from the union of the sets of the six different regions in central Africa (n = 1442 and 1458 for TC and TN, respectively, after the removal of the 6 x 20 spiking samples for each region; see below). This set can be written as

$$\mathcal{C} = \bigcup_{1}^{6} \mathrm{G}_i$$

where G_i represents the data of the ith region.

(b) Six regional sets G_i, (n = 80–718 after removal of 20 spiking samples for every set; see below).

(c) Six regional spiking sets (K_i): for each complete regional set, 20 samples were selected using the k-means sampling algorithm (Næs, 1987; Stevens and Ramirez-Lopez, 2020).

2.5.2 Modeling strategies

Three different scenarios were compared which are related to the scale of the geographical extrapolation:

– Strategy 1: MBL predictions for the C set were computed from A. This scenario represents an extreme case of extrapolation (from the geographical perspective) since no samples from the entire central African area are present in the AfSIS set Figure A1, which is the only data used to build the predictive models. In addition, the MIR data from the central African (C) set originates from a different spectrometer type than the one used for scanning the AfSIS samples.

– Strategy 2: Predictions for every G_i are computed by using MBL models built from the pooled AfSIS data A together with the data from the remaining five regions $\mathcal{C}_i$, i.e. $\mathrm{A} \cup \mathcal{C}_i$, where

$$\mathcal{C}_i = \bigcup_{\substack{j=1 \\ j \notin i}}^{6} \mathrm{G}_j$$

Although in this case there is also extrapolation (from the geographical perspective), is not as extreme as in Strategy 1.

– Strategy 3: Strategy 2 was repeated, but in this case, extrapolation was avoided by using the spiking samples from the same geographical region; Each regional set G_i was predicted by the pooled AfSIS data, the data of the remaining regions and the respective spiking set, i.e. $\mathrm{A} \cup \mathcal{C}_i \cup \mathrm{K}_i$.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

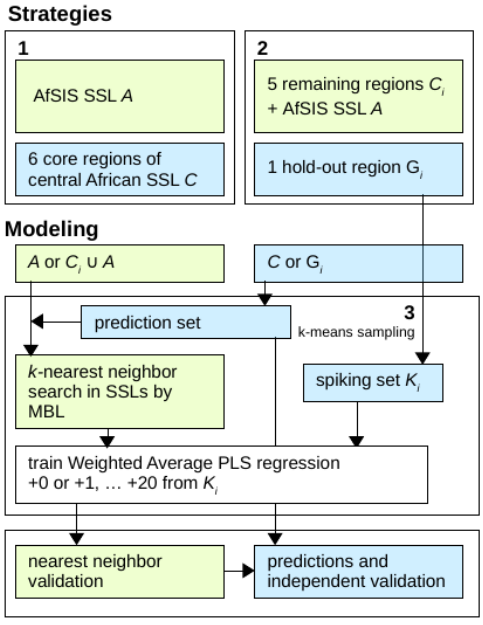

Figure 2. Flow chart representing the work flow of the modeling strategies with three different levels of geographic extrapolation. Strategy 1: Predictions of the six selected core-regions using only the AfSIS soil spectral library (SSL; without any central African soils), Strategy 2: Predicting each hold-out region by the pooled remaining five regions (adding closer samples) together with the AfSIS SSL and Strategy 3: avoiding extrapolation by adding 1 to 20 spiking samples to the models regional models of Strategy 2.

2.6 Predictive modeling

We used Memory-based learning (MBL) as our predictive modeling approach. MBL describes a family of (non-linear) machine learning methods designed to handle complex spectral datasets (Ramirez-Lopez et al., 2013). In the chemometrics literature, MBL is also known as local modeling. This type of learning method does not attempt to fit a general (global) predictive function using all available data. Instead, a new and unique function ($\hat{f}_i$) is built on-demand, every time a new prediction for a given response variable is required. This new function is built using only a subset of relevant observations from a reference set that are queried through k-nearest neighbour search. The MBL method implemented for this study uses a spectral nearest neighbour search based on a moving window correlation dissimilarity. To measure the dissimilarity (d) between two spectra (x_i and x_j), the following equation is used:

$$d(\mathrm{x}_i, \mathrm{x}_j; w) = \frac{1}{2w} \sum_{k=1}^{p-w} 1 - \rho(\mathrm{x}_{i,\{k:k+w\}}, \mathrm{x}_{j,\{k:k+w\}})$$

where ρ represents the Pearson's correlation function and w the window size. After nearest neighbor retrieval, our MBL method fits a local model using the Weighted Average Partial Least Squares (WA-PLS) regression algorithm proposed by

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Shenk et al. (1997). In WA-PLS, the final prediction is a weighted average of multiple predictions generated by PLS models built from different PLS factors. The weight for each component is calculated as follows:

$$w_j = \frac{1}{s_{1:j} \times g_j}$$

where $s_{1:j}$ is the root mean square of the spectral residuals of the new observation when a total of j pls components are used and g_j is the root mean square of the regression coefficients corresponding to the jth PLS component (see Shenk et al. (1997) for more details).

The number of neighbors to retrieve was optimized using the nearest neighbor (NN) cross-validation (Ramirez-Lopez et al., 2013). Using this method, for each observation to be predicted, its nearest neighbor was excluded from the group of neighbors and then a WA-PLS model is fitted using the remaining ones. This model is then used to predict the value of the response variable of the nearest observation. These predicted values are finally cross-validated with the actual values (see Ramirez-Lopez et al. (2013) for additional details). To avoid overfitting, the region was used as a grouping factor, which was a 'sentinel site' for the AfSIS library and a 'province' or 'district' of the particular country for the CSSL. Samples from the same sampling region were consequently assigned to the same fold when dividing them into hold-out and validation sets. Neighborhood sizes varying from 150 to 500 neighbors in increments of 10 were tested. The best model and the optimal number of neighbours were determined by the minimal RMSE (Equation 8) of the nearest neighbour validation, where $\boldsymbol{n}$ is the number of neighbours used for the model, $\boldsymbol{y_i}$ is the measured value of the hold-out neighbor, and $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ is the value predicted by the remaining neighbours.

Subsequently, 1 to 20 spiking samples were added from the target region and forced into the neighbourhood of every observation and thus used in the predictive models, independent from their distances to the validation set. The stepwise spiking was applied to test the effect of spiking in general and to find the smallest number of samples required for satisfying model performances.

2.7 Model validation and prediction accuracy

For model validation, the RMSE statistics of the nearest neighbor validation described in the previous section were used. Prediction accuracy of the seven sets (the combined 6 regions $\mathcal{C}$ and the six individual regional sets G_i; see above), which is the so-called independent or external validation, was also calculated using RMSE (Equation 8), where in this case $\boldsymbol{y_i}$ is the actual measured reference value and $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ the prediction of the final model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{8}$$

Model validation and prediction performance were additionally evaluated using the Ratio of Performance to the InterQuartile distance (RPIQ) as suggested by Bellon-Maurel et al. (2010). The interquartile range of the observed reference data is divided by the RMSE of the nearest neighbor validation or by the RMSE of the prediction ($\mathrm{RMSE}_{\mathrm{pred}}$), respectively. The RPIQ is useful because it does not make any assumptions about the distribution of the reference data.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

3 Results

The sample archive of the CSSL covered a wide range of TC and TN contents (Table 3). Seven of the 10 regions (Bas-Uélé, Equateur, Haut-Katanga, Kabarole, Kongo Central, Tshopo, Tshuapa) had mean TC and TN of 1.09–1.77 % and 0.07–0.17 %, respectively. Maximum TC and TN values for these seven regions were 5.67 % and 0.51 %, respectively. The three regions of South Kivu and North Kivu in Eastern DRC and Iburengerazuba in Western Rwanda had significantly higher TC and TN contents, with TC and TN means of 2.63–31.02 % and 0.17–1.93 %, respectively. Volcanic soils from North Kivu had the highest TC and TN contents. The AfSIS SSL had generally lower TC and TN means of 1.2 % and 0.08 %, respectively.

Table 3. Summary of the reference data for total carbon (TC) and total nitrogen(TN) of the two soil spectral libraries (SSL). The central Africal soil spectral library (CSSL) is divided into the main regions, with the six core regions (Haut-Katanga, South Kivu, Tshopo, Tshuapa, Iburengerazuba, Kabarole) selected for the spectral analyses in in the first six rows. The remaining four regions (Équateur, Bas-Uélé, North Kivu, Kongo-Central) are presented, but were not further analysed in this study.

SSL	Covered region	TC [%]					TN [%]				
		n	Mean	Median	Min	Max	n	Mean	Median	Min	Max
CSSL	Haut-Katanga (DRC)	119	1.13	0.97	0.13	3.47	119	0.11	0.10	0.04	0.29
	South Kivu (DRC)	367	3.54	2.93	0.60	18.21	368	0.31	0.24	0.07	1.50
	Tshopo (DRC)	134	1.38	1.24	0.40	5.67	149	0.10	0.09	0.02	0.45
	Tshuapa (DRC)	738	1.26	1.16	0.37	4.74	738	0.10	0.09	0.02	0.39
	Iburengerazuba (RWA)	104	2.63	2.27	0.15	9.38	104	0.17	0.15	0.01	0.55
	Kabarole (UGA)	100	1.77	1.19	0.08	5.38	100	0.17	0.12	0.01	0.51
	Équateur (DRC)	12	1.32	1.02	0.12	5.05	12	0.07	0.07	0.02	0.14
	Bas-Uélé (DRC)	49	1.09	0.96	0.27	2.84	49	0.09	0.07	0.02	0.22
	North Kivu (DRC)	4	31.02	31.97	18.96	41.17	4	1.93	1.80	1.20	2.92
	Kongo-Central (DRC)	40	1.68	1.24	0.34	5.50	40	0.14	0.12	0.04	0.49
AfSIS SSL	Sub-Saharan Africa	1902	1.24	0.78	0.08	11.29	1902	0.08	0.05	0.00	0.66

3.1 Principal components and spectral variability in the two libraries

A principal component analysis (PCA) was conducted on the pre-processed spectra of both libraries. The first three principal components account for 70 % of the spectral variability. These components indicate that bulk of CSSL samples are within the spectral domains of the AfSIS SSL as their PCA scores overlap (Figure 3). The overlapping, however, is less evident for the spectra of the South Kivu region and, to a lower extent, for the samples of the Iburengerazuba region, which suggests that the type of soils in these regions might not be very well represented by the samples in the AfSIS SSL. Note that these two regions are geographically close (Figure 1).

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

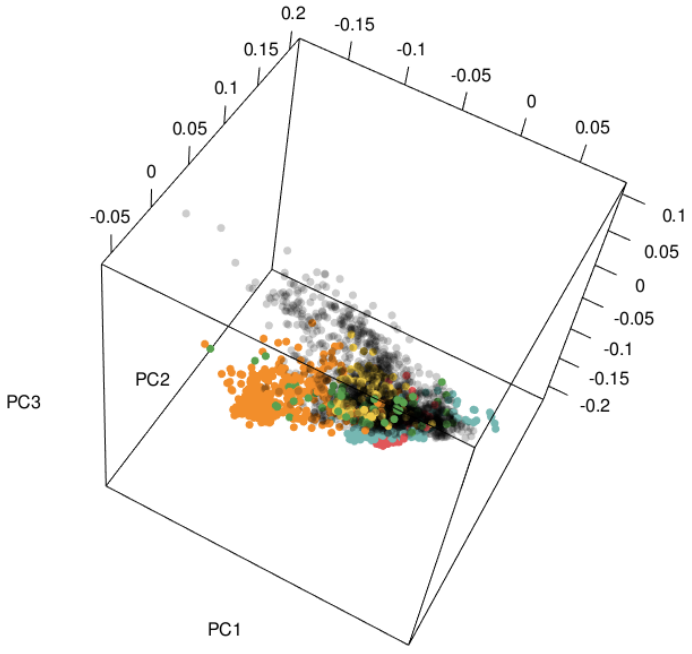

Figure 3. Score plot of the the first three pricipal components of the pre-processed MIR spectra. The six central African hold-out regions (colours) and the AfSIS SSL (black). The regions South Kivu (orange) and Iburengerazuba (green) are covering an area, which is neither represented by the AfSIS SSL, nor by the remaining four central African regions.

3.2 Predictive performance of the three strategies

265 The prediction results for the three strategies are presented in Table 4 and Figure 4. In general, MBL retrieved good predictive results for all the strategies and for both TC and TN. As expected, the predictions for the South Kivu and Iburengerazuba regions showed the lowest accuracy levels. This was expected as the principal component analysis indicate that the sols of these regions might not be properly represented by the AfSIS library.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Table 4. Statistics of the independent validations of the predictions of total carbon and total nitrogen for each region and three strategies. Strategy 1: Predictions of the combined six regions by the AfSIS soil spectral library (SSL), Strategy 2: Predictions of the individual regions by the remaining five regions together with the AfSIS SSL, Strategy 3: Spiking six regional models from Strategy 2 with 20 samples from each target area.

Strategy	Region	Total carbon [%]					Total nitrogen [%]				
		n_{pred}	$RMSE_{pred}$	$R^2{}_{pred}$	ME_{pred}	$RPIQ_{pred}$	n_{pred}	$RMSE_{pred}$	$R^2{}_{pred}$	ME_{pred}	$RPIQ_{pred}$
Strategy 1	Haut-Katanga	99	0.60	0.79	0.50	1.62	99	0.08	0.31	0.07	0.59
	South Kivu	347	0.86	0.94	0.35	2.43	348	0.17	0.85	0.13	1.10
	Tshopo	114	0.73	0.47	0.26	0.96	129	0.05	0.52	0.03	0.93
	Tshuapa	718	0.38	0.71	0.21	1.84	718	0.04	0.68	0.03	1.37
	Iburengerazuba	84	0.87	0.84	0.45	2.60	84	0.08	0.81	0.06	2.13
	Kabarole	80	0.57	0.86	0.11	3.95	80	0.07	0.84	0.05	2.86
Strategy 2	Haut-Katanga	99	0.42	0.72	0.18	2.30	99	0.03	0.59	0.00	1.50
	South Kivu	347	0.89	0.95	0.47	2.36	348	0.12	0.89	0.07	1.55
	Tshopo	114	0.54	0.64	0.03	1.31	129	0.03	0.72	0.01	1.49
	Tshuapa	718	0.41	0.78	0.22	1.71	718	0.03	0.77	0.01	1.88
	Iburengerazuba	84	0.80	0.86	0.27	2.84	84	0.05	0.82	0.00	3.21
	Kabarole	80	0.86	0.83	0.43	2.65	80	0.06	0.86	0.04	2.90
Strategy 3	Haut-Katanga	99	0.36	0.80	0.14	2.72	99	0.03	0.71	0.01	1.87
	South Kivu	347	0.73	0.95	0.15	2.86	348	0.09	0.89	0.03	2.05
	Tshopo	114	0.49	0.69	0.01	1.43	129	0.03	0.75	0.00	1.62
	Tshuapa	718	0.32	0.80	0.09	2.22	718	0.02	0.79	0.00	2.25
	Iburengerazuba	84	0.63	0.91	0.11	3.57	84	0.04	0.91	0.00	4.45
	Kabarole	80	0.41	0.94	0.17	5.48	80	0.04	0.91	0.02	4.27

3.2.1 Strategy 1: Predicted central African soils by AfSIS SSL

The prediction performance for TC and TN for the six regions of central Africa ($\mathcal{C}$) are characterized by errors ($RMSE_{pred}$) ranging from 0.38–0.87 % and 0.04–0.17 % respectively. The best prediction accuracies were achieved for the regions Tshuapa, Kabarole, Haut-Katanga and Tshopo. For three of these regions with low $RMSE_{pred}$ (Tshopo, Haut-Katanga and Tshuapa), the goodness of fit was less precise than for the other regions with $R^2{}_{pred}$ of 0.47–0.79 and 0.31–0.68 for TC and TN, respectively and $RPIQ_{pred}$ of 0.96–1.84 for TC and 0.59–1.36 for TN. For South Kivu, samples with high TC and TN contents (> 10 % TC and > 0.5 % TN) are deviating from the 1:1 line. Moreover, TC predictions in three regions (Haut-Katanga, Tshopp, Tshuapa and Iburengerazuba) as well as TN predictions in all six regions showed a clear underestimation trend (Figure 4). This might be caused by one or the combination of the two following effects: *i)* spectral offset and/or multiplicative effects in the spectra

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

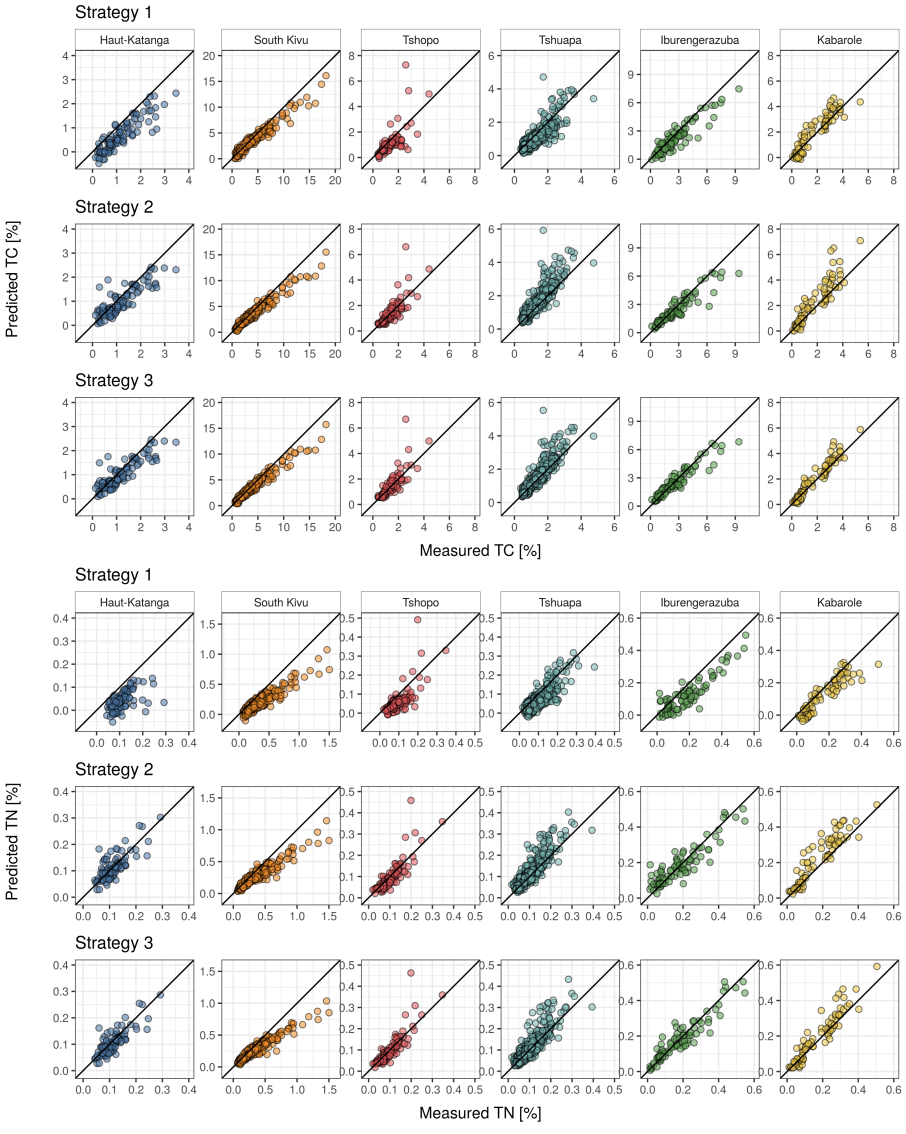

Figure 4. Predicted vs. measured total carbon (TC) and total nitrogen (TN) for soil samples of six hold-out regions using Memory-based learning; Strategy 1: predicting the six regions together by the AfSIS SSL, Strategy 2: Predicting each individual hold-out region by the remaining five regions together with the AfSIS SSL and Strategy 3: Spiking each model of Strategy 2 with local samples from the target regions. A 1:1 line is indicated as a visual aid.

(due to instrument differences) that might not have been completely accounted by the pre-pocessing methods *ii)* differences between the conventional laboratory analyses used to obtain the reference property values.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

3.2.2 Strategy 2: Regional predictions by soil spectral libraries

The predictive performance in this strategy exhibited errors ($RMSE_{pred}$) ranging between 0.41–0.89 % and 0.03–0.12 % for TC and TN respectively (Table 4). Similarly to strategy 1, the most accurate predictions were obtained for the regions Haut-Katanga, Tshopo and Tshuapa, where the errors ($RMSE_{pred}$) were below or equal to 0.54 % and 0.03 % for TC and TN, respectively. In comparison to strategy 1, the $RMSE_{pred}$ for Haut-Katanga and Tshopo regions were reduced by 0.2 %, while they were about the same for Tshuapa, South Kivu and Iburengerazuba. Kabarole was the only region, where the $RMSE_{pred}$ increased in strategy 2 compared to strategy 1 (Figure 2, Table 4). When compared to strategy 1, the TN prediction errors were consistently lower. This might be due to the inclusion of CSSL samples in the training set for this strategy. By doing so, variance coming from instrument and reference laboratory differences is then discarded from the local models. The R^2_{pred} of the TC predictions indicate that the precision of such models was, in general, equal or slightly better for the strategy 2 than for the strategy 1. Also the $RPIQ_{values}$ for the TC predictions tended to be the same as in strategy 1 or slightly higher, except for region Kabarole where $RPIQ_{pred}$ was reduced from 3.95 to 2.65 for strategy 1 and strategy 2, respectively. For TN, with the exception of South Kivu, all regional predictions resulted in better regression fits than when using the AfSIS SSL only for predictions, which is demonstrated by the higher R^2_{pred} and $RPIQ_{pred}$ values. R^2_{pred} and $RPIQ_{pred}$ for TN had a range of 0.59–0.89 and 1.50–3.21, respectively. In Figure 4, the improved prediction accuracy and better fits are visible especially for region Haut-Katanga. Also for the other five regions, the underestimation of TC and TN contents was reduced or even removed compared to strategy 1.

3.2.3 Strategy 3: Spiking of the regional models

Spiking the regional models with up to 20 local samples K_i (Figure 2) consistently returned the lowest prediction errors for all the regions (Table 4, Figure 5). Especially for the Ugandan region Kabarole the spiking effect was markedly large and $RMSE_{pred}$ values were reduced from 0.60 % to 0.36 % and 0.06 % to 0.04 % for TC and TN, respectively. The $RMSE_{pred}$ values for three regions Haut-Katanga, Tshopoa and Tshuapa were smaller compared to strategy 2, but the differences were relatively small (< 0.1 % for TC and < 0.01 for TN). With 20 spiking samples, $RMSE_{pred}$ for TC and TN contents for South Kivu could be reduced from 0.89 % to a minimal $RMSE_{pred}$ 0.73 % TC and from 0.12 % to a minimal $RMSE_{pred}$ of 0.09 % TN, respectively. The prediction error ($RMSE_{pred}$ for TC in the Iburengerazuba region could be reduced, but as in South Kivu, it remained relaviely large (> 0.6 %). Comparing the effect of 20 spiking samples with the previous strategies 1 and 2, the predictions could better be fitted to the measured values (higher R^2_{pred} and $RPIQ_{pred}$ values).

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

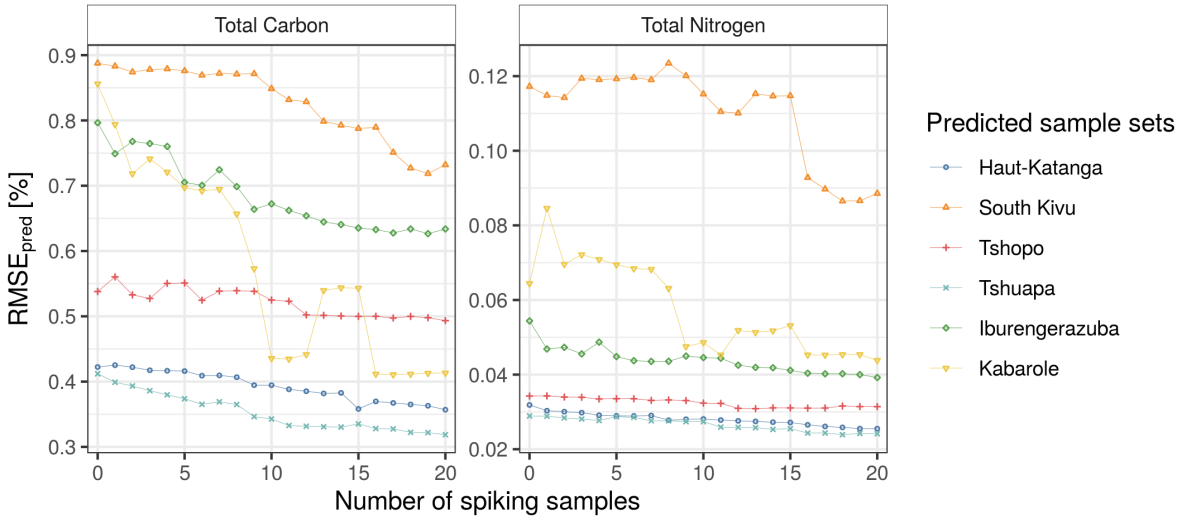

Figure 5. Root Mean Square Error of predicted total carbon (left) and total nitrogen (right; $RMSE_{pred}$) for the six hold-out regions of central Africa built from pooled AfSIS data together with the five remaining regions and the spiking samples. The 20 spiking samples were selected from each particular target area and stepwise added to the predictive models. Stepwise addition was done in order to find the lowest number of spiking samples that reduces the prediction accuracy to a satisfactory tolerance level.

4 Discussion

4.1 Using soil spectral libraries in geographically different domains

We showed that TC and TN in six regions of our CSSL can be accurately predicted, leveraging existing SSLs informed by soils from completely different geographical areas using MBL methods (Table 4, Figure 4). The advantage of using MBL is that it finds spectrally similar observations for every new observation to fit specific models. The spectral similarity is in fact reflecting the similarity between observations) in terms of soil composition which information is largely contained in the MIR features. This means that the predictive success of MBL models largely depends on the quality of the spectra dissimilarity methods used to find the spectral neighbors. In other words, MBL can be described as a method driven by compositional similarity search.

For central African soils together $\mathcal{C}$ by the AfSIS SSL (A) and for each hold-out region G_i predicted by the pooled AfSIS SSL together with the remaining five hold-out regions ($\mathrm{A} \cup \mathcal{C}_i$; Figure 2), MBL models were able to find similar samples and could accurately model and predict a new set without any additional local calibration samples. The improved prediction accuracy (lower $RMSE_{pred}$) when reducing extrapolation (strategy 2) can be explained by the addition of soils to the library that are more similar to the hold-out region. The AfSIS SSL is missing data for most of central Africa (Figure A1); for example, none of the tropical forest soils with high contents of organic carbon with distinctive mineral-organic composition are covered by this large-scale SSL. This, naturally, impacts the generalization ability of any predictive model or modeling strategy. Two regions (South Kivu and Iburenerazuba) show large variability in TC and TN contents (Table 3). Both sites

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

contain samples from both tropical forests and agricultural fields, from diverse altitudes (Table 2) and parent materials and have therefore transformed under a variety of environmental conditions. We conclude that the particularly high soil diversity in these two regions in terms of soil biogeochemical properties introduces additional complexity in the soil spectral prediction workflow. To improve predictions for these diverse regions, more data particularly with high TC and TN values are needed for calibrating the CSSL, and ultimately deliver better regional estimates using local methods (i.e., memory-based learning). High diversity in organic compounds and their stabilization in soils (i.e. organo-mineral association, complexation, aggregation) can introduce non-linear relationships that are difficult to predict with linear calibration models (i.e., memory-based learning in combination with PLS regression). Similarly high RMSEs have been shown in other studies for samples with organic C higher than 15 % (Nocita et al., 2014). As in our study, these high errors were attributed to low sample numbers with high organic C contents. The creation of subsets from large spectral libraries via spectral similarities has been shown to be effective to train calibration models (e.g., Wetterlind and Stenberg, 2010; Clairotte et al., 2016; Tziolas et al., 2019; Dangal et al., 2019; Sanderman et al., 2020). Hence, in order to reduce uncertainties for regions in central Africa that are diverse in terms of soil chemical composition, in particular for the Eastern Congo Basin, there is an urgent need for filling the existing gaps in the continental library by gathering more data on the ground.

4.2 Effect of spiking with local samples on prediction performance

The effect of spiking of the calibration models with local target samples was smaller than expected (Figure 5). Although spiking could reduce $RMSE_{pred}$ somewhat for two regions (Iburengerazuba, South Kivu and Kabarole, Table 4), the effect was rather small for the remaining regions. Regions that occupied the same score space of the first two principal components as the corresponding other regions and the AfSIS SSL (Figure 3) showed only a minimal effect from spiking (Figure 1). This is especially true for the Tshuapa, Tshopo and Haut-Katanga regions. In these regions, similar spectra were apparently already available and the MBL found the required neighbours to build accurate models and predict TC and TN. For South Kivu and Iburengerazuba, the predictions could not be improved by adding the other regions to the AfSIS SSL, but spiking with samples from the target area could slightly improve their results. However, the prediction error ($RMSE_{pred}$ remained relatively high (Figure 5). On one hand, no other region and also not the AfSIS SSL cover the same score space as these two regions and on the other hand, the variability of soil properties within these two regions is large which also minimizes the effect of spiking. Even though spiking is described as particularly effective in improving performance of small sized models (Guerrero et al., 2010), spiking, in our study, did not have as strong of an effect as reported by earlier studies (e.g., Guerrero et al., 2014; Seidel et al., 2019; Barthès et al., 2020; Wetterlind and Stenberg, 2010).

4.3 Suggestions for building new models and extending the existing spectral library

Our regional predictions of TC and TN show promising results when analyzing soils from geographically distinct areas in central Africa that are not covered by the continental AfSIS SSL (Figure A1). The addition of geographically proximal regions to the large-scale library, which are included in our CSSL, improved prediction accuracy significantly. This improvement underlines the usability of spectral libraries for new regions in general but encourages also the future amendment of currently

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

existing libraries to improve accuracy. To improve future soil analyses and to extend the geographical area covered by the SSL, we suggest the following workflow:

1. **Preprocessing**: Different spectral pre-processing methods influence model and prediction performance. We suggest selecting the best pre-processing strategies using spectral projections and minimizing the reconstruction error (see subsection 2.4).

2. **Estimate uncertainty for new samples**: When analyzing new soil samples from a region which is not covered by the SSL, samples with different composition and hence chemical properties are more likely to be introduced. Samples with high distances in the score space to the SSL cannot be predicted accurately with a high certainty, since they are often highly divergent from the SSL. A preliminary graphical inspection of resampled and pre-processed spectra can already allow for recognition of differences. A further dimension reduction (e.g. with a PCA) with a subsequent 2D or 3D visualization of the first factors provides additional insights into dissimilarity.

3. **Reference analysis for independent validation**: If the new samples are from a completely new region or the new sample set trends to differ from the SSL, a certain number of validation samples is recommended to test for prediction accuracy. The number is dependent on the similarity/dissimilarity to the SSL.

4. **Search for nearest neighbors and train a model**: Run an MBL algorithm to find the nearest neighbors of the new set and train a subsequent weighted average PLS regression.

5. **Model validation**: For predicting soil TC and TN and quantifying the error of these predictions in new geographical regions, a new model validation is required. The nearest neighbor validation is a suitable method, as demonstrated in this study.

Our CSSL is freely available to use and build upon at GitHub. As shown with the AfSIS SSL, the application of already existing libraries and the extrapolation to new regions is accurate and suitable to estimate soil properties. However, to make predictions more accurate, especially for more diverse, heterogeneous and complex soils, more data is required. As demonstrated, the addition of new geographical regions improves the overall prediction accuracy when more proximal central African regions were added to the large-scale library. These results encourage the use and amendment of existing libraries, rather than the construction of new, separate, and extensive databases. Given the existing distribution of samples in the new CSSL, it is especially important to increase the number of forest soils with high TC content, which represent a large portion of the Congo Basin. The future enlargement of the CSSL, preferably facilitated by our suggested workflow, is crucial to fill the gap of soil information in this highly understudied part of the world.

5 Conclusions

Our study presents the results and workflow for building the first central African SSL for predicting soil properties (TC and TN) using lab based MIR spectroscopy in a crucial but understudied area of the African continent. Extrapolations were possible

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

for central Africa and for all the hold-out regions. Our results further demonstrate how MBL algorithms are useful to find spectral similarities and reduce the need for spiking when a new set covers the same score space as the existing library. These encouraging insights highlight the utility of spectral libraries for future applications, since they are not necessarily limited to certain geographical areas. Our approach of augmenting a smaller SSL with a continental SSL, even when scanned on a different instrument, lead to highly accurate predictions for new regions. The CSSL fills an appreciable continental gap of the continental scale AfSIS SSL and contributes an important range of soil variability, particularly from lowland tropical forests. However, in order to improve the accuracy of predicting soil organic matter across regions, especially for soil compartments with high TC and TN contents, our study highlights the need to extend the existing library into new regions. The inclusion of more samples and regions, in particular with more (vayring) data of humid tropical forest soils is crucial to fill existing gaps. Also combining spectral libraries will allow fast analyses of soil samples and provide spatially explicit data across humid tropical Africa. Improved knowledge of soil properties is a major step to maintain ecosystem services that promote human and ecological well-being.

Code and data availability. Data and R codes are available on our GitHub repository 'ssl-central-africa' and can also can be found under Zenodo with the DOI 10.5281/zenodo.4351254 (Summerauer, 2020) to reproduce our results presented in the submitted manuscript. Raw data can be provided to the reviewers upon reasonable request.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Appendix A: Supplementary Figures

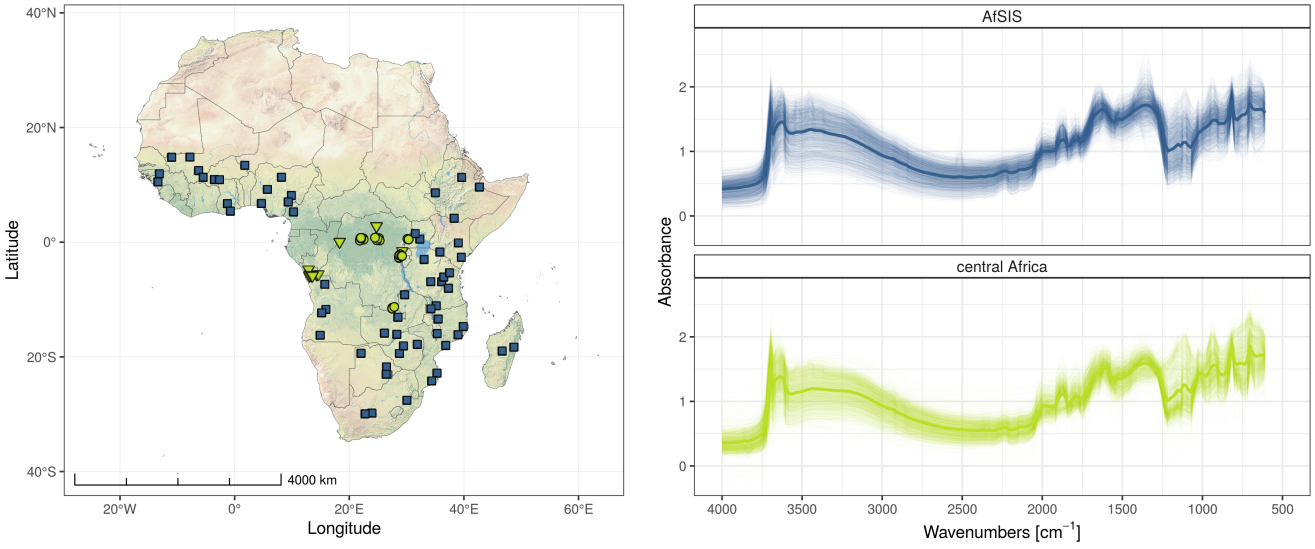

Figure A1. Location of samples used from the two spectral libraries from central Africa and the continental library from the African Soil Information Service (AfSIS; left). The samples used from the core-regions and further analysed in this study are presented with a ○ symbol, the remaining samples of the central African library with a △ symbol and the AfSIS SSL with □ symbol. The average resampled spectrum of each library are shown (bold line) along with the individual resampled sample spectra (transparent lines; right).

Author contributions. J.S. conceived the study. L.S., P.B. and L.R.-L. were the main contributors to the conceptualization, methodology (modeling strategies) and data analyses. MattiB. supported the conceptualization, provided technical support and project coordination. P.B., L.R.-L., S.D., MattiB., MarijnB. and J.S. helped with writing of the manuscript. S.D. and J.S. supervised the project. L.S., MarijnB., B.B., M.R., P.B., E.K., K.V.O., B.V., D.C., A.B.H, P.M., A.S., K.S., B.B.M., E.V.R., G.B., S.D. substantially contributed to the work by organizing, preparing and providing soil samples and the corresponding reference data. All co-authors revised the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are truly grateful to all collaborators in the Democratic Republic of Congo, in Rwanda and in Uganda who made this study possible and helped us with the organization and coordination of numerous field campaigns. We would like to express our gratitude to all the farmers and their families for their hospitality and help during the soil sampling. We would like to warmly thank the teams of the International Institute of Tropical Agriculture (IITA) in Bukavu (Kalambo), in Kinshasa and in Nairobi, the World Agroforestry centre (ICRAF) in Nairobi, the University of Lubumbashi, the Catholic University of Bukavu, the Mountains of The Moon University Fort Portal and

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

of the African Wildlife Foundation (AWF) in Kinshasa for their support and generosity. Finally, we would like to acknowledge the Walter
415 Hochstrasser foundation for their generous financial support of a unique master thesis for six months in Djolu, the province of Tshuapa. Thanks also to Heather Maclean and Travis Drake, for the additional editing of the manuscript.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

References

Baert, G.: Properties and chemical management aspects of soils on different parent rocks in the Lower Zaire, Doctoral thesis, Ghent University, Ghent, Belgium, 1995.

Baert, G., Van Ranst, E., Ngongo, M., Kasongo, E., Verdoodt, A., Mujinya, B., and Mukalay, J.: Guide des Sols en R.D. Congo. Tome II : Description et Données Physico-chimiques de Profils Types., Imprimé par l'Ecole Technique Salama, Lubumbashi, R.D. Congo, 2009.

Baert, G., Ranst, E. V., Ngongo, M., and Verdoodt, A.: Soil Survey in DR Congo – from 1935 until today, Meded. Zitt. K. Acad. overzeese Wet, Brussel, 59, 345–362, 2013.

Barthès, B. G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Saby, N. P., Le Cadre, E., Etayo, A., and Chevallier, T.: Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration – The case of soil inorganic carbon prediction by mid-infrared spectroscopy, Geoderma, 369, 114 272, https://doi.org/10.1016/j.geoderma.2020.114272, https://linkinghub.elsevier.com/retrieve/pii/S0016706119320749, 2020.

Baumann, P.: simplerspec: Soil and plant spectroscopic model building and prediction, https://github.com/philipp-baumann/simplerspec, r package version 0.1.0.9001, 2020.

Baumgartner, S., Barthel, M., Drake, T. W., Bauters, M., Makelele, I. A., Mugula, J. K., Summerauer, L., Gallarotti, N., Cizungu Ntaboba, L., Van Oost, K., Boeckx, P., Doetterl, S., Werner, R. A., and Six, J.: Seasonality, drivers, and isotopic composition of soil CO2 fluxes from tropical forests of the Congo Basin, Biogeosciences, 17, 6207–6218, https://doi.org/10.5194/bg-17-6207-2020, https://bg.copernicus.org/articles/17/6207/2020/, 2020.

Bauters, M., Ampoorter, E., Huygens, D., Kearsley, E., De Haulleville, T., Sellan, G., Verbeeck, H., Boeckx, P., and Verheyen, K.: Functional identity explains carbon sequestration in a 77-year-old experimental tropical plantation, Ecosphere, 6, art198, https://doi.org/10.1890/ES15-00342.1, http://doi.wiley.com/10.1890/ES15-00342.1, 2015.

Bauters, M., Verbeeck, H., Doetterl, S., Ampoorter, E., Baert, G., Vermeir, P., Verheyen, K., and Boeckx, P.: Functional Composition of Tree Communities Changed Topsoil Properties in an Old Experimental Tropical Plantation, Ecosystems, 20, 861–871, https://doi.org/10.1007/s10021-016-0081-0, http://link.springer.com/10.1007/s10021-016-0081-0, 2017.

Bauters, M., Verbeeck, H., Rütting, T., Barthel, M., Bazirake Mujinya, B., Bamba, F., Bodé, S., Boyemba, F., Bulonza, E., Carlsson, E., Eriksson, L., Makelele, I., Six, J., Cizungu Ntaboba, L., and Boeckx, P.: Contrasting nitrogen fluxes in African tropical forests of the Congo Basin, Ecological Monographs, 89, e01 342, https://doi.org/10.1002/ecm.1342, http://doi.wiley.com/10.1002/ecm.1342, 2019a.

Bauters, M., Vercleyen, O., Vanlauwe, B., Six, J., Bonyoma, B., Badjoko, H., Hubau, W., Hoyt, A., Boudin, M., Verbeeck, H., and Boeckx, P.: Long-term recovery of the functional community assembly and carbon pools in an African tropical forest succession, Biotropica, 51, 319–329, https://doi.org/10.1111/btp.12647, https://onlinelibrary.wiley.com/doi/abs/10.1111/btp.12647, 2019b.

Bauters, M., Moonen, P., Summerauer, L., Doetterl, S., Wasner, D., Griepentrog, M., Mumbanza, F. M., Kearsley, E., Ewango, C., Boyemba, F., Six, J., Muys, B., Verbist, B., Boeckx, P., and Verheyen, K.: Soil nutrient depletion and tree functional composition shift following repeated clearing in secondary forests of the Congo Basin, Ecosystems, (in press), 2021.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Present and future Köppen-Geiger climate classification maps at 1-km resolution, Scientific Data, 5, 180 214, https://doi.org/10.1038/sdata.2018.214, http://www.nature.com/articles/sdata2018214, 2018.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., and McBratney, A.: Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy, TrAC Trends in Analytical Chemistry, 29, 1073–1081, https://doi.org/10.1016/j.trac.2010.05.006, https://linkinghub.elsevier.com/retrieve/pii/S0165993610001585, 2010.

Birgé, H. E., Bevans, R. A., Allen, C. R., Angeler, D. G., Baer, S. G., and Wall, D. H.: Adaptive management for soil ecosystem services, Journal of Environmental Management, 183, 371–378, https://doi.org/10.1016/j.jenvman.2016.06.024, 2016.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P., Bernoux, M., and Barthès, B. G.: National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy, Geoderma, 276, 41–52, https://doi.org/10.1016/j.geoderma.2016.04.021, https://linkinghub.elsevier.com/retrieve/pii/S001670611630180X, 2016.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D., and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, Nature, 494, 341–344, https://doi.org/10.1038/nature11882, 2013.

Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A., and Hansen, M. C.: Classifying drivers of global forest loss, Science, 361, 1108–1111, https://doi.org/10.1126/science.aau3445, https://www.sciencemag.org/lookup/doi/10.1126/science.aau3445, 2018.

Dangal, S., Sanderman, J., Wills, S., and Ramirez-Lopez, L.: Accurate and Precise Prediction of Soil Properties from a Large Mid-Infrared Spectral Library, Soil Systems, 3, 11, https://doi.org/10.3390/soilsystems3010011, http://www.mdpi.com/2571-8789/3/1/11, 2019.

Doetterl, S., Asifiwe, R. W., Bamba, F., Bukombe, B., Cooper, M., Hoyt, A., Kidinda, K. L., Maier, A., Mainka, M., Mayrock, J., Muhindo, D., Mukotanyi, S. M., Nabahungu, L., Reichenbach, M., Stegmann, A., Summerauer, L., Unseld, R., Wilken, F., and Fiener, P.: Organic matter cycling along geochemical, geomorphic and disturbance gradients in vegetation and soils of African tropical forests and cropland - Project TropSOC DATABASE_v1.0, SOIL DISCUSSION, (in review), 2021a.

Doetterl, S., Baert, G., Bauters, M., Boeckx, P., Cadisch, G., Cizungu, L., Hoyt, A., Kabaseke, C., Kalbitz, K., Mujinya, B. B., Rewald, B., Six, J., Vanlauwe, B., van Oost, K., Verheyen, K., Vogel, C., Wilken, F., and Fiener, P.: Organic matter cycling along geochemical, geomorphic and disturbance gradients in vegetation and soils of African tropical forests and cropland – Project TropSOC, SOIL DISCUSSION, (in review), 2021b.

Don, A., Schumacher, J., and Freibauer, A.: Impact of tropical land-use change on soil organic carbon stocks - a meta-analysis, Global Change Biology, 17, 1658–1670, https://doi.org/10.1111/j.1365-2486.2010.02336.x, 2011.

FAO and ITTO: The state of forests in the Amazon Basin, Congo Basin and Southeast Asia : a report prepared for the Summit of the Three Rainforest Basins, Tech. rep., UN Food and Agriculture Organization (FAO) and the International Tropical Timber Organization (ITTO), Brazzaville, Republic of Congo, 31 May-3 June, 2011, 2011.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, 37, 4302–4315, https://doi.org/10.1002/joc.5086, https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086, 2017.

Gallarotti, N., Barthel, M., Pujol Pereira, E. I., Bauters, M., Boeckx, P., Mohn, J., Baumgartner, S., Drake, T. W., Longepierre, M., Cizungu, L. N., Verhoeven, E., Mugula, J. K., Makelele, I. A., and Six, J.: N2O reduction in tropical forest soils of the Congo Basin, ISMEJ, (submitted), 2021.

Goyens, C., Verdoodt, A., Van De Wauw, J., Baert, G., Van Engelen, V., Dijkshoorn, J., and Van Ranst, E.: Base de Données Numériques sur les SOls et le TERrain (SOTER) de l'Afrique Centrale (RD Congo, Rwanda et Burundi), Etude et Gestion des Sols, 14, 207–218, 2007.

Guerrero, C., Zornoza, R., Gómez, I., and Mataix-Beneyto, J.: Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy, p. 12, 2010.

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., Zornoza, R., Ruiz-Sinoga, J. D., and Kuang, B.: Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the

spiking subset: Spiking and extra-weighting to improve soil organic carbon predictions with NIR, European Journal of Soil Science, 65, 248–263, https://doi.org/10.1111/ejss.12129, http://doi.wiley.com/10.1111/ejss.12129, 2014.

Heri-Kazi, A. B.: Caractérisation de l'état de dégradation des terres par l'érosion hydrique dans le Sud-Kivu montagneux à l'Est de la R.D. Congo, Thèse doctorale, Université Catholique de Louvain, Louvain La Neuve, 2020.

Imerzoukene, S. and Van Ranst, E.: Une banque de données pédologiques et son S.I.G. pour une nouvelle politique agricole au Rwanda, Meded. Zitt. K. Acad. overzeese Wet, Brussel, 47, 299–325, 2002.

Janik, L. J., Merry, R. H., and Skjemstad, J. O.: Can mid infrared diffuse reflectance analysis replace soil extractions?, Australian Journal of Experimental Agriculture, 38, 681, https://doi.org/10.1071/EA97144, http://www.publish.csiro.au/?paper=EA97144, 1998.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E.: Hole-filled seamless SRTM data V4., 2008.

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Tahar, G., Thiombiano, L., Van Ranst, E., Yemefack, M., and Zougmore, R.: Soil Atlas of Africa, Publication Office of the European Union, Luxembourg, european commission edn., series: JRC soil atlas series, 2013.

Kearsley, E., de Haulleville, T., Hufkens, K., Kidimbu, A., Toirambe, B., Baert, G., Huygens, D., Kebede, Y., Defourny, P., Bogaert, J., Beeckman, H., Steppe, K., Boeckx, P., and Verbeeck, H.: Conventional tree height–diameter relationships significantly overestimate aboveground carbon stocks in the Central Congo Basin, Nature Communications, 4, 2269, https://doi.org/10.1038/ncomms3269, http://www.nature.com/articles/ncomms3269, 2013.

Kearsley, E., Verbeeck, H., Hufkens, K., Van de Perre, F., Doetterl, S., Baert, G., Beeckman, H., Boeckx, P., and Huygens, D.: Functional community structure of African monodominant *Gilbertiodendron dewevrei* forest influenced by local environmental filtering, Ecology and Evolution, 7, 295–304, https://doi.org/10.1002/ece3.2589, http://doi.wiley.com/10.1002/ece3.2589, 2017.

Lin, D., An, X., and Zhang, J.: Double-bootstrapping source data selection for instance-based transfer learning, Pattern Recognition Letters, 34, 1279–1285, https://doi.org/10.1016/j.patrec.2013.04.012, https://linkinghub.elsevier.com/retrieve/pii/S0167865513001621, 2013.

Lobsey, R., C., Viscarra Rossel, R. A., Poudier, P., and Hedley, C. B.: rs-local data-mines information from spectral libraries to improve local calibrations, European Journal of Soil Science, 68, 840–852, 2017.

Minten, K.: Development Of a Business Plan For Production And Export Of Green Coffee Beans From The Equateur Province In The Democratic Republic Of The Congo, Master's thesis, Ghent University, Ghent, Belgium, 2017.

Moonen, P. C., Verbist, B., Boyemba Bosela, F., Norgrove, L., Dondeyne, S., Van Meerbeek, K., Kearsley, E., Verbeeck, H., Vermeir, P., Boeckx, P., and Muys, B.: Disentangling how management affects biomass stock and productivity of tropical secondary forests fallows, Science of the Total Environment, 659, 101–114, https://doi.org/10.1016/j.scitotenv.2018.12.138, 2019.

Mujinya, B., Van Ranst, E., Verdoodt, A., Baert, G., and Ngongo, L.: Termite bioturbation effects on electro-chemical properties of Ferralsols in the Upper Katanga (D.R. Congo), Geoderma, 158, 233–241, https://doi.org/10.1016/j.geoderma.2010.04.033, https://linkinghub.elsevier.com/retrieve/pii/S0016706110001539, 2010.

Mujinya, B., Mees, F., Erens, H., Dumon, M., Baert, G., Boeckx, P., Ngongo, M., and Van Ranst, E.: Clay composition and properties in termite mounds of the Lubumbashi area, D.R. Congo, Geoderma, 192, 304–315, https://doi.org/10.1016/j.geoderma.2012.08.010, https://linkinghub.elsevier.com/retrieve/pii/S0016706112003059, 2013.

Mujinya, B., Adam, M., Mees, F., Bogaert, J., Vranken, I., Erens, H., Baert, G., Ngongo, M., and Van Ranst, E.: Spatial patterns and morphology of termite (Macrotermes falciger) mounds in the Upper Katanga, D.R. Congo, CATENA, 114, 97–106, https://doi.org/10.1016/j.catena.2013.10.015, https://linkinghub.elsevier.com/retrieve/pii/S0341816213002609, 2014.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Mujinya, B. B.: Effects of Macrotermes termites on the mineralogical and electro-chemical properties of Ferralsol materials in the Upper Katanga (D.R. Congo), Doctoral thesis, Ghent University, Ghent, Belgium, 2012.

Mujinya, B. B., Mees, F., Boeckx, P., Bodé, S., Baert, G., Erens, H., Delefortrie, S., Verdoodt, A., Ngongo, M., and Van Ranst, E.: The origin of carbonates in termite mounds of the Lubumbashi area, D.R. Congo, p. 11, 2011.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., and Montanarella, L.: Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach, Soil Biology and Biochemistry, 68, 337–347, https://doi.org/10.1016/j.soilbio.2013.10.022, https://linkinghub.elsevier.com/retrieve/pii/S0038071713003660, 2014.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J. M., Shepherd, K. D., Stenberg, B., Towett, E. K., Vargas, R., and Wetterlind, J.: Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring, in: Advances in Agronomy, vol. 132, pp. 139–159, Elsevier, https://doi.org/10.1016/bs.agron.2015.02.002, https://linkinghub.elsevier.com/retrieve/pii/S0065211315000425, 2015.

Næs, T.: The design of calibration in near infra-red reflectance analysis by clustering, Journal of Chemometrics, 1, 121–134, https://doi.org/10.1002/cem.1180010207, http://doi.wiley.com/10.1002/cem.1180010207, 1987.

Padarian, J., Minasny, B., and McBratney, A.: Transfer learning to localise a continental soil vis-NIR calibration model, Geoderma, 340, 279–288, https://doi.org/10.1016/j.geoderma.2019.01.009, https://linkinghub.elsevier.com/retrieve/pii/S0016706118305639, 2019.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2020.

Ramirez-Lopez, L.: resemble: Regression and Similarity Evaluation for Memory-Based Learning in Spectral Chemometrics, https://CRAN.R-project.org/package=resemble, r package version 2.1.1, 2020.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., and Scholten, T.: The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets, Geoderma, 195-196, 268–279, https://doi.org/10.1016/j.geoderma.2012.12.014, https://linkinghub.elsevier.com/retrieve/pii/S0016706112004314, 2013.

Ramirez-Lopez, L., Wadoux, A. M. J., Franceschini, M. H. D., Terra, F. S., Marques, K. P. P., Sayão, V. M., and Demattê, J. A. M.: Robust soil mapping at the farm scale with vis–NIR spectroscopy, European Journal of Soil Science, 70, 378–393, https://doi.org/10.1111/ejss.12752, https://onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12752, 2019.

Rinnan, A.: Pre-processing in vibrational spectroscopy – when, why and how, Analytical Methods, 6, 7124–7129, https://doi.org/10.1039/C3AY42270D, http://xlink.rsc.org/?DOI=C3AY42270D, 2014.

Sanderman, J., Savage, K., and Dangal, S. R.: Mid-infrared spectroscopy for prediction of soil health indicators in the United States, Soil Science Society of America Journal, 84, 251–261, https://doi.org/10.1002/saj2.20009, https://onlinelibrary.wiley.com/doi/abs/10.1002/saj2.20009, 2020.

Savitzky, A. and Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Analytical Chemistry, 36, 1627–1639, https://doi.org/10.1021/ac60214a047, https://pubs.acs.org/doi/abs/10.1021/ac60214a047, 1964.

Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., and Vohland, M.: Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations, Geoderma, 354, 113 856, https://doi.org/10.1016/j.geoderma.2019.07.014, https://linkinghub.elsevier.com/retrieve/pii/S0016706119304537, 2019.

Shenk, J. S., Westerhaus, M. O., and Berzaghi, P.: Investigation of a LOCAL Calibration Procedure for near Infrared Instruments, Journal of Near Infrared Spectroscopy, 5, 223–232, https://doi.org/10.1255/jnirs.115, http://journals.sagepub.com/doi/10.1255/jnirs.115, 1997.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Shepherd, K. D. and Walsh, M. G.: Infrared Spectroscopy—Enabling an Evidence-Based Diagnostic Surveillance Approach to Agricultural and Environmental Management in Developing Countries, Journal of Near Infrared Spectroscopy, 15, 1–19, https://doi.org/10.1255/jnirs.716, http://journals.sagepub.com/doi/10.1255/jnirs.716, 2007.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., and Viscarra Rossel, R. A.: Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations, Science China Earth Sciences, 57, 1671–1680, https://doi.org/10.1007/s11430-013-4808-x, http://link.springer.com/10.1007/s11430-013-4808-x, 2014.

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., and Zhou, Y.: Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral library: vis-NIR predictions of soil carbon with scL-PLSR, European Journal of Soil Science, 66, 679–687, https://doi.org/10.1111/ejss.12272, http://doi.wiley.com/10.1111/ejss.12272, 2015.

Sila, A. M., Shepherd, K. D., and Pokhariyal, G. P.: Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties, Chemometrics and Intelligent Laboratory Systems, 153, 92–105, https://doi.org/10.1016/j.chemolab.2016.02.013, https://linkinghub.elsevier.com/retrieve/pii/S0169743916300351, 2016.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., and McLaughlin, M. J.: The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties, Applied Spectroscopy Reviews, 49, 139–186, https://doi.org/10.1080/05704928.2013.811081, http://www.tandfonline.com/doi/abs/10.1080/05704928.2013.811081, 2014.

Stevens, A. and Ramirez-Lopez, L.: prospectr: Miscellaneous Functions for Processing and Sample Selection of Spectroscopic Data, https://CRAN.R-project.org/package=prospectr, r package version 0.2.0, 2020.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael, B.: Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy, PLoS ONE, 8, e66 409, https://doi.org/10.1371/journal.pone.0066409, https://dx.plos.org/10.1371/journal.pone.0066409, 2013.

Summerauer, L.: Sustainable Agricultural Intensification Methods of Cassava Based Systems for Improving Livelihoods and Forest Conservation in the Congo Basin, Master's thesis, ETH Zurich, Zurich, Switzerland, 2017.

Summerauer, L.: laura-summerauer/ssl-central-africa: Codes and data for manuscript submission (submission version), https://doi.org/10.5281/ZENODO.4351254, https://zenodo.org/record/4351254, 2020.

Tsakiridis, N. L., Theocharis, J. B., Panagos, P., and Zalidis, G. C.: An evolutionary fuzzy rule-based system applied to the prediction of soil organic carbon from soil spectral libraries, p. 18, 2019.

Tyukavina, A., Hansen, M. C., Potapov, P., Parker, D., Okpa, C., Stehman, S. V., Kommareddy, I., and Turubanova, S.: Congo Basin forest loss dominated by increasing smallholder clearing, Science Advances, 4, https://doi.org/10.1126/sciadv.aat2993, 2018.

Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., and Zalidis, G.: A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation, Geoderma, 340, 11–24, https://doi.org/10.1016/j.geoderma.2018.12.044, https://linkinghub.elsevier.com/retrieve/pii/S0016706118307006, 2019.

UNESCO World Heritage Centre: World Heritage in the Congo Basin, Tech. rep., Paris, France, 2010.

Van Ranst, E., Verdoodt, A., and Baert, G.: Soil Mapping in Africa at the Crossroads: Work to Make up for Lost Ground, Meded. Zitt. K. Acad. overzeese Wet, Brussel, 56, 147–163, 2010.

Veldkamp, E., Schmidt, M., Powers, J. S., and Corre, M. D.: Deforestation and reforestation impacts on soils in the tropics, Nature Reviews Earth & Environment, https://doi.org/10.1038/s43017-020-0091-5, http://www.nature.com/articles/s43017-020-0091-5, 2020.

https://doi.org/10.5194/soil-2020-99
Preprint. Discussion started: 8 January 2021

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B., Bartholomeus, H., Bayer, A., Bernoux, M., Böttcher, K., Brodský, L., Du, C., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C., Knadel, M., Morrás, H., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E. R., Sanborn, P., Sellitto, V., Sudduth, K., Rawlins, B., Walter, C., Winowiecki, L., Hong, S., and Ji, W.: A global spectral library to characterize the world's soil, Earth-Science Reviews, 155, 198–230, https://doi.org/10.1016/j.earscirev.2016.01.012, https://linkinghub.elsevier.com/retrieve/pii/S0012825216300113, 2016.

Vollset, S. E., Goren, E., Yuan, C.-W., Cao, J., Smith, A. E., Hsiao, T., Bisignano, C., Azhar, G. S., Castro, E., Chalek, J., Dolgert, A. J., Frank, T., Fukutaki, K., Hay, S. I., Lozano, R., Mokdad, A. H., Nandakumar, V., Pierce, M., Pletcher, M., Robalik, T., Steuben, K. M., Wunrow, H. Y., Zlavog, B. S., and Murray, C. J. L.: Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study, The Lancet, p. S0140673620306772, https://doi.org/10.1016/S0140-6736(20)30677-2, https://linkinghub.elsevier.com/retrieve/pii/S0140673620306772, 2020.

Vågen, T.-G., Winowiecki, L. A., Desta, L., Tondoh, E. J., Weullow, E., Shepherd, K., and Sila, A.: Mid-Infrared Spectra (MIRS) from ICRAF Soil and Plant Spectroscopy Laboratory: Africa Soil Information Service (AfSIS) Phase I 2009-2013, https://doi.org/10.34725/DVN/QXCWP1, https://doi.org/10.34725/DVN/QXCWP1, edition: V1 Section: 2020-08-10 01:44:20.122, 2020.

Wetterlind, J. and Stenberg, B.: Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples, European Journal of Soil Science, 61, 823–843, https://doi.org/10.1111/j.1365-2389.2010.01283.x, http://doi.wiley.com/10.1111/j.1365-2389.2010.01283.x, 2010.

Wold, S., Trygg, J., Berglund, A., and Antti, H.: Some recent developments in PLS modeling, Chemometrics and Intelligent Laboratory Systems, 58, 131–150, https://doi.org/10.1016/S0169-7439(01)00156-3, https://linkinghub.elsevier.com/retrieve/pii/S0169743901001563, 2001.

WRB, I. W. G.: World Reference Base for Soil Resources 2006, World Soil Resources Report No 103, Food and Agriculture Organization of the United Nations, Rome, 2006.