Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring

Philipp Baumann[1, 4], Anatol Helfenstein[1, 2], Andreas Gubler[3], Armin Keller[4], Reto Giulio Meuli[3], Daniel Wächter[3], Juhwan Lee[5], Raphael Viscarra Rossel[6], and Johan Six[1]

[1]Institute of Agricultural Sciences, Department of Environmental Systems Science (D-USYS), ETH Zürich, Switzerland
[2]Soil Geography and Landscape Group, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands
[3]Swiss Soil Monitoring Network (NABO), Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland
[4]Swiss Competence Center for Soils (KOBO), School of Agricultural, Forest and Food Sciences HAFL, Bern University of Applied Sciences BFH, Bern, Switzerland
[5]Department of Smart Agro-industry, Gyeongsang National University, Jinju, 52725, Republic of Korea
[6]Soil and Landscape Science, School of Molecular and Life Sciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia

Correspondence: Philipp Baumann (baumann-philipp@protonmail.com), Current Address: Swiss Competence Center for Soils (KOBO), School of Agricultural, Forest and Food Sciences HAFL, Bern University of Applied Sciences BFH, Bern, Switzerland

Abstract.

Information on soils' composition and physical, chemical and biological properties is paramount to elucidate agroecosystem functioning in space and over time. For this purposes we developed a national Swiss soil spectral library (SSL; $n = 4374$) in the mid-infrared (mid-IR), calibrating 16 properties from legacy measurements on soils from the Swiss biodiversity monitoring program ($n = 3778$; 1094 sites) and the Swiss long-term monitoring network (NABO; $n = 596$; 71 sites). General models were trained with the interpretable rule-based learner CUBIST, testing combinations of $\{5, 10, 20, 50, 100\}$ ensembles of rules (committees) and $\{2, 5, 7, 9\}$ nearest-neighbors used for local averaging with repeated 10-fold cross-validation grouped by location. To evaluate the information in spectra to facilitate long-term soil monitoring at a plot-level, we conducted 71 model transfers for the NABO sites to induce locally relevant information from the SSL, using the data-driven sample selection method RS-LOCAL.Ten soil properties were estimated with discrimination capacity suitable for screening ($R^2 \geq 0.72$; ratio of performance to interquartile distance (RPIQ) ≥ 2.0), out of which total carbon (C), organic C (OC), total N, pH, and clay showed accuracy eligible for accurate diagnostics ($R^2 > 0.8$; RPIQ ≥ 3.0). CUBIST and the spectra estimated total C accurately with the root mean square error (RMSE) = $8.4\,\mathrm{g\,kg^{-1}}$ and the RPIQ = 4.3 while the measured range was $1 – 583\,\mathrm{g\,kg^{-1}}$, and OC with RMSE = $9.3\,\mathrm{g\,kg^{-1}}$ and RPIQ = 3.4 (measured range $0 – 583\,\mathrm{g\,kg^{-1}}$). Compared to the general statistical learning approach, the local transfer approach — using two respective training samples — on average reduced the RMSE of total C per site fourfold. We found that the selected SSL subsets were highly dissimilar compared to validation samples, in terms of both their spectral input space and the measured values. This suggests that data-driven selection with RS-LOCAL leverages chemical diversity in composition rather than similarity. Our results suggest that mid-IR soil estimates were sufficiently accurate to support many soil applications that require a large volume of input data, such as precision agriculture, soil C accounting and

monitoring, and digital soil mapping. This SSL can be updated continuously, for example with samples from deeper profiles and organic soils, so that the measurement of key soil properties becomes even more accurate and efficient in the near future.

1 Introduction

Soils provide a manifold of functions within terrestrial ecosystems, many of which are vital for humankind. To quantify these functions from the soils' composition and properties, one typically relies on physical, chemical and biological laboratory analytical measurements. Doing this consumes both financial resources and time. For example, repeated measurements are needed to describe soil functioning in changing environments, for example in response to agronomic management. Soil visible (vis) and infrared (IR) spectroscopic measurements and modeling have become indispensable tools to gather quick, relatively accurate, and inexpensive estimates of soil properties, both on the field and in the laboratory (Nocita et al., 2015; Viscarra Rossel et al., 2016, 2017). Once soil chemical and physical properties are calibrated to the spectra, a single mid-IR ($4000-500\,\mathrm{cm}^{-1}$; $2500-25000\,\mathrm{nm}$) or vis–NIR ($25000-4000\,\mathrm{cm}^{-1}$; $400-2500\,\mathrm{nm}$) measurement can be used to estimate multiple soil properties of new samples. Soil is a complex matrix with many organic and mineral components. This yields spectra with absorptions that overlap and contain many and often highly correlated variables. Hence, to successfully develop calibrations and make predictions for attributes related to soil composition on more samples, statistical learning methods are needed to find and use relationships between these variables and measured attributes. It is important to consider that the diversity in spectral characteristics typically reflects soils' chemical and physical composition. Since the soil composition is influenced by the soil forming factors — soil parent material, climate, topography, organisms, and age of soils (Dokuchaev, 1899; Jenny, 1941) — these factors provide further means of causally interpreting and judging the applicability of the method for a particular set of soils. Compared to the NIR, mid-IR offers a more accurate characterization of soils' chemistry since this region contains the fundamental vibrations with more defined peaks (Janik et al., 1998; Viscarra Rossel et al., 2006).

A soil spectral library (SSL) can be defined as a well-ordered and harmonized collection of soil samples, their spectra, analytical reference measurements, contextual information, and additional metadata on samples and methods used. A central question behind the development of large SSLs is how to achieve accurate local predictions based on established collections of soil information — for example within a new landscape, ecosystem, farm, field, or plot in a new region — where reference data of only a few observations are available. More recently, SSLs that span large geographical extents are being developed (Sila et al., 2016; Viscarra Rossel et al., 2016; Padarian et al., 2019b; England and Viscarra Rossel, 2018; Briedis et al., 2020; Angelopoulou et al., 2020; Dangal et al., 2019). These efforts are motivated by the prospect that soil spectroscopy can supplement many conventional methods of soil analysis. A range of predictive modeling strategies and algorithms have been tested for soil spectral analysis, among others involving tools from chemometrics (e.g., partial least squares (PLS) regression) (Janik and Skjemstad, 1995), traditional machine learning (e.g., regression tree methods) (Viscarra Rossel and Webster, 2012), to convolutional neural networks (CNNs) (Padarian et al., 2019a, b; Tsakiridis et al., 2020).

There are two main strategies for estimating properties of new soils using spectra. The first one is to calibrate one general or global model that is applied to predict new samples, and the other is to derive local calibrations by conditioning on a specific set of observations and features of the SSL to new data based on soil knowledge and/or algorithms. However, empirical evaluation of local and global methods are needed in different contexts where data on soil attributes is needed (i.e. soil study, soil mapping project). Such studies or applications should consider the "no-free-lunch" theorems for machine learning and optimization (Wolpert, 1996; Wolpert and Macready, 1997): there is no single algorithm or combinations of them that work best under all situations or applications.

General statistical learning makes use of all available training data to construct one parametric model. In contrast, local learning methods combine different learning methods, supervised and/or unsupervised, and together with domain knowledge produce more modular forms of learning (Solomatine, 2008). The available training set can be a subset and algorithmic sub-models can thereby be optimized to more accurately predict new single observations or groups of them. Local learning has also been termed *transfer learning*. Transfer learning is a general expression for adapting previous knowledge gained from existing data (i.e., model representation) for a new prediction case (Pratt et al., 1993; Pratt and Thrun, 1997; Thrun and Pratt, 1998). It has been defined as a transfer from knowledge in (a) source task(s) or domain(s)—here, an SSL—to a target domain (Pan and Yang, 2010), and thus comprising soils from new locations in this case.

The soil spectroscopy community has suggested several approaches to achieve local calibrations based on an established SSL and its feature space. One example is augmenting (spiking) SSLs with a few unweighted (Guerrero et al., 2010; Seidel et al., 2019) or extra-weighted (Guerrero et al., 2014, 2016) local samples. Other studies calibrated separate models on partitions of training data that were derived from applying certain criteria (i.e., geographical region, terrain attributes, parent material, soil type, land use, spectra-based clustering) (Sila et al., 2016; Ogen et al., 2019). Still others used memory-based learning based on spectral similarity, extracting useful information from compositional relatedness of soils (Ramirez-Lopez et al., 2013; Clairotte et al., 2016; Hong et al., 2019; Dangal et al., 2019) or additionally geographic proximity (Tziolas et al., 2019). These all produce individual models for each sample to be predicted. Memory-based learning combines both *lazy learning*, where a subset of stored samples are only retrieved and modeled when new samples are predicted, and *local learning* principles, where modeled subsets define points within a local neighborhood (Dietterich et al., 1993). The spectrum-based learner developed by Ramirez-Lopez et al. (2013) is a prominent memory-based method for which each new prediction sample forms its own target domain. The selection of source instances is governed by spectral similarity. Therefore, the spectrum-based learner is also considered a transfer learning method. Another approach used by Padarian et al. (2019a) was re-training weights within specific layers of a deep CNN using local (target) sets, which were spectral soil data sets per country (parameter-transfer approach). Finally, the selection of matching SSL samples using the resampling-based selection RS-LOCAL algorithm has also been used (Lobsey et al., 2017). Lobsey et al. (2017) showed that this data-driven transfer approach outperforms most other current methods for deriving local estimates. Still, despite these promising learners, transferring the useful information contained within large and diverse SSLs, and their resulting calibrations onto new, local target areas with unique soil characteristics remains challenging due to soil complexity.

RS-LOCAL obtains locally-relevant information by selecting specific rows (instances) from the training set and transfer them to the prediction set. RS-LOCAL is an example of an instance or sample transfer approach. It heavily relies on sampling and performance-driven reduction of the library, yielding a subset of samples that can accurately estimate the properties of soils in the local target task. We wanted to investigate this promising new method for local soil estimation and monitoring in Switzerland because it makes no prior assumptions on which samples from the library could be useful. This makes it potentially more accurate and as well more flexible to new local soil contexts than when creating constraints with similarity measures. A further advantage for large SSLs is that it removes samples that might be spectrally similar but cause inaccurate calibrations (i.e., erroneous measurements or spectra with confounding effects). Such a local approach however requires an well-established and sufficiently diverse SSL in order to extract useful soils that are locally relevant.

Thus, our first goal was to develop a national mid-IR SSL with reference measurements for Switzerland to deliver 16 key chemical and physical soil proxies. This SSL includes soils and their analysis data from the long-term Swiss soil monitoring network (NABO; 71 agricultural sites with times series measurements, $n = 596$) and the Swiss biodiversity monitoring (BDM) network (1094 grid-locations, $n = 3778$). This is the first operational SSL for Switzerland in the mid-IR that allows means for spectral estimation with sufficient existing soil diversity. The second goal was to develop general rule-based models for all available soil properties using the CUBIST algorithm. Further, we wanted to infer important spectral regions in the models and their chemical associations, which we illustrated with the estimation of total carbon (C) contents.

For soil monitoring and also for determining C stocks, it is crucial to obtain locally accurate spectral estimates of key soil properties such as organic C contents, from high soil variability of large SSLs and over time. This was our motivation to design a predictive transfer workflow that was adaptive to soils' composition and properties for each long-term monitoring site. Hence, our third goal was to leverage the SSL with its spatial and temporal variability of soils to derive local calibrations by transfer learning with RS-LOCAL. Specifically, we aimed for showing local models' capacity to reproducing time-series measurements (starting from 1985) of soil C at the Swiss agricultural long-term monitoring sites based on spectral analyses and two calibration samples per site. To the best of our knowledge, there is no study yet that has evaluated the usefulness of a large and diverse SSL for systematic soil monitoring. We therefore wanted to design a local calibration strategy using transfer learning, that would be effective in reducing (conditional) errors at monitoring plots compared to the general rules derived in the first aim. Furthermore, we had a strong interest in identifying the mechanisms, considering both soil knowledge and data distributions, of how such a local transfer would induce locally-adaptive soil estimation.

In brief, our work addresses three objectives: (1) developing a national SSL, (2) building general prediction models using CUBIST, and (3) build site-specific (local) prediction models using RS-LOCAL.

2 Material and Methods

2.1 Soils and data sets

To establish the Swiss SSL, we obtained soil samples and reference data from two different sources: 1) the Swiss soil monitoring network (NABO), and 2) the Swiss Biodiversity Monitoring (BDM) program (Bundesamt für Umwelt (BAFU), 2014) (Figure 1). The NABO currently consists of 108 sites where soils are being continuously measured every five years since 1985 for long-term soil monitoring. Out of the 108 sites, we selected 71 sites under agricultural management—comprising of arable land (33 sites), permanent grassland (26 sites), and special crops (11 sites; horticulture)—and one protected area. For the mid-infrared SSL, we used 596 NABO soil samples from 6 campaigns conducted between 1985 and 2015.

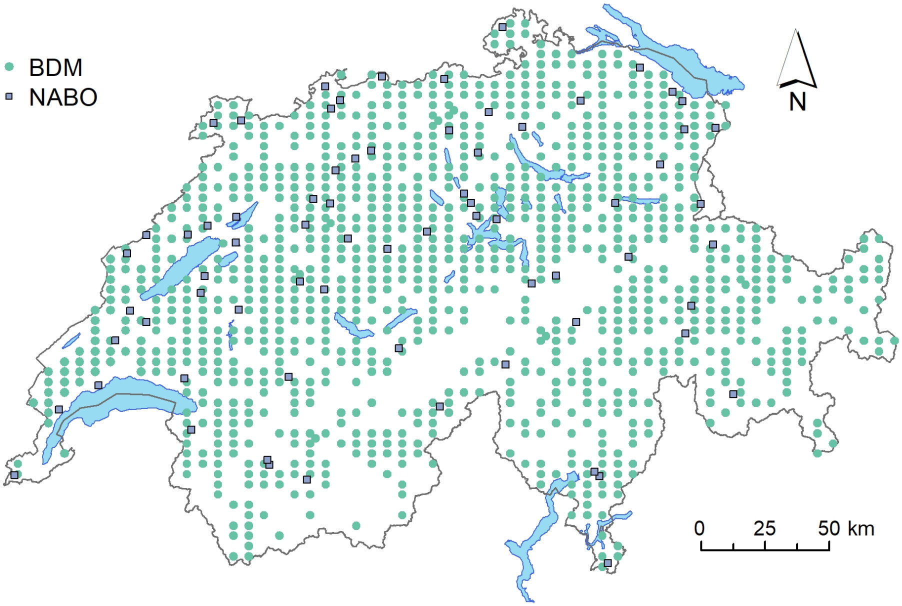

Figure 1. Swiss map with sampling locations of mid-infrared spectral library including the sites of the Biodiversity Monitoring Program (BDM; $6 \times 4\,\mathrm{km}$; $n = 1094$) and the National Soil Monitoring Network (NABO; $n = 71$). 71 NABO sites ($10\,\mathrm{m}\times\,10\,\mathrm{m}$) were sampled with a grid-based stratified design. 1094 BDM samples were obtained from single sampling events. The NABO sites have been continuously sampled and measured in five-year intervals since 1985.

The plots at the NABO sites covered $10\,\mathrm{m}\times\,10\,\mathrm{m}$ each. These were repeatedly sampled for 0–20 cm soil depth. Four replicate samples were taken by stratified random sampling and bulking four times 25 cores from 100 sub-areas of $1\,\mathrm{m}^2$ to account for small-scale soil variability. Desaules et al. (2010) and Gubler et al. (2019) detailed the sample collection and data harmonization process of the measurements. The soils of the BDM were sampled at 0–0.2 m depth from positions on a regular grid of 6 km ×4 km laid over Switzerland (a total of 1094 locations). The points that were not sampled were inaccessible; these were mostly in the alpine regions. Each sampled location included four sub-samples that were taken at the intersection of the four cardinal directions from the center point and the circumference of a circle with a radius of 3 m to 3.5 m (Meuli et al., 2017). Due to it's design covering all major geographic regions in Switzerland—the Jura mountain range, the Central Plateau, and the Alps—the BDM sampling campaign comprises a major part of the biogeochemical diversity of soils and predominant land use types in Switzerland. The wide coverage of soil conditions are an important source of soil chemical variability.

2.2 Chemical reference analysis

Data on chemical and physical soil properties were previously measured and provided by the NABO group. All laboratory soil analyses for the 16 properties were based on the protocols of the Swiss Standard Method (Agroscope, 1996). Mineral elements were determined by extraction with 1:10 ammonium acetate-EDTA solution (AAE10; method Agroscope 1996). The measured properties were total C, organic C (OC), total nitrogen (N), pH ($CaCl_2$), $CaCO_3$, clay, silt, sand, CEC_{pot}, P(AAE), K(AAE), Ca(AAE), Mg(AAE), Cu(AAE), Zn(AAE), and Fe(AAE). For samples of BDM and for the more recent NABO sampling campaigns five and six (years 2009 – 2014), the total C and N measurements were done with dry combustion (Leco TruSpec). For campaigns one through four (years 1985 – 2014), the OC contents determined with wet oxidation using a modified Walkley-Black method were transformed into dry combustion equivalents, using site-specific robust linear regressions (complementary data of campaigns five and six) (Gubler et al.). Carbonates were determined by volumetric calcimetry using hydrochloric acid (HCl) for digestion. Organic C was obtained by difference of total C and carbonate-C when pH was greater than 6.5. Inorganic (carbonate) C was calculated with $0.12 \times CaCO_3$. The texture was determined by the pipette method. The pH was measured in $CaCl_2$ using a 1:2 volumetric ratio of soil to water. For CEC_{pot}, the exchangeable elements were extracted with a 0.05 N-0.025 N HCl-H_2SO_4 solution, which was buffered with triethanolamine for soil samples with pH > 5.9. All soil properties were referenced to dry weight by water correction after drying at 105 °C. All chemical analyses of NABO soils were done on four bulked replicates per site and sampling event. For BDM locations, four spatial replicates were measured each.

2.3 Measuring and processing spectra

All milled soil samples from the NABO and the BDM archive (n = 4374; with a particle size below 100 μm) were measured with the Vertex 70 Fourier-transform spectrometer from Bruker (Bruker Optik GmbH, Ettlingen, Germany) at ETH Zurich, using a high-throughput accessory (HTS-XT) and custom 24-well plates tailored to diffuse reflectance measurements. The mid-IR spectrometer was equipped with a KBr beamsplitter and a Mercury Cadmium Telluride (MCT) detector, which was permanently cooled with liquid nitrogen during the measurements. The reflectance spectra were acquired between 7500 cm^{-1} (1333.3 nm) and 600 cm^{-1} (16666.7 nm) at an effective resolution of 2 cm^{-1}, and trimmed to the mid-IR range between 3996 cm^{-1} and 600 cm^{-1} before further processing (see below).

Each soil sample was measured twice. The soil surface was flattened evenly and without compression by the thin round middle part of the spatula. The first measurement position of the 24-well plate contained a gold (Au) reference surface, which produced a single reflectance spectrum for normalizing the reflectance of the 23 following soil measurements. The "atmospheric compensation" routine implemented in the Bruker OPUS software was used to eliminate unwanted absorptions of H_2O vapour continuum and CO_2 gas in the measurement chamber, based on the single channel reference spectrum measured once on each plate. All single channel reflectance spectra were obtained by averaging 32 internal measurements.

The resulting reflectance spectra (R; background referenced) were converted to apparent absorbance (A)by $A = log_{10}(1/R)$. Then, an average spectrum per sample was produced by calculating the mean of all spectral variables for the measured replicates. Finally, the spectrum offset and further scatter effects were reduced and the features were transformed with a Savitzky-Golay (Savitzky and Golay, 1964) first derivative smoother using a window size of 35 variables (70 cm^{-1}) and third order polynomial fit. Finally, we selected every 8th spectral variable to reduce redundancy in the spectra (collinearity) and produce more parsimonious spectral estimates of soil properties. This resulted in 209 variables between 634 cm^{-1} and 3962 cm^{-1}, which formed the predictors for the subsequent general and local transfer modeling.

2.4 Data processing and statistical computing

All spectral and reference data were processed and modeled with the R software environment for statistical computing and graphics version 3.6.0 (R Core Team, 2019). We used the caret (Kuhn, 2020) R package to streamline the statistical learning process. Basic data transformations such as data preparation and aggregation were done using the tidyverse (Wickham, 2019) set of packages and data.table (Dowle and Srinivasan, 2019). The spectral data were handled and processed with the simplerspec (Baumann, 2019) and prospectr (Stevens and Ramirez-Lopez, 2013) packages.

2.5 General soil estimation: Rules for the entire SSL

The general soil estimation was done by rules trained with the CUBIST (Quinlan, 1993) learner, separately developed for each analytical soil measure. We chose this algorithm since it has shown excellent performance for modelling soil information and developing SSLs with rather large soil variability and multicollinear spectral variables (Bui et al.; Viscarra Rossel and Webster, 2012; Stevens et al., 2013; Miller et al.; Peng et al.; Viscarra Rossel et al., 2016; Dangal et al., 2019; Padarian et al., 2019b), and because its interpretation is mechanistically more intuitive as it is a form of data partitioning (simple conditions and linear equations). CUBIST first forms model trees using basic mechanisms of M5 (Quinlan, 1992). CUBIST is a form of rule-based decision tree with piecewise linear models. Wang and Witten (1996) outlined detailed principles behind the construction of the model trees and derivation of rules, and Viscarra Rossel and Webster (2012) described it for soil spectroscopic modeling.

A CUBIST prediction rule is a unique set of conditions, "if-then" logical statements, together with the associated ordinary linear regression model. During training, the condensed regression equations are made for samples in the terminal nodes. All preceding split variables are potentially allowed for regression in a final node; however, some of them are pruned or combined in the rules. The smoothed regression equation with the selected variables allows then to predict an individual new observation. CUBIST features two empirical parameters that can improve predictions: committees and neighbors. Committees are ensembles of rules that are created by successive construction of trees, which correct predictions of preceding rules and thereby lower predictive errors by averaging. When neighbors are used (maximum 9), a new training sample is predicted using both unweighted or weighted averages of the measured values of the nearest neighbors using all features in the training set and the prediction of the new sample using the training rule(s).

2.5.1 Model development and validation

We tested a full-factorial combination of $\{5, 10, 20, 50, 100\}$ committees of rules and $\{2, 5, 7, 9\}$ neighbors to tune the CUBIST models. To get realistic estimates of the models' general performance, we defined a grouped ten-fold cross-validation scheme that treated the entire site (e.g., for total C: NABO: 71 sites; BDM: 1079 sites) as independent in the modeling data sets. This made all observations from a site the unit of prediction, making the procedure equivalent to external cross-validation.

To reduce the bias-variance trade-off in the assessment, we repeated five times the grouped ten-fold cross-validation (CV) procedure (Friedman et al., 2008; Kuhn and Johnson, 2013). The division into training and validation proportions of the data was done in consistent and repeatable manner (pseudo-random number generation). We considered this site grouping factor as prior information when cross-validation segments were created, so that samples from a particular site were only present within one segment (fold) of a cross-validation split. This grouped assignment prevented that the relationships were trained on the model fitting sets and prevented a particular site from leaking into the testing segments, yielding reliable generalization errors.

We tested the correspondence of mid-IR and model-derived predictions ($\hat{x_i}$) and measured standard reference measurements (x_i) with common regression metrics. We cross-validated the inaccuracy of the models with the root mean square error (RMSE). The mean squared error (MSE) was further decomposed into mean error (ME) or bias and the standard deviation of the error (SDE) or imprecision, so that $\mathrm{RMSE}^2 = \mathrm{ME}^2 + \mathrm{SDE}^2$ (Viscarra Rossel and McBratney, 1998). To describe the linear dependency between measurements and modeled values and give a relative goodness of fit, the coefficient of determination (R^2) from linear regression was also reported. All metrics were aggregated from five estimates from independent resampling repeats. We reported mean values and standard deviations to provide uncertainties of the estimates.

2.5.2 Deriving important spectral variables

The importance of each spectral variable was assessed based on its usage in the rule conditions and the model for CUBIST. We used the recursive feature elimination (RFE) method, a backwards variable-selection algorithm described by Guyon et al. (2002), to test whether the modeling can be simplified and to find most important spectral features. Soil reflectance spectra typically contain many correlated and potentially redundant variables. Therefore, constraining them to relevant subsets that feed into the modeling can further improve predictive accuracy and reduce computation time and storage for model updates. We recursively eliminated subsets of variables with low CUBIST variable importance, calculated as the average relative usage frequencies of a particular variable in split conditions and regressions.This step-wise variable reduction was based on the following predefined subset sizes S_i, starting with the full set at $i = 1$ and ending with the most important predictor at $i = 30$:

$$S_i = \{209, 150, 120, 105, 90, 75, 60, 50, 40, 35, 30, 25, 20, 17, 14, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\} \qquad (1)$$

The dropped variables at each specific reduction step received identical importance ranks, from 30 (least important variables) to 1 (most important variable). Importance ranks were determined with step-wise variable reduction because model-based importance of a given input variable can substantially change when some correlated variables occur more frequently than others. Otherwise, using CUBIST importance measure on the entire spectrum would confound the importance of relevant but highly correlated variables. Since RFE is a wrapper method of variable selection, external test sets (resampling) was needed to exclude selection bias in estimating subset performance (RMSE) (Kuhn and Johnson, 2013). For this purpose, we nested another inner layer of resampling for RFE within the five times repeated 10-fold CV scheme. Importance ranks of variables and outer test RMSEs were averaged from the 50 CV folds. To decrease computation time, we conducted the RFE with 5 CUBIST committees. The RFE procedure and the resampling setup is explained further in the appendix 3.2.1.

2.6 Local soil estimation for plot-level monitoring

We defined a local soil estimation scenario where a new long-term monitoring site was initiated at time zero (t_0). Each one of the 71 NABO sites was assumed to be novel while the remaining ones were established with spectral and reference data records. We therefore conducted 71 separate sample selections from the SSL, each yielding different transfer subsets of the SSL, to test spectral-based soil monitoring using the Swiss mid-IR SSL presented here.We calibrated models at each site using two local samples per given site and a relevant subset of the remaining Swiss SSL (see description below).

The two local samples were chosen from pooled samples at t_0 (first two out of maximum four replicates), or in addition at t_1 if there was only one sample in t_0. Figure 2 illustrates the local modeling workflow. All other samples per given site besides the two chosen during calibration (in other words the successive time-series measurements at a monitoring plot) were used as local validation samples (N_{site}). The respective samples from the remaining SSL included spectra and reference measurements from all BDM samples and NABO samples, excluding the ones from the respective target site. We used only two calibration samples per NABO site to capture the predictive mechanisms at site-level because we wanted to avoid overoptimistic local assessment; both local calibration and validation samples were repeated soil measurements, and are otherwise — if not adequately handled in the calibration sampling strategy — at risk of over-fit when soils' composition and relevant properties show constant trends over time.

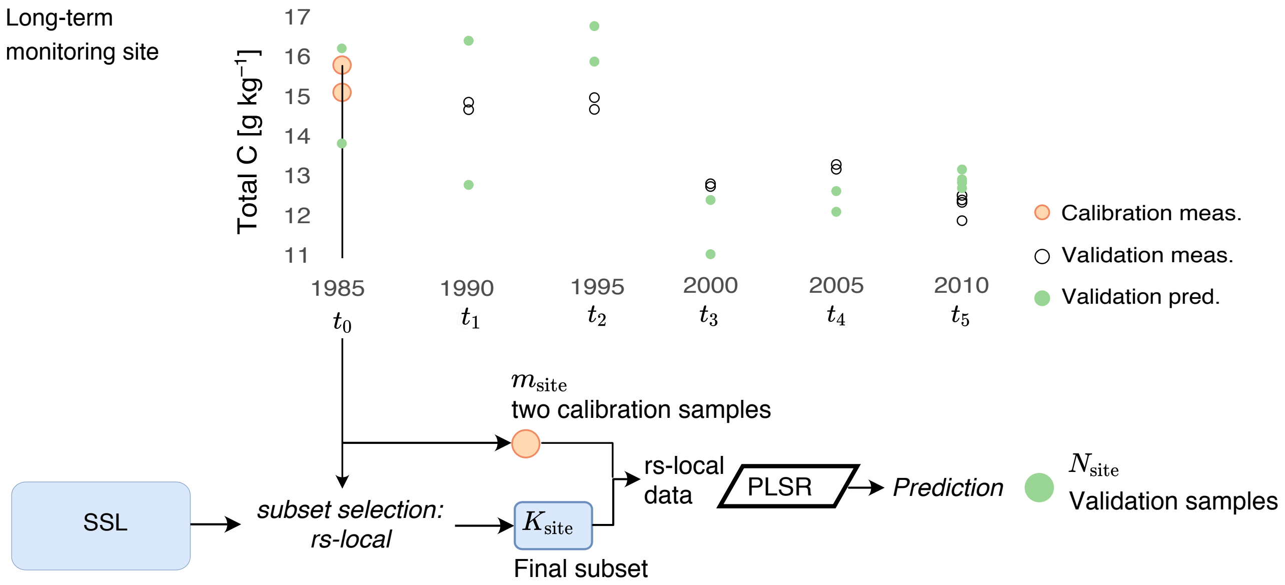

Figure 2. Conceptual scheme illustrating the transfer of the soil spectral library (SSL) to a long-term monitoring site using the RS-LOCAL approach. The local calibration samples and a subset of the SSL are used to calibrate a partial least squares regression (PLSR) model, which predicts the local validation samples.

For each of the 71 sites, the spectral relevant samples from the remaining Swiss SSL were selected using the RS-LOCAL algorithm (see Lobsey et al. (2017)). The site-specific samples (m_{site}) denote local calibration samples from a NABO plot. The recursive reductions of the initial training data, which determined the finally yielded subsets (K_{site}) were driven by model performance (RMSE) for the two local calibration samples. For each NABO site, the corresponding K_{site} set was spiked with the two local calibration samples. On this combined $m_{\text{site}} + K_{\text{site}}$ data set, a final PLSR model, locally adapted for the monitoring plot by optimization on the calibration samples, was developed using 10-fold cross-validation. Finally, the local validation spectra (N_{site}) were predicted using the most accurate calibration model.

The search algorithm RS-LOCAL has three empirical parameters to control the samples that are selected for the local transfer from the SSL (Lobsey et al., 2017). Parameter k is both the number of samples drawn from the original and reduced library without replacement and the number of samples of the returned SSL subset. Parameter b is the number of times k samples are randomly drawn from the remaining data at iteration i of the performance-driven library reduction. Parameter r is the proportion of samples, which are consistently in weakest models, that are removed at it each reduction step. The configuration of the RS-LOCAL search was optimized for each NABO site. For each site, we ran separate RS-LOCAL runs, testing a full-factorial combination of empirical parameter sets $k = \{30, 50, 150\}$, $b = \{10, 20, 50\}$, $r = \{0.05, 0.1, 0.2\}$. The RS-LOCAL procedure is based on partial least squares regression (PLSR) (Wold et al., 1983). For the RS-LOCAL tuning during the subset selection procedure and final calibrations, we tested 1 to 10 PLSR components. The finally selected optimal subset per site yielded the smallest RMSE on the two local calibration samples, and was therefore used to predict the local validation samples.

2.6.1 Uncertainty of spectral monitoring uncertainty: CUBIST vs. RS-LOCAL transfer

To compare performance of the CUBIST approach and RS-LOCAL transfer, errors and concordance of both methods were conditionally assessed per individual NABO ($n = 71$) site. For CUBIST, grouped cross-validation hold-outs were used. Thereby, the two respective local calibration samples m_{site} were excluded, so that the test configuration was identical to the local transfer scenario. In addition to the mentioned assessment statistics, the ratio of performance to interquartile distance (Bellon-Maurel et al., 2010) (RPIQ; 75th and 25th percentiles) was used for relative comparisons between the local transfer and rule-based model because it is robust to non-normal and skewed distributions of measured values.

2.6.2 Evaluating the predictive mechanisms behind the local transfer

For each of the 71 statistical transfers at a plot-level, we quantified the similarity between the selected data sources K_{site} (from SSL) and the respective local target domain $\{N_{\text{site}}\}$ (local validation) by multivariate distances across the spectral input variables. The distance of single observations within $\{K_{\text{site}}; N_{\text{site}}\}$ was referenced to the center of all data, which lead to two respective distributions of distance measures for these sets of points and per site. This procedure involved two steps: 1) compress the input data to reduce the "curse of dimensionality" (Bellman, 1961) and be able to discriminate similarity with spectra (with many dimensions, distance to nearest neighbor becomes similar to distance to farthest neighbor); 2) calculate Mahalanobis distance using a robust method (see below; (Varmuza and Filzmoser, 2016), so that the location and scatter were influenced by the main data rather than by atypical observations.

To condense the spectral information over the entire SSL, Savitzky-Golay preprocessed spectra that included all observations with C elemental measurements were mean-centered, scaled, and then transformed by PCA using singular value decomposition. Dimensionality reduction was necessary to avoid computationally singular values during the subsequent calculation of the covariance matrix (for the Mahalanobis distance). The first ten principal components that explained $86.5\,\%$ of the variation in preprocessed spectra were kept for distance calculations. Finally, the Mahalanobis distance of all the observations to their center was computed with robust estimates for both the center and the covariance matrix of the selected PCA scores, using the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984; Hubert and Debruyne, 2010).

3 Results

3.1 Summary of reference measurements

The samples from the Swiss soil monitoring network (NABO) exhibited the highest variability across samples for total C and OC ($n = 592$; Table 1). Organic C ranged from 1 $\mathrm{g\,kg^{-1}}$ to 583 $\mathrm{g\,kg^{-1}}$. The texture of the soils varied considerably. The pH had values between 3.5 and 7.6 and the soils were slightly acidic overall with a median of 5.8. Compared to the NABO data set, the soils from the BDM program covered a wider set ($n = 3723$ for total C) and range of measured soil properties. The measured range of total C for BDM (1 – 583 $\mathrm{g\,kg^{-1}}$) extended further than that for the NABO. The distribution of pH values was similar in the NABO and BDM sets. The BDM data included also the available cations extracted by AAE (see Table 1). The median $\mathrm{CEC_{pot}}$ was almost equivalent to the value of the NABO sites (24 vs. 23 $\mathrm{cmol(+)\,kg^{-1}}$). Exchangeable Ca showed the largest coefficient of variation (CV = 1.56) among the measured properties of the BDM set. All soil properties except pH and $\mathrm{CEC_{pot}}$ were positively or neutrally (sand) skewed, for both NABO and BDM data sets, respectively.

3.2 General soil estimation with CUBIST modeling

For most of the properties, minimal cross-validated errors were achieved with 100 committees and 9 neighbors. The rule-based models explained a large proportion of the variation ($R^2 > 0.9$) in properties that typically have a strong link to total C (organic C, N) (Table 3; Figure 3). Clay was accurately estimated (RMSE = 47 $\mathrm{g\,kg^{-1}}$; RPIQ = 3.0; range = 0–602 $\mathrm{g\,kg^{-1}}$), whereas sand and silt were less accurately estimated. The pH was accurately estimated (RMSE = 0.3; RPIQ = 6.5). Our models discriminated a large proportion in the measured variation of Ca and Mg (ammonium acetate-EDTA) in the mid-IR (R^2 = 0.97 and 0.79; RPIQ = 2.4 and 1.2). Reference values of potential cation exchange capacity ranged from 0 – 136 $\mathrm{cmol(+)\,kg^{-1}}$ and were modeled with an RMSE of 7 $\mathrm{cmol(+)\,kg^{-1}}$ ($R^2 = 0.72$; RPIQ = 2.0). However, the extractable nutrients P, K, Cu and Zn were insufficiently explained by mid-infrared spectral rules ($R^2 = 0.05 – 0.1$; RPIQ = 0.4 – 0.9). Nonetheless, the rules achieved nearly unbiased property estimates over all measurements. We found marginal local bias at the largest values, mostly for variables with positively skewed distributions such as total C (Table 3; Figure 3).

Overall, out of the 16 available soil properties, total C, total N, total $\mathrm{CaCO_3}$, Ca and Mg (ammonium acetate-EDTA), OC, $\mathrm{CEC_{pot}}$, pH, sand and clay (10) were modeled with relatively good discrimination capacity in the measured ranges (Figure 3).

3.2.1 Model interpretation and filtering with variable importance

Figure 4 shows that the test RMSE of total C first slightly decreased and then steadily increased from all (209) to less spectral variables using CUBIST and RFE. The lowest error ($\mathrm{RMSE_{test}}$ = 8.10 $\mathrm{g\,kg^{-1}}$ total C) of spectroscopic estimation was achieved with the spectra with 105 variables. For the subsequent variable reduction steps, model performance steadily dropped until one wavenumber was left ($\mathrm{RMSE_{test}}$ = 18.8 $\mathrm{g\,kg^{-1}}$ total C).

The spectral feature between 1786 $\mathrm{cm^{-1}}$ and 1754 $\mathrm{cm^{-1}}$ was the most important one for the estimation of total C with CUBIST (Figure 4). The twelve spectral variables with the best importance ranks across all RFE iterations and test sets de-

Table 1. Summary statistics of the measured soil properties of the Swiss soil spectral library derived from the sample archive of the Swiss soil monitoring network (NABO) and the Swiss Biodiversity Monitoring (BDM) program. Total C = total carbon, OC = organic carbon, CEC$_{pot}$ = potential cation exchange capacity.

	n	#Locations	Min	Max	Median	Mean	SD	Skewness
NABO								
Total C [g kg^{-1}]	572	71	11	273	33	40	35	3.84
OC [g kg^{-1}]	592	71	11	273	30	37	34	4.03
N [g kg^{-1}]	572	71	1.1	19.9	3.2	3.6	2.6	3.12
pH	574	71	3.5	7.6	5.7	5.8	0.9	0.09
Clay [g kg^{-1}]	80	55	30	590	220	231	105	0.73
Silt [g kg^{-1}]	81	55	150	800	380	397	125	0.95
Sand [g kg^{-1}]	80	55	40	820	400	371	166	0.01
CEC$_{pot}$ [cmol(+) kg^{-1}]	121	58	7	136	23	26	17	4.10
BDM								
Total C [g kg^{-1}]	3723	1079	1	583	41	55	49	4.04
OC [g kg^{-1}]	3652	1068	0	583	37	50	48	4.55
N [g kg^{-1}]	3724	1079	0.0	26.4	3.2	3.8	2.5	2.91
pH	3765	1094	2.6	8.0	5.6	5.6	1.3	-0.10
CaCO$_3$ [g kg^{-1}]	1565	455	0	885	36	107	144	1.80
Clay [g kg^{-1}]	787	785	5	602	194	213	108	0.71
Silt [g kg^{-1}]	787	785	105	708	303	309	95	0.74
Sand [g kg^{-1}]	787	785	5	817	419	411	168	-0.08
CEC$_{pot}$ [cmol(+) kg^{-1}]	674	190	0	94	24	26	12	1.39
P (AAE) [g kg^{-1}]	417	417	1	1047	19	40	77	7.99
K (AAE) [g kg^{-1}]	417	417	22	1255	106	136	115	4.06
Ca (AAE) [mg kg^{-1}]	417	417	141	96250	3927	12226	19127	2.20
Mg (AAE) [mg kg^{-1}]	417	417	16	3196	161	232	259	5.35
Cu (AAE) [mg kg^{-1}]	417	417	2	73	6	8	5	5.32
Zn (AAE) [mg kg^{-1}]	417	417	1	131	4	6	9	8.52
Fe (AAE) [mg kg^{-1}]	417	417	84	1640	342	387	194	1.93

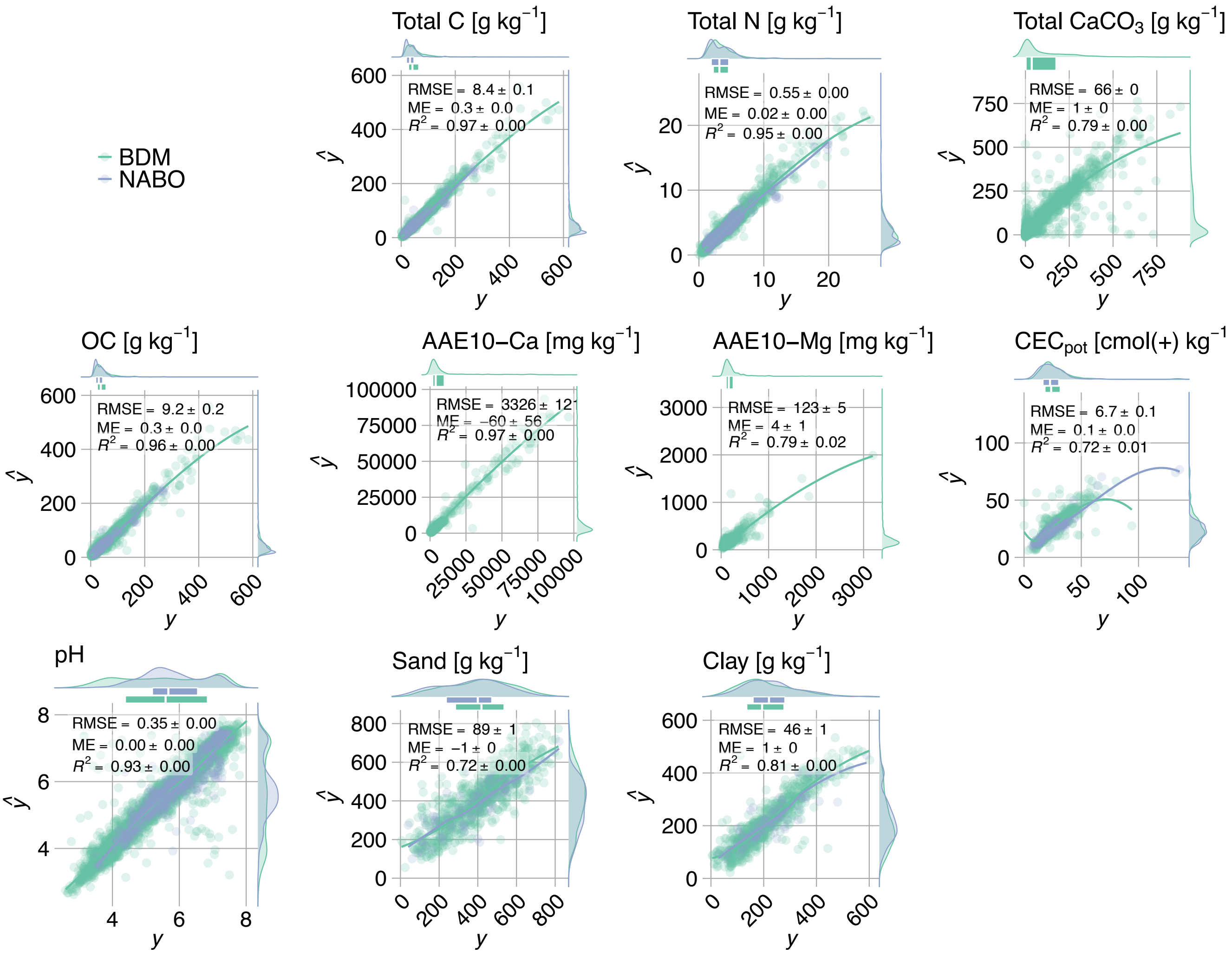

Figure 3. Agreement between measured and mid-infrared predicted values that was obtained from CUBIST models. Models' performance was assessed by site grouped cross-validation holdouts (five times repeated ten-fold). The lines obtained with local regression (LOESS) smoothing indicate the trends in predictions. Models of soil properties with $R^2 \geq 0.72$ are shown (see Table 3 for more detailed model summaries).

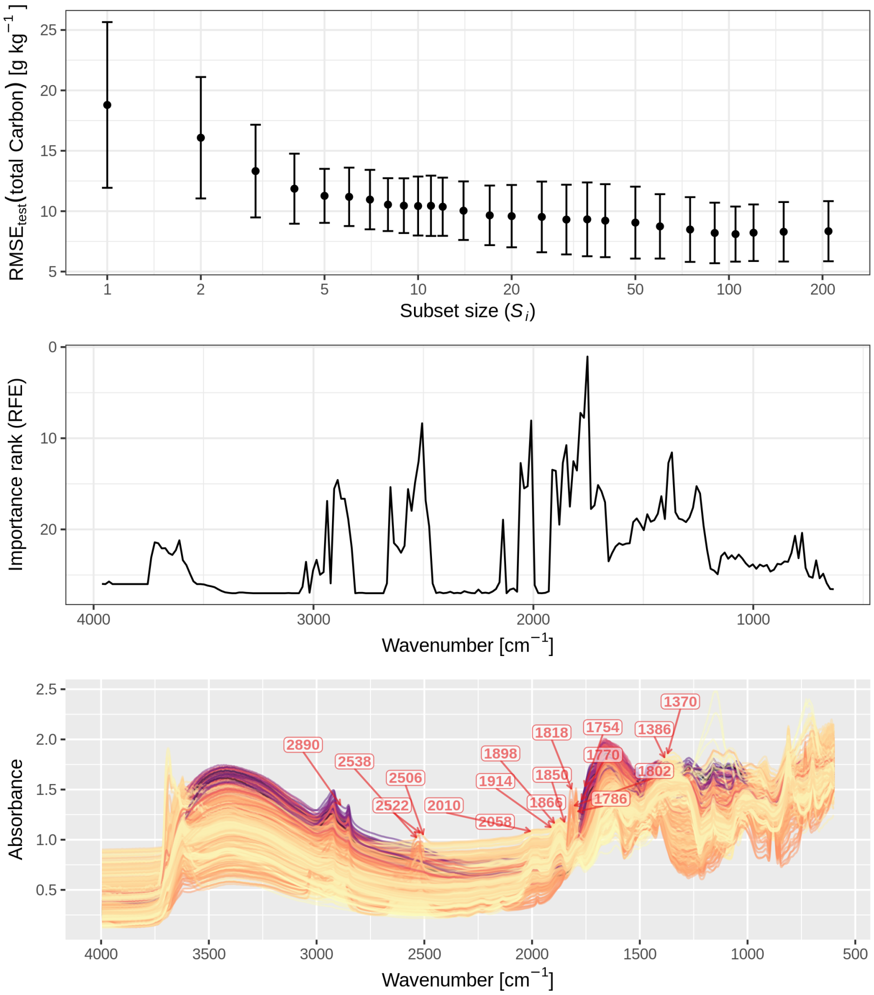

Figure 4. *Top:* Root mean square error (RMSE) of mid-infrared estimates of total C that CUBIST produced at the respective subsets of spectral variables. The performance profile was obtained with a recursive feature elimination (RFE) procedure. The error bars represent the standard deviations of the test RMSE derived with nested cross-validation ($n = 50$). *Middle:* Average importance ranks across the spectrum. Lower rank values indicate higher importance for the estimation of total C. Ranks were determined with RFE. *Bottom:* Mid-infrared absorbance spectra of the Swiss soil spectral library ($n = 4295$; with corresponding total carbon (C) measurements determined by dry combustion). The unprocessed absorbance spectra are annotated with the 17 most influential spectral variables (wavenumbers) in the CUBIST model (average importance rank < 15); these had the highest mean importance ranking determined by the recursive feature elimination procedure.

Table 3. Cross-validated mid-IR estimates of 18 soil properties derived from rule-based CUBIST models that were developed using all available data. The 10-fold cross-validation procedure was grouped by the site and repeated 5 times to achieve many test-train data combinations and provide realistic model assessment with generalization. #C and #N denote the number of committees and neighbors for CUBIST. Total C = total carbon , OC = organic carbon, N = total nitrogen, and CEC$_{pot}$ = potential cation exchange capacity.

	n	Range	#C	#N	#Rules	RMSE	ME	SDE	R^2	RPIQ
Total C [g kg^{-1}]	4295	1–583	100	9	6–26	8.4 ± 0.1	0.3 ± 0.0	8.4 ± 0.1	0.97 ± 0.00	4.3 ± 0.0
OC [g kg^{-1}]	4244	0–583	100	9	4–24	9.3 ± 0.2	0.0 ± 0.0	9.2 ± 0.2	0.96 ± 0.00	3.4 ± 0.1
N [g kg^{-1}]	4296	0.0–26.4	100	9	9–21	0.55 ± 0.00	0.00 ± 0.00	0.55 ± 0.00	0.95 ± 0.00	4.3 ± 0.0
pH	4339	1.0–8.0	100	9	4–18	0.3 ± 0.0	0.0 ± 0.0	0.3 ± 0.0	0.93 ± 0.00	6.5 ± 0.0
CaCO$_3$ [g kg^{-1}]	1565	0–885	10	9	1–5	6.6 ± 0.0	0.1 ± 0.0	6.6 ± 0.0	0.79 ± 0.00	2.6 ± 0.0
Clay [g kg^{-1}]	867	5–602	100	9	2–6	47 ± 1	1 ± 0	47 ± 1	0.81 ± 0.00	3.0 ± 0.0
Silt [g kg^{-1}]	868	100–800	100	9	1–14	71 ± 1	−0 ± 0	71 ± 0	0.51 ± 0.01	1.7 ± 0.0
Sand [g kg^{-1}]	867	5–820	100	9	1–6	89 ± 1	−1 ± 0	89 ± 1	0.72 ± 0.00	2.8 ± 0.0
CEC$_{pot}$ [cmol(+) kg^{-1}]	795	0–136	50	9	1–10	7 ± 0	0 ± 0	7 ± 0	0.72 ± 0.01	2.0 ± 0.0
P (AAE) [mg kg^{-1}]	417	1–1047	50	9	1–8	77 ± 1	−3 ± 1	77 ± 1	0.05 ± 0.00	0.4 ± 0.0
K (AAE) [mg kg^{-1}]	417	1–1255	100	9	1–12	111 ± 2	−3 ± 1	111 ± 2	0.10 ± 0.01	0.8 ± 0.0
Ca (AAE) [mg kg^{-1}]	417	1–96250	100	9	6–14	3326 ± 121	−60 ± 56	3325 ± 121	0.97 ± 0.00	2.4 ± 0.1
Mg (AAE) [mg kg^{-1}]	417	1–3196	100	9	1–13	123 ± 5	4 ± 1	123 ± 5	0.79 ± 0.02	1.2 ± 0.0
Cu (AAE) [mg kg^{-1}]	417	1–73	50	9	1–7	5 ± 0	−0 ± 0	5 ± 0	0.10 ± 0.01	0.9 ± 0.0
Zn (AAE) [mg kg^{-1}]	417	1–131	100	9	1–13	9 ± 0	−0 ± 0	9 ± 0	0.06 ± 0.01	0.4 ± 0.0
Fe (AAE) [mg kg^{-1}]	417	1–1640	50	5	1–7	167 ± 3	1 ± 1	167 ± 3	0.28 ± 0.02	1.2 ± 0.0

rived from the subset sizes were (starting with best): 1754 cm^{-1} $(\mathrm{mean(rank)} = 1.04)$, 1786 cm^{-1} , 1770 cm^{-1} , 2010 cm^{-1}, 2506 cm^{-1}, 1850 cm^{-1}, 1370 cm^{-1}, 2522 cm^{-1}, 1818 cm^{-1}, 1866 cm^{-1}, 2058 cm^{-1}, 1386 cm^{-1} $(\mathrm{mean(rank)} = 12.7$; Figure 4).

3.3 Accuracy of the local transfer models compared to the general model

For the example site *65 COR*, the best performance of RS-LOCAL was achieved with 55 samples from the SSL (K), 10 sampling events (B) of size K at each iteration, and 10 % reduction (r) at each iteration (Figure 5). Therefore, 55 transfer samples from the SSL were combined with two site calibration samples previously used to supervise the selection from the data source, to form a PLS regression calibration model for the estimation of the site validation samples (see Figure 5 panel *a* right). Compared to the target observations from the site (right part of panels *a* and *b*; measured range = 11.9–16.0 g kg^{-1} C), the selected instances were heterogeneous with regard to their characteristic patterns in raw spectra, their preprocessed feature space, and their measurements (range = $8.7–97.7\,\mathrm{g\,kg^{-1}}$ C). The selected instances covered a significant proportion of the first two components in the feature space of the entire SSL.

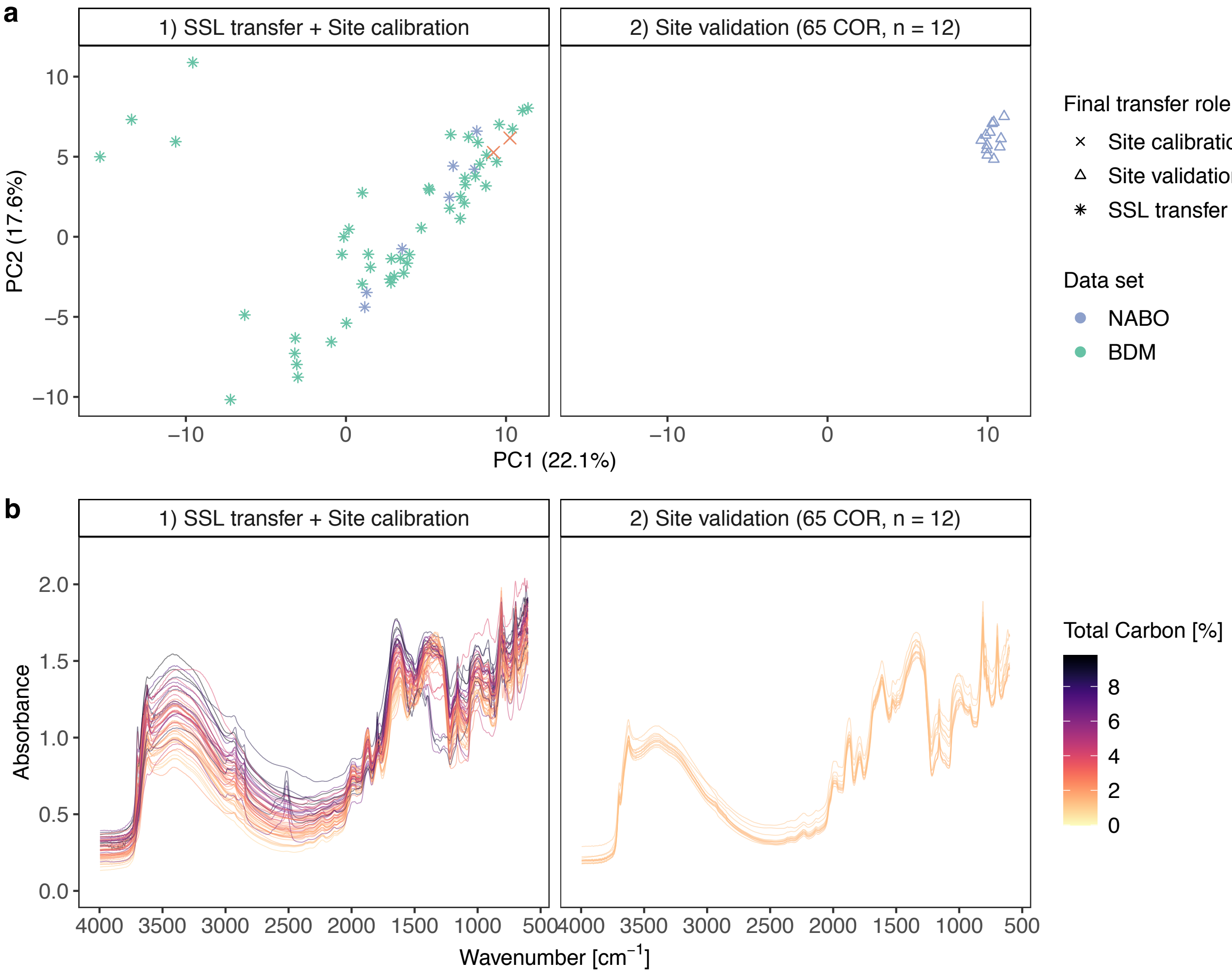

Figure 5. Illustration of the site-specific transfer modeling of total carbon (C), using RS-LOCAL for the example site *65 COR* of the Swiss soil monitoring network (NABO). Panel *a* contains the principal components subspace (PC1 and PC2) of the Savitzky-Golay first derivative mid-IR spectra, and panel *b* outlines the corresponding absorbance spectra (unprocessed for illustration), which are coloured by the total C content. The left subplots show the SSL transfer samples ($n = 55$) that were selected from the soil spectral library (n = 4281; excluding all NABO calibration samples). This subset was most accurate when predicting the two calibration samples under the mechanisms RS-LOCAL and their optimal tuning configuration for the site ($\{K = 50; B = 10; r = 0.1\}$). The right panels shows the time series data for the validation samples of the NABO site "65 COR".

The RMSE on the site validation samples ($\mathrm{RMSE}_{N_{\mathrm{site}}}$) at the final subsets varied between 0.01 $\mathrm{g\,kg^{-1}}$ C and 10.73 $\mathrm{g\,kg^{-1}}$ C and for all tuning parameter combinations and sites, and between 0.01 and 3.02 $\mathrm{g\,kg^{-1}}$ C for the best subsets per site (Figure A1).

The local approach reduced the error of the rule-based approach on average by factor 4.4 (Figure 6; $\mathrm{mean}(\mathrm{RMSE}_{\mathrm{rs-local}}) = 0.7\,\mathrm{g\,kg^{-1}}$ C $\mathrm{mean}(\mathrm{RMSE}_{\mathrm{cubist}}) = 3.1\,\mathrm{g\,kg^{-1}}$ C). The local transfer was more accurate for the majority of NABO sites (69 out of 71 sites). The linear dependency between modeled and measured values was higher for the local transfer compared to the general model (53 out of 71 sites). Moreover, RS-LOCAL produced on average 1.3 times less biased estimates of total C per site for 52 out of 69 sites in terms of absolute values ($|\,\mathrm{ME}\,| = 0.1\,\mathrm{g\,kg^{-1}}$ C vs. $0.5\,\mathrm{g\,kg^{-1}}$ C). The ratio of performance to inter-quartile distance (RPIQ) confirmed that local learning in the mid-infrared was able to better discriminate developments of total C over time, relative to its measured distribution. Overall, local learning with two local calibration samples and targeted SSL selections allowed for better estimations than the generic CUBIST approach on average (RPIQ = 3.08 vs. 1.00; RPIQ larger for 66 out of 71 sites). Across all validation data points of the NABO set, the RS-LOCAL transfer was 5.6 times more accurate for total C than the general rules in terms of RMSE and RPIQ (RMSE = 0.9$\mathrm{g\,kg^{-1}}$ C; RPIQ = 31.7)

3.3.1 Predictive mechanisms behind the local transfer

The samples used for the transfer process (RS-LOCAL data) of the example site *COR 65* showed high spectral dissimilarity along the first 2 PCs, explaining 39.8 % of preprocessed spectral variance (Figure 5). Compared to the entire SSL with total C measurements available (the source domain prior selection; range PC1: -41.4 to 13.0; range PC2: -19.0 to 30.0), the selected transfer samples of this site occupied a region of major variation in the PC space (range PC1: -15.4 to 11.4; range PC2:-10.2 to 10.9). The two local calibration samples and the 12 validation samples on the upper right corner were close to each other in the PC1-PC2 subspace (Figure 5, panel *a*, *left* and *right*; range PC1: 9.2 to 11.0; range PC2: 4.9 to 7.5). Not only the absorbance spectra but also the corresponding C reference values were highly variable compared to the exemplary NABO site (Figure 5, panel *b*; 7.3–117.8 $\mathrm{g\,kg^{-1}}$ C for $K_{\mathrm{rs-local}}$, and 11.9–16.0 $\mathrm{g\,kg^{-1}}$ C for the plot of this site). This particular target monitoring site indicated that RS-LOCAL selected soils from the SSL with a relatively large spectral diversity and a wide range of total C.

The instances selected by RS-LOCAL filled a substantial proportion of the SSL's feature space (Figure 7), confirming the trend of site *65 COR*. We found that RS-LOCAL yielded a quite wide selection of relevant samples from the SSL with reference to both the total C range and a wide coverage of spectral features expressed with robust multivariate locations. The spectral estimations of the site validation sets that resulted from RS-LOCAL-based transfers did neither show trends in the mode or spread for distributions of C measurements nor in the ones from their spectral distances. The measured distributions of K_{site} SSL subsets and N_{site} local validation samples for further key soil properties related to the chemical composition (OC, pH, $\mathrm{CEC_{pot}}$, clay and $\mathrm{CaCO_3}$) were also markedly different, confirming the local transfer of quite heterogeneous soils (Table 5). For example, standard deviations of the 0 %, 25 %, 50 %, and 75 % percentile differences between the transfer sets selected the SSL and the samples from the respective NABO site were on average between 18 $\mathrm{g\,kg^{-1}}$ and and 66 $\mathrm{g\,kg^{-1}}$ for measured C and OC, respectively. Further, the measured clay and $\mathrm{CaCO_3}$ contents were markedly different between the RS-LOCAL selection and the local validation sets (mean absolute median differences of 85 $\mathrm{g\,kg^{-1}}$ clay and 89 $\mathrm{g\,kg^{-1}}$ $\mathrm{CaCO_3}$). This

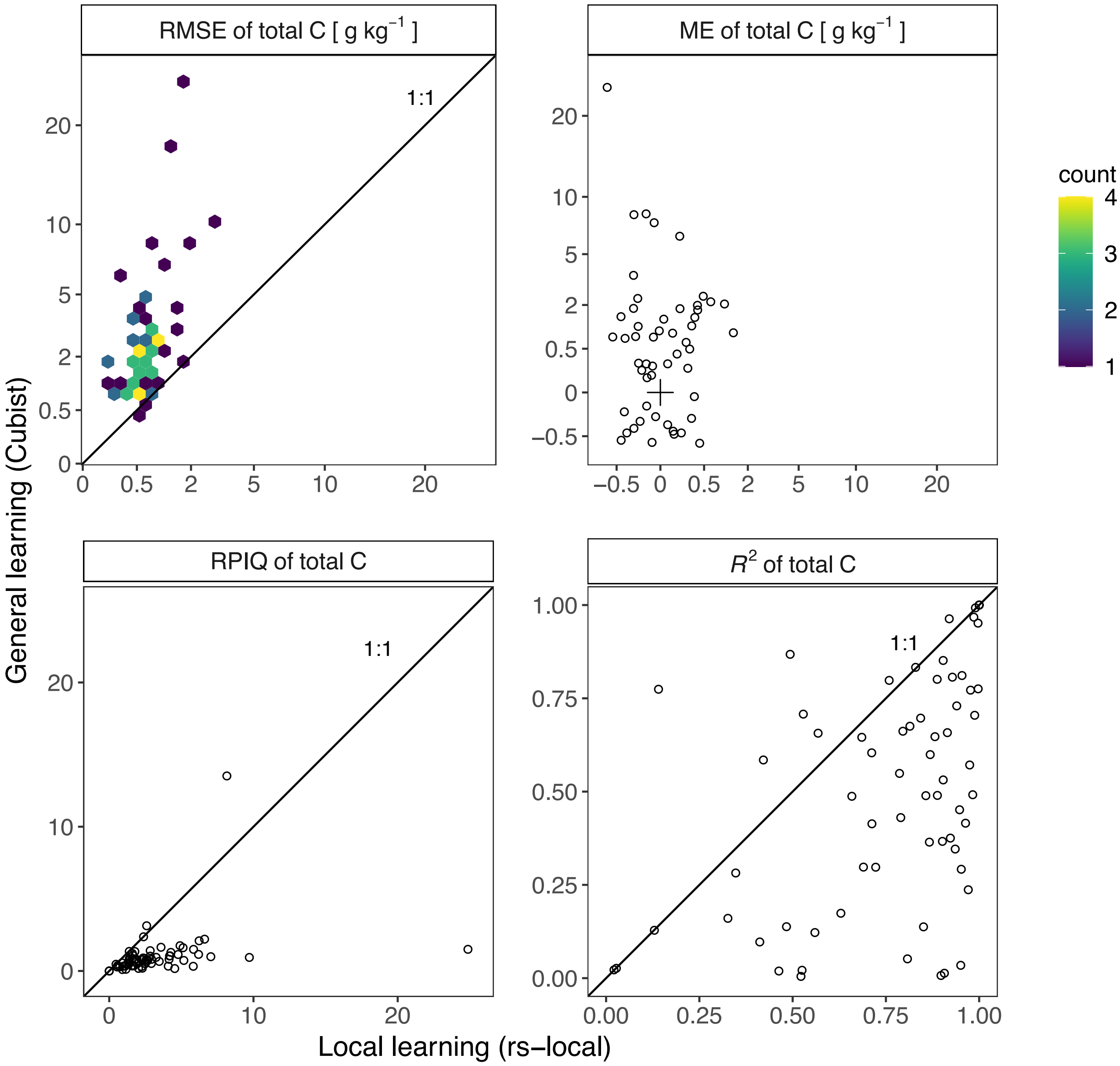

Figure 6. Model assessment of the estimated total carbon (C) of 71 NABO sites for the general learning with CUBIST (y-axis) vs. local learning transfer with RS-LOCAL (x-axis). The four panels depict the root-mean-square-error (RMSE), the mean error (ME), the ratio of performance to interquartile distance (RPIQ) and R^2. The 1:1-line emphasizes the difference between the two approaches.

Table 5. Standard deviations (SD) of the absolute differences of percentiles ($P_0, P_{25}, P_{50}, P_{75}, P_{100}$) of final RS-LOCAL subsets (K_{site}) and corresponding site validation samples (N_{site}) the across 71 long-term monitoring sites. The aggregated values for six measured soil properties are shown. Total C = total carbon, OC = organic carbon, CEC$_{\text{pot}}$ = potential cation exchange capacity.

	$\text{SD}\left(\lvert P_X(K_{\text{site}}) - P_X(N_{\text{site}})\rvert\right)$				
	$\text{SD}\left(\lvert\Delta P_0\rvert\right)$	$\text{SD}\left(\lvert\Delta P_{25}\rvert\right)$	$\text{SD}\left(\lvert\Delta P_{50}\rvert\right)$	$\text{SD}\left(\lvert\Delta P_{75}\rvert\right)$	$\text{SD}\left(\lvert\Delta P_{100}\rvert\right)$
Total C [g kg^{-1}]	18	22	25	31	61
OC [g kg^{-1}]	25	23	22	22	66
pH	0.8	0.7	0.5	0.6	0.9
CEC$_{\text{pot}}$ [cmol(+) kg^{-1}]	9.8	10.3	10.6	10.6	23.1
Clay [g kg^{-1}]	110	97	85	68	90
$CaCO_3$ [g kg^{-1}]	74	79	89	98	178

findings correspond with the dissimilar selection compared to the local target samples found in the PCA space of preprocessed spectra.

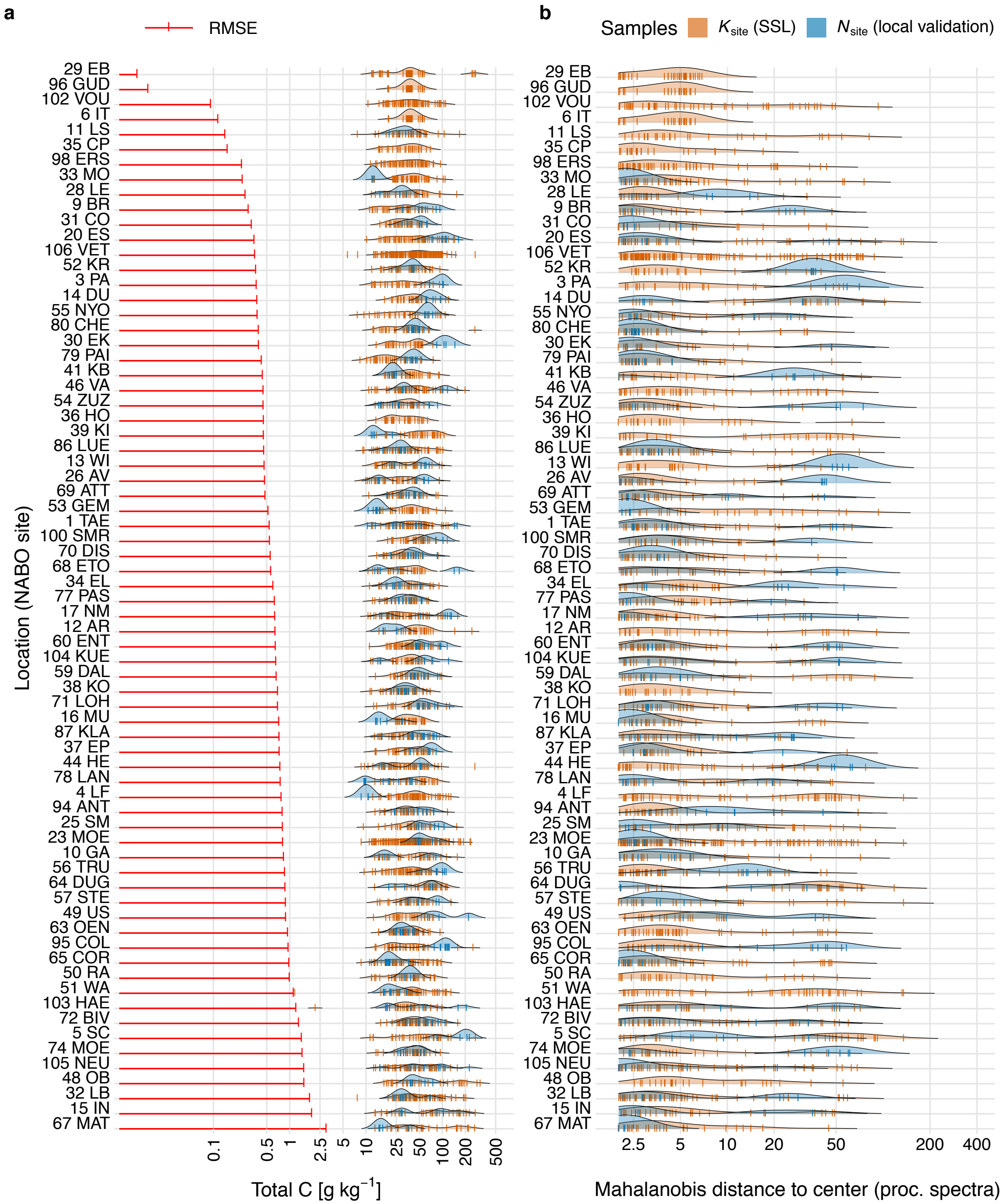

Figure 7. Analyzing the mechanisms behind the individual adaptive transfer realized with RS-LOCAL. Panel *a*: The left horizontal bars show the root mean square error (RMSE) of mid-infrared predictions of the temporal validation set N_{site} of time series of total carbon (C) for each of the 71 NABO sites, which was calculated without the two respective calibration samples. The blue density plots depict the distribution of the site-specific validation samples, and the brown vertical bars show the measured values of C for the final subsets of SSL used for the transfer (K_{site}). Panel *b*: The distribution of the robust distances from the PCA center of Savitzky-Golay preprocessed spectra of the entire soil spectral library compared to the subset of instances involved in the individual transfer modeling (K_{site}) and validation samples (N_{site}) (similarity in site-specific vs. final RS-LOCAL selection), computed with the Minimum Covariance Determinant (MCD) estimator.

4 Discussion

4.1 General soil estimation with the Swiss SSL

Many of the chemical properties with distinct links to soil organic matter and the key minerals (e.g., clays and quartz) were discriminated well with mid-IR CUBIST models (Table 3; Figure 3). Specifically, the models estimated total C, OC, N, pH, texture, AAE10-Ca, and AAE10-Mg with $R^2 > 0.7$. This suggests that the majority of developed models are useful for applications that require soil proxies in order to manage land resources. For example, CEC$_{\text{pot.}}$ (RMSE = 7.0 $\mathrm{cmol(+)\,kg^{-1}}$), as well as pH (RMSE = 0.3), have high ecological importance for nutrient availability in ecosystems. In agriculture, both measures are key factors for soil fertility and nutrient recommendations.

The accuracy of our estimates for the properties that have direct chemical links, through compound-associated absorptions, were mostly comparable to established continental or country-specific mid-IR SSLs. For example, Clairotte et al. (2016) achieved RMSE = 2 $\mathrm{g\,kg^{-1}}$ for OC using mid-IR and the spectrum-based learner for local predictions, while Sila et al. (2016) reported RMSE = 4 $\mathrm{g\,kg^{-1}}$. The accuracy of our general OC estimates was lower (RMSE = 9.3 $\mathrm{g\,kg^{-1}}$; RPIQ = 3.4), which we explain by the relatively large range of measured values and variable mineralogy (Stenberg and Rossel (2010)). We found that total C had more CUBIST rules per committee than OC (Table 3), indicating total C, which also included inorganic C (mostly $\mathrm{CaCO_3}$), leverages more chemical constituents and latent absorptions for its estimation. In spite of lower parsimony, slightly more accurate estimates of total C were achieved (RMSE = 8.4 $\mathrm{g\,kg^{-1}}$; RPIQ = 4.3).

The majority of soil properties were most accurately estimated with the maximum tested 100 committees and 9 neighbors. Instance-based correction with similar data in the the training set, (nearest-)neighbors, yielded considerably higher accuracy for total C (e.g., RMSE = 8.9 $\mathrm{g\,kg^{-1}}$ for 20 committees and 2 neighbors vs. RMSE = 8.1 $\mathrm{g\,kg^{-1}}$ for 20 committes and 9 neighbors; model evaluation across cross-validation folds; results not shown). The number of rules give a first proxy for model complexity and the complementary of spectral features that are involved in prediction. The range in number of rules across the ensembles was widest for total C (6 – 26), similar for OC (4 – 24), medium wide for CEC$_{\text{pot}}$ (1 –10), and very narrow for $\mathrm{CaCO_3}$ (1 – 5) to give specific examples. Viscarra Rossel and Webster (2012) report comparably less rules (medium: 21; range all properties: 5 – 64) for OC and relatively similar number of rules for CEC (15). Nonetheless, such comparisons have to be done with care because the NIR range has a less pronounced representation of functional groups than the mid-IR range, and because temperate soils have fundamental differences in chemical composition compared to more weathered tropical soils. For our mid-IR SSL, we were surprised that the rules for OC were similarly complex as the ones for total C; in fact, we also could not find any clear partitioning in the rules with respect to measured ranges and spectral patterns (exploratory analysis not shown) in contrast to Viscarra Rossel and Webster (2012). In fact, this is different from the general patterns found by Viscarra Rossel and Webster (2012), where the rules clearly partitioned the data into distinct measured distributions. Last but not least, the diversity in rules for total C and OC of the general estimation approach makes the soil diversity selected from the library and we found for site-specific local transfer even less exotic (see section 4.2).

The variable importance assessment of the spectroscopic models revealed five major regions of features with particularly high predictive influence for total C: 2890 $\mathrm{cm^{-1}}$, 2522 $\mathrm{cm^{-1}}$, 2010 $\mathrm{cm^{-1}}$, 1754 $\mathrm{cm^{-1}}$, and 1370 $\mathrm{cm^{-1}}$ (Figure 4). We attribute the two absorption peaks near 2890 cm^{-1} to C$-$H stretching vibrations of organic matter (Skjemstad and Dalal, 1987), which were also relatively important for estimating C in other studies (e.g., Janik and Skjemstad (1995); Viscarra Rossel and McBratney (1998)). The important variable at 2522 cm^{-1} is indicative of C$=$O absorption due to the carbonyl group present in carbonates (e.g., calcite) (Nguyen et al., 1991; Soriano-Disla et al., 2014). The three important absorptions between 2010 cm^{-1} and 1786 cm^{-1} result from three consecutive Si-O-Si (overtone and combination) absorptions, which are indicative of quartz. However, the most important absorptions near 1754 cm^{-1} showed no distinct peak but an edge feature. This is in accordance with Sila et al. (2016), which identified this region as being most relevant for estimating total C with a (general) random forest model developed from the SSL of the African Soil Information project. This region is close to the C$=$O stretching vibration of the carboxyl group that occurs around 1725 – 1720 cm^{-1} (Madari et al., 2006), which is further confirmed by the high importance of these vibrations found by Janik and Skjemstad (1995). The last relatively important region around 1370 cm^{-1} was also an edge feature with no distinctly visible peak of chemical group assigned, which, however, might be influenced by the adjacent carboxylate (COO^{-}) or $\breve{}-\mathrm{CH}$ absorptions at 1400 – 1350cm^{-1} of aliphatic compounds such as humic acids (Madari et al., 2006; Parikh et al., 2014). In summary, the CUBIST-RFE variable importance analysis enabled us to link characteristic absorptions of typically prominent functional groups of soil organic and inorganic C compounds, and as well quartz absorptions as indirect correlative features of predictive relevance, with our general-model based estimates of total C.

Because the rule-based models we developed can estimate ten soil properties reasonably well ($R > 0.6$; RPIQ > 2.0; Figure 3), the Swiss SSL will be useful for new soils when new reference measurements for model adaptation are relatively scarce or not available. Thereby, the Swiss SSL will be cost and time efficient for characterizing soils of similar composition in the near future. The new predictions can further be augmented with straightforward model interpretation, which allows chemical inference of pedological aspects to provide means of model applicability. Although the combined BDM and NABO set comprises a large soil variability in Switzerland, the diversity of subsoils at depths greater than 20 cm — mostly in terms of the mineral composition — as well as peat and forest soils are probably not yet represented sufficiently in the SSL. We therefore must continuously update the present SSL with more and deeper soil horizons in the near future.

4.2 Local transfer from the SSL for soil monitoring at plot-scale

The local estimates of total C that were derived with RS-LOCAL selection were substantially better on average (RMSE = 0.7 $\mathrm{g\,kg^{-1}}$ C) as those derived using all of the data and general CUBIST models (RMSE = 3.1 $\mathrm{g\,kg^{-1}}$ C; Figure 6). The data-driven estimation at plot-scale further considerably reduced bias and increased R^2 compared to the general CUBIST rules.

Our third goal was to analyze the characteristics of soils that were selected from the SSL and used for establishing locally-adaptive models tailored to the respective long-term monitoring sites. Surprisingly, the RS-LOCAL subsets selected from the SSL had rather dissimilar spectra in the robust PCA space (Figure 5; Figure 7); their distances to the center had a wide distribution compared to the local samples. The K_{site} subsets accordingly covered a large proportion of the spectral input space. The likely dissimilar chemical composition of soils was also reflected in the reference measurements of total C. We conducted a broader analysis to interpret the soil context of the selected samples with further soil compositional covariates (OC, pH, $\mathrm{CEC_{pot}}$, clay, $\mathrm{CaCO_3}$), which also did not resemble the soil characteristics of the local monitoring sites (see Figure

5). These findings together with the accurate validation results clearly indicate that dissimilarity and diversity in soils can also provide the means for fitting locally-adaptive models.

Nevertheless, we can yet only speculate about how and why such diverse calibration sets are able to leverage accurate local calibrations. One hypothesis is that by increasing the range and variability in spectral variables and measurements a model can become quite stable in the central range of local refererence measurements because a larger range of input variables is considered; thereby, the RS-LOCAL subsets that are selected from the SSL and used for PLS regression would stabilize and reduce the errors of the local samples. We imagine that we leverage a similar mechanism as in simple linear regression, where narrowing the range of the independent variable (x) in the training samples would decrease the accuracy of intermediate values of the independent variable. We therefore need to further look into the details of spectral dissimilar learning, for example, also investigating the relevance of specific spectral features for local spectral transfers. The inherent working principles of RS-LOCAL are in contrast to the spectrum-based learner (SBL) or other forms of memory-based learning that utilize similar samples to infer sample-specific predictions based on existing training data (Lin and Vitter, 1994; Ramirez-Lopez et al., 2013). Our approach could describe a data-driven phenomenon, which implies that spectra can help to estimate a set of unrelated new soils. Another possibility is that there is in fact a pedological explanation that could be elucidated with more soil covariates such as mineralogy.

Local soil characterization is simpler, quicker and cheaper when a large proportion of properties of new soils are estimated by spectroscopy. Our results suggest the importance of optimizing the transfer of relevant information present in large SSLs to minimize the required amount of conventional laboratory analyses of new soils. Soil chemical and physical heterogeneity can be substantial in large SSLs. Therefore, such data variation can be beneficial for future predictions of properties of soils. However, learning a single general model over a heterogeneous training set, and obtaining parameter estimates optimized with a global measure of goodness of fit can introduce bias and inaccuracy to local (soil) estimation (Hand and Vinciotti, 2003; Ramirez-Lopez et al., 2013; Lobsey et al., 2017). Although the highest estimation accuracy could be achieved with only soils of the target study area (Stenberg and Rossel, 2010; Guerrero et al., 2016), it is impractical and inefficient to derive a single spectral prediction model with those. It requires 1) a large volume of reference measurements for a reasonably accurate multivariate calibration, and 2) it does not utilize already existing soil information.

Currently, the Swiss long-term soil monitoring uses a spatially representative sampling and then bulking the soils into four replicates for reference measurements (Desaules et al., 2010; Gubler et al., 2019). When the long-term monitoring would be augmented with mid-IR spectroscopy, one could make spectral measurements on all subsamples, rather than only on bulked samples, which would deliver spatially-explicit information and reduce nuisance factors from different sampling conditions. If not constrained economically (separate drying, sieving, and milling of sub-samples), a spectral workflow could thus allow to account for small-scale soil variability and reduce bias in measurements to robustly estimate temporal soil changes. For example, there is currently a relatively large variability in C measurements between the bulked replicate samples at one time point (Gubler et al., 2019). Our results suggest that unbiased spectral measurements eventually mediate such inconsistencies. Although Gubler et al. (2019) reported only minor changes for the ensemble of permanent cropland or cropland-meadow monitoring sites (30), there were four sites with declining trends and nine sites with increasing trends in OC (-11% to $+16\%$

relative change per decade, respectively). Here, the trend of spectroscopic predictions could be investigated with respect to specific research questions on agronomic management-induced changes, further physicochemical soil characterization (e.g., OC fractions).

Relatively precise and unbiased geographically-local estimates of soil properties from diverse and large SSLs can be achived by a handful of data-driven statistical approaches that are currently popular in the soil science community (Viscarra Rossel and Webster, 2012; Ramirez-Lopez et al., 2013; Guerrero et al., 2014; Lobsey et al., 2017; Tsakiridis et al., 2020). Among the methods, we tested RS-LOCAL Lobsey et al. (2017) in our local soil monitoring scenario. Compared to memory-based learning, such as SBL (Ramirez-Lopez et al. (2013)), RS-LOCAL does not precondition the choice of useful subsets based on similarity in the input dimensions, here spectra, when performing the selection of SSL samples. The RS-LOCAL method is applied to exhaustively sample instances from the SSL without replacement, while it preferably selects those that perform well on the local target set, using PLS regression. An advantage of the method is that it can deal better with erroneous spectra as well as inaccurate and imprecise analytical reference measurements in the SSL, because it filters them as irrelevant instances. Besides chemometric and classical machine learning approaches, convolutional neural networks are being popularized for modeling SSLs with large soil variability (e.g., Liu et al. (2018); Padarian et al. (2019a, b); Tsakiridis et al. (2020)). There seems to be a small performance gain of a multi-output CNN with a similarity-based error correction using neighbors compared to the SBL (Tsakiridis et al. (2020); RMSE = $11.0\,\mathrm{g\,kg^{-1}}$ vs. $11.7\,\mathrm{g\,kg^{-1}}$ for OC). Despite the current development of interpretation methods in deep learning, CUBIST and PLSR modeling employed in both in the SBL and RS-LOCAL offer easier interpretation with comparable accuracy to CNNs.

Transfer learning or local learning introduces a new paradigm to supervised learning: model building is governed by the intended model application and thus coupled to it (Hand and Vinciotti, 2003). This contrasts general-model application, where the inference process is separated from the prediction of new data. Including local samples and their local data characteristics is necessary in order that a combined search and learning algorithm has a chance to capture predictive mechanisms. At the same time, the selection process and the partial data dependence within the predictive unit, the site, requires a careful assessment scheme to prevent a potential selection bias in the assessment of the approach. To account for this, we kept the respective site-specific local tuning and calibration set — whose hold-out performance directed the iterative search process and the reduction of the SSL — at minimum size of two observations at t_0 or in addition t_1 when only one measurement was available from the first sampling (see Figure 2).

4.3 Future applications and updates of the SSL

We found that data-driven modeling with selection of spectral dissimilar soils (see Figure 7) is accurate for inducing local predictions of total C (Figure 6). Hence, there is the need to further improve data-driven selection using RS-LOCAL, i.e., by further optimizing the current version of the algorithm. To address this need, we could use combined memory-based or lazy learning strategies (Stanfill and Waltz, 1986; Lin and Vitter, 1994; Ramirez-Lopez et al., 2013) to optimize with more data-driven transfer methods (Pan and Yang, 2010) in terms of reducing the time needed to evaluate suitable subsets of the SSL for a new application. To give an example, some similarity criteria or clustering before doing calibration sampling could be used as prior information for reducing the SSL size to obtain the final subsets. In principle, the sample reduction could also be done with algorithms that can deal with non-linear relationships between spectra and soil properties, such as random forest or CUBIST. Another extension is to further filter spectral features and to do data compression to make the local modeling faster and even more adaptive to local conditions.

Our results showed that a transfer of the SSL to individual monitoring sites yielded very low bias and was accurate. This indicates that mid-IR spectroscopy and SSLs have the potential to give quick and relatively precise soil property estimates for soil monitoring. Nevertheless, the sites of the NABO long-term-monitoring program has not undergone substantial changes in OC (Gubler et al., 2019). Up to now, although major changes C content and organic composition should yield a spectral response, spectral changes in OC have mostly been reported along chronosequences (i.e., Awiti et al. (2008)), and only rarely for changes within individual plots over time (Deng et al., 2013). Hence, to address this, we propose to further investigate to what extent mid-IR spectroscopy can detect changes of OC considering small-scale variability and different agronomic management practices. This could for example be achieved with a study using soils from a long-term field trial, that shows sufficient temporal changes to be detected with spectroscopy.

The current SSL includes soils that contain between 0 and $583\,\mathrm{g\,kg^{-1}}$ total C and OC (Table 1). Because organic soils can have up to $500\,\mathrm{g\,kg^{-1}}$ OC, and because more than 98% of the samples are mineral soils, organic soils are underrepresented in the current Swiss SSL. For this reason, Helfenstein et al. (2020) evaluated the present Swiss SSL for a regional transfer based on new organo-mineral soils from two peat land regions in Switzerland. Although the range of total C measured was large (14 – $520\,\mathrm{g\,kg^{-1}}$ C) and the soils were diverse, as few as 5 or 10 site-specific tuning samples were sufficient to estimate the validation samples with reasonably accurately (RMSE = $< 30\,\mathrm{g\,kg^{-1}}$ C; RPIQ > 3.4); this was comparable to a local-only calibration with 50 samples. Helfenstein et al. found considerably lower conditional prediction errors ($< 10\,\mathrm{g\,kg^{-1}}$) when considering measurements of $< 100\,\mathrm{g\,kg^{-1}}$; this suggests that increasing the amount and compositional complexity of organic soils in the library has potential for more accurately characterizing diverse soil ecoregions with soil high organic matter contents.

Our results suggest that the present mid-IR SSL has great potential for applications that require soil data in high temporal and spatial coverage (i.e., for deriving quantitative indicators of soil quality for spatial planning or for soil-related environmental research). Mid-infrared spectral modeling was able to estimate many soil properties accurately with rather large variation in measurements explained (Figure 3), making them suitable for agronomic diagnosis and the assessment of soil functions in various landscapes. Currently, fine grained soil information of properties and function across agricultural lands in Switzerland is still scarce and often challenging to harmonize (i.e., measurement methods) because legacy maps are at varying levels of detail and quality (Keller et al., 2018; Grêt-Regamey et al., 2018). For example, only $13\,\%$ ($127000\,\mathrm{ha}$) of soil in agricultural land has been mapped with soil attributes of sufficient quality to evaluate its potential for crop production (Rehbein et al., 2020). Soil properties are also insufficiently mapped nationwide from point into space, depth and over time to regionally model soil processes, or to evaluate site-specific effects of agricultural practices on soils (i.e., soil C dynamics). Therefore, we suggest to couple infrared spectral estimation with traditional soil surveys and digital soil mapping to speed up the collection of soil information in Switzerland and elsewhere. This will offer means to test and further extend this SSL, so that only minimal amounts of costly and time consuming traditional laboratory analyses will be needed for characterizing and mapping soils' properties and functions in the next decades.

5 Conclusions

We developed the Swiss mid-IR SSL ($n = 4374$), using legacy soils and reference measurements of 16 properties, from 71 long-term monitoring sites (national soil monitoring; NABO) and 1094 locations sampled from a regular grid over Switzerland (biodiversity monitoring program; BDM). The trained CUBIST models — a general modeling approach using all data — were able to explain a relatively large proportion ($R^2 \geq 0.72$; RPIQ ≥ 2.0) of measured variance for ten of the properties. Total C, OC, total N, pH, CEC_{pot}, and clay content were estimated with high discrimination capacity ($R^2 > 0.8$; RPIQ $>$ 3.0). Total C was estimated with a cross-validated RMSE = $84\,\text{g}\,\text{kg}^{-1}$ at a measured range of $0-583\,\text{g}\,\text{kg}^{-1}$, and OC with RMSE$=9.3\,\text{g}\,\text{kg}^{-1}$ at the same measured range. Compared to the general CUBIST approach, the local transfer yielded on average 4.4 times more accurate estimates of total C with the mean RMSE$=0.7\,\text{g}\,\text{kg}^{-1}$ C, which is a substantial improvement of local estimates at plot-scale. Our similarity analysis revealed that local learning with subset selection based on RS-LOCAL produced a chemically diverse calibration set rather than narrowing down soil diversity for local modeling, as it is for example the case in memory-based learning. The developed national mid-IR SSL offers rapid soil estimates which are key inputs for many applications requiring soil information, such as digital soil mapping, agronomic diagnostics and precision farming, soil C accounting and monitoring, etc. The created mid-IR SSL and both local and general models can be updated with new soil records, which will allow to cover more soils conditions and will require less and less soil laboratory reference measurements in relation to spectral measurements for monitoring, mapping and modeling new soils.

Appendix A: Figures and tables in appendices

A1 Recursive feature elimination for interpreting general soil estimation with CUBIST

The recursive feature elimination (RFE) procedure started with the initial set of $S_1 = 209$ predictive variables that resulted after processing the spectra (see section 2.3). The following subset sizes S_i representing the number of spectral variables that are retained after each i^{th} variable elimination step were defined and evaluated within the RFE procedure:

$$S_i = \{209, 150, 120, 105, 90, 75, 60, 50, 40, 35, 30, 25, 20, 17, 14, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\} \tag{A1}$$

The first variable elimination step (i = 1) started with tuning a full CUBIST model derived from $S_1 = 209$ possible predictors using 10-fold cross-validation, then calculating the CUBIST model usage statistics for all predictors, next sorting all predictors from highest to lowest importance, and lastly dropping $S_1 - S_2 = 59$ of the least important predictors. For the next iteration ($i = 2$) and the following ones, we repeated this model fitting and variable reduction procedure with $S_2 = 150$ predictors and the preceding subsets, until the most important predictive variable ($S_{30} = 1$) was left at the last iteration ($i = 30$).

Variable selection is in addition prone to overoptimistic model assessment when resampling subsets (i.e., cross-validation) are used for two purposes, here model building and selection. This selection bias due to data leakage is well-documented for so-called *wrapper methods* of variable selection like RFE (Ambroise and McLachlan, 2002; Kuhn and Johnson, 2013), and occurs if these two tasks are not sufficiently separated by using independent data sets for each of them; this becomes especially more important when many predictive variables in relation to to relatively few observations are used, as it the case for our spectra.

To provide realistic predictive generalization of the RFE method, the aforementioned iterative selection procedure was done within an internal cross-validation scheme so that independent data were used to test the performance of the variable selection on the outer data segments. These outer cross-validation segments served external validation. To quantify the uncertainty of the models using the reduced variable sets and specifically variable selection, the outer cross-validation layer that served cross-validation was repeated five times, leading to five independent estimations per sample.

A2 Tuning profile of the RS-LOCAL parameters for local predictive transfers

The most relevant samples from the SSL at each respective NABO long-term monitoring plot were empirically selected at the RS-LOCAL configuration that yielded the lowest RMSE on two calibration samples per plot (Figure A1; performance profile). Time-series validation on the remaining samples of each site was separated from the optimization in the transfer workflow (see Figure 2).

Code and data availability. The data and the code to reproduce the results of this manuscript are available upon reasonable request.

Author contributions. Philipp Baumann drafted the manuscript, conceptualized spectral modeling scenario, carried out the data analysis, and wrote the code. Juhwan Lee and Johan Six developed the concept of having a Swiss Spectral Library. All co-authors contributed to the manuscript. NABO provided the milled soil archive of NABO and BDM, together with the chemical reference data and metadata. Andreas Gubler created the map of locations. Raphael Viscarra Rossel helped guide the spectral modeling experiments and provided the implementation of RS-LOCAL in R.

Competing interests. The authors declare no competing interests.

Acknowledgements. We express thanks to Leonardo Ramirez-Lopez, Laura Summerauer, and Jonas Anderegg for their valuable inputs that helped to improve the manuscript. This research was funded by ETH grants. We also would like to express gratitude to the Federal Office of Environment for commissioning and funding the soil analyses conducted within the BDM.

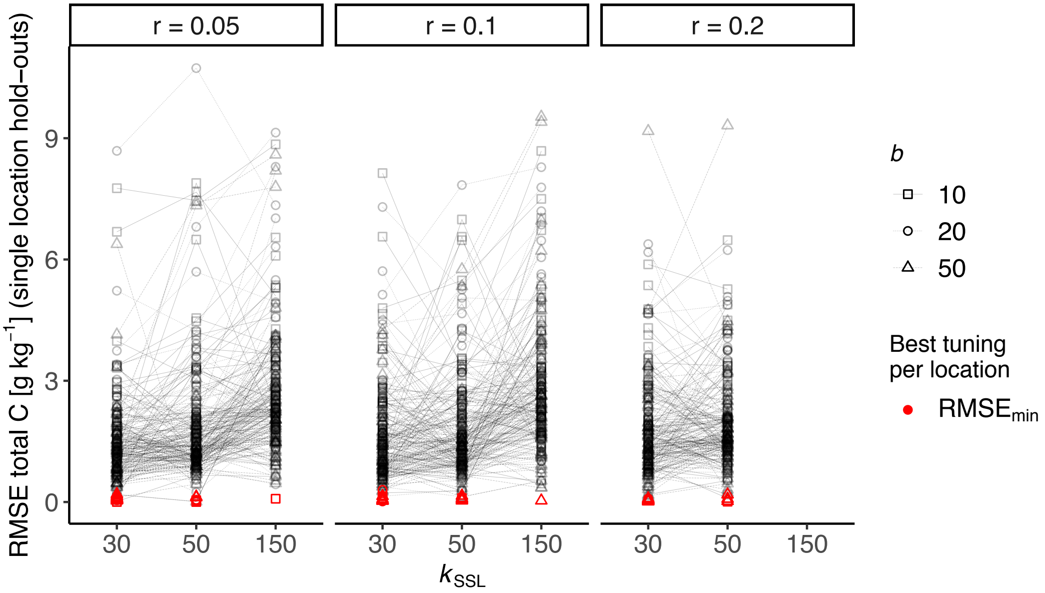

Figure A1. Performance profile of the 27 empirical parameter combinations of RS-LOCAL tested on each of the 71 NABO sites. The root mean squared error (RMSE) of the plot-level transfer was assessed with the first two calibration samples for each time series of total carbon (C) (see Figure 1 for an illustration of the setup of the local predictive transfer)

References

Agroscope: Referenzmethoden von Agroscope, 1996.

Ambroise, C. and McLachlan, G. J.: Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data, Proceedings of the National Academy of Sciences, 99, 6562–6566, https://doi.org/10.1073/pnas.102102699, 2002.

Angelopoulou, T., Balafoutis, A., Zalidis, G., and Bochtis, D.: From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review, Sustainability, 12, 443, https://doi.org/10.3390/su12020443, 2020.

Awiti, A. O., Walsh, M. G., Shepherd, K. D., and Kinyamario, J.: Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence, Geoderma, 143, 73–84, https://doi.org/10.1016/j.geoderma.2007.08.021, http://www.sciencedirect.com/science/article/pii/S0016706107002625, 2008.

Baumann, P.: philipp-baumann/simplerspec: Beta release simplerspec 0.1.0 for zenodo, https://doi.org/10.5281/zenodo.3303637, https://doi.org/10.5281/zenodo.3303637, 2019.

Bellman, R.: Adaptive Control Processes: A Guided Tour, Princeton University Press, http://www.jstor.org/stable/j.ctt183ph6v, 1961.

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, http://www.sciencedirect.com/science/article/pii/S0165993610001585, 2010.

Briedis, C., Baldock, J., de Moraes Sá, J. C., dos Santos, J. B., and Milori, D. M. B. P.: Strategies to Improve the Prediction of Bulk Soil and Fraction Organic Carbon in Brazilian Samples by Using an Australian National Mid-Infrared Spectral Library, Geoderma, 373, 114 401, https://doi.org/10.1016/j.geoderma.2020.114401, 2020.

Bui, E. N., Henderson, B. L., and Viergever, K.: Knowledge Discovery from Models of Soil Properties Developed through Data Mining, 191, 431–446, https://doi.org/10.1016/j.ecolmodel.2005.05.021.

Bundesamt für Umwelt (BAFU): Biodiversitätsmonitoring Schweiz BDM, 2014.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P. A., 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, 2016.

Dangal, S. R. 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, https://www.mdpi.com/2571-8789/3/1/11, 2019.

Deng, F., Minasny, B., Knadel, M., McBratney, A., Heckrath, G., and Greve, M. H.: Using Vis-NIR Spectroscopy for Monitoring Temporal Changes in Soil Organic Carbon, Soil Science, 178, 389–399, https://doi.org/10.1097/SS.0000000000000002, https://journals.lww.com/soilsci/Fulltext/2013/08000/Using_Vis_NIR_Spectroscopy_for_Monitoring_Temporal.2.aspx, 2013.

Desaules, A., Ammann, S., and Schwab, P.: Advances in long-term soil-pollution monitoring of Switzerland, Journal of Plant Nutrition and Soil Science, 173, 525–535, https://doi.org/10.1002/jpln.200900269, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jpln.200900269, 2010.

Dietterich, T. G., Wettschereck, D., Atkeson, C. G., and Moore, A. W.: Memory-Based Methods for Regression and Classification, in: Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993], edited by Cowan, J. D., Tesauro, G., and Alspector, J., pp. 1165–1166, Morgan Kaufmann, 1993.

Dokuchaev, V.: Report to the Transcaucasian Statistical Committee on Land Evaluation in General and Especially for the Transcaucasia. Horizontal and Vertical Soil Zones. (In Russian.) Off. Press Civ, Affairs Commander-in-Chief Cacasus, Tiflis, Russia, 1899.

Dowle, M. and Srinivasan, A.: data.table: Extension of ‘data.frame‘, https://CRAN.R-project.org/package=data.table, r package version 1.12.8, 2019.

England, J. R. and Viscarra Rossel, R. A.: Proximal Sensing for Soil Carbon Accounting, SOIL, 4, 101–122, https://doi.org/10.5194/soil-4-101-2018, https://www.soil-journal.net/4/101/2018/, 2018.

Friedman, J., Hastie, T., and Tibshirani, R.: The elements of statistical learning, Springer series in statistics Springer, Berlin, second edition edn., http://statweb.stanford.edu/~tibs/book/preface.ps, 2008.

Grêt-Regamey, A., Kool, S., Bühlmann, L., and Kissling, S.: Eine Bodenagenda für die Raumplanung, Thematische Synthese TS3 des Nationalen Forschungsprogramms «Nachhaltige Nutzung der Ressource Boden» (NFP 68), Swiss National Science Foundation (SNF), Bern, 2018.

Gubler, A., Wächter, D., and Schwab, P.: Homogenisation of Series of Soil Organic Carbon: Harmonising Results by Wet Oxidation (Swiss Standard Method) and Dry Combustion., 62.

Gubler, A., Wächter, D., Schwab, P., Müller, M., and Keller, A.: Twenty-five years of observations of soil organic carbon in Swiss croplands showing stability overall but with some divergent trends, Environmental Monitoring and Assessment, 191, 277, https://doi.org/10.1007/s10661-019-7435-y, 2019.

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, Geoderma, 158, 66–77, https://doi.org/10.1016/j.geoderma.2009.12.021, 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, 2014.

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., Zornoza, R., and Viscarra Rossel, R. A.: Do We Really Need Large Spectral Libraries for Local Scale SOC Assessment with NIR Spectroscopy?, Soil and Tillage Research, 155, 501–509, https://doi.org/10.1016/j.still.2015.07.008, http://linkinghub.elsevier.com/retrieve/pii/S0167198715001567, 2016.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines, Machine Learning, 46, 389–422, https://doi.org/10.1023/A:1012487302797, 2002.

Hand, D. J. and Vinciotti, V.: Local Versus Global Models for Classification Problems: Fitting Models Where it Matters, The American Statistician, 57, 124–131, https://doi.org/10.1198/0003130031423, 2003.

Helfenstein, A., Baumann, P., Viscarra Rossel, R., Gubler, A., Oechslin, S., and Six, J.: Quantifying Soil Carbon in Temperate Peatlands Using a Mid-IR Soil Spectral Library, 7, 193–215, https://doi.org/10.5194/soil-7-193-2021.

Helfenstein, A., Baumann, P., Viscarra Rossel, R., Gubler, A., Oechslin, S., and Six, J.: Predicting Soil Carbon by Efficiently Using Variation in a Mid-IR Soilspectral Library, submitted, 2020.

Hong, Y., Chen, S., Liu, Y., Zhang, Y., Yu, L., Chen, Y., Liu, Y., Cheng, H., and Liu, Y.: Combination of Fractional Order Derivative and Memory-Based Learning Algorithm to Improve the Estimation Accuracy of Soil Organic Matter by Visible and near-Infrared Spectroscopy, CATENA, 174, 104–116, https://doi.org/10.1016/j.catena.2018.10.051, http://www.sciencedirect.com/science/article/pii/S0341816218304867, 2019.

Hubert, M. and Debruyne, M.: Minimum Covariance Determinant, WIREs Computational Statistics, 2, 36–43, https://doi.org/10.1002/wics.61, 2010.

Janik, L. J. and Skjemstad, J. O.: Characterization and analysis of soils using mid-infrared partial least-squares .2. Correlations with some laboratory data, Australian Journal of Soil Research, 33, 637–650, https://doi.org/10.1071/sr9950637, https://www.publish.csiro.au/sr/sr9950637, publisher: CSIRO PUBLISHING, 1995.

Janik, L. J., Skjemstad, J. O., and Merry, R. H.: 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.

Jenny, H.: Factors of Soil Formation, McGraw-Hill Book Co., New York, 1941.

Keller, A., Franzen, J., Knüsel, P., Papritz, A., and Zürrer, M.: Bodeninformations-Plattform Schweiz (BIP-CH), Thematische Synthese TS4 des Nationalen Forschungsprogramms «Nachhaltige Nutzung der Ressource Boden» (NFP 68), Swiss National Science Foundation (SNF), Bern, 2018.

Kuhn, M.: caret: Classification and Regression Training, https://CRAN.R-project.org/package=caret, r package version 6.0-85, 2020.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling, Springer New York, New York, NY, 2013.

Lin, J.-H. and Vitter, J. S.: A Theory for Memory-Based Learning, Machine Learning, 17, 143–167, https://doi.org/10.1023/A:1022667616941, 1994.

Liu, L., Ji, M., and Buchroithner, M.: Transfer Learning for Soil Spectroscopy Based on Convolutional Neural Networks and Its Application in Soil Clay Content Mapping Using Hyperspectral Imagery, Sensors (Basel, Switzerland), 18, https://doi.org/10.3390/s18093169, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6165490/, 2018.

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., and Hedley, C. B.: RS-LOCAL Data-Mines Information from Spectral Libraries to Improve Local Calibrations: RS-LOCAL Improves Local Spectroscopic Calibrations, European Journal of Soil Science, https://doi.org/10.1111/ejss.12490, 2017.

Madari, B. E., Reeves, J. B., Machado, P. L., Guimarães, C. M., Torres, E., and McCarty, G. W.: Mid- and near-Infrared Spectroscopic Assessment of Soil Compositional Parameters and Structural Indices in Two Ferralsols, Geoderma, 136, 245–259, https://doi.org/10.1016/j.geoderma.2006.03.026, 2006.

Meuli, R. G., Wächter, D., Schwab, P., Kohli, L., and Zimmermann, R.: Connecting Biodiversity Monitoring with Soil Inventory Data – A Swiss Case Study, BGS Bulletin, 38, 3, 2017.

Miller, B. A., Koszinski, S., Wehrhan, M., and Sommer, M.: Comparison of Spatial Association Approaches for Landscape Mapping of Soil Organic Carbon Stocks, 1, 217–233, https://doi.org/10.5194/soil-1-217-2015.

Nguyen, T., Janik, L., and Raupach, M.: Diffuse Reflectance Infrared Fourier Transform (DRIFT) Spectroscopy in Soil Studies, Soil Research, 29, 49, https://doi.org/10.1071/SR9910049, 1991.

Nocita, M., Stevens, A., van Wesemael, B., Brown, D. J., Shepherd, K. D., Towett, E., Vargas, R., and Montanarella, L.: Soil Spectroscopy: An Opportunity to Be Seized, Global Change Biology, 21, 10–11, https://doi.org/10.1111/gcb.12632, 2015.

Ogen, Y., Zaluda, J., Francos, N., Goldshleger, N., and Ben-Dor, E.: Cluster-Based Spectral Models for a Robust Assessment of Soil Properties, Geoderma, 340, 175–184, https://doi.org/10.1016/j.geoderma.2019.01.022, http://www.sciencedirect.com/science/article/pii/S0016706118316045, 2019.

Padarian, J., Minasny, B., and McBratney, A. B.: 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, http://www.sciencedirect.com/science/article/pii/S0016706118305639, 2019a.

Padarian, J., Minasny, B., and McBratney, A. B.: Using Deep Learning to Predict Soil Properties from Regional Spectral Data, Geoderma Regional, 16, e00 198, https://doi.org/10.1016/j.geodrs.2018.e00198, http://www.sciencedirect.com/science/article/pii/S2352009418302785, 2019b.

Pan, S. J. and Yang, Q.: A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering, 22, 1345–1359, https://doi.org/10.1109/TKDE.2009.191, 2010.

Parikh, S. J., Goyne, K. W., Margenot, A. J., Mukome, F. N. D., and Calderón, F. J.: Chapter One - Soil Chemical Insights Provided through Vibrational Spectroscopy, in: Advances in Agronomy, edited by Sparks, D. L., vol. 126, pp. 1–148, Academic Press, https://doi.org/10.1016/B978-0-12-800132-5.00001-8, 2014.

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., and Greve, M. H.: Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra, 10, e0142 295, https://doi.org/10.1371/journal.pone.0142295.

Pratt, L. and Thrun, S.: Guest Editors' Introduction, Machine Learning, 28, 5–5, https://doi.org/10.1023/A:1007322005825, 1997.

Pratt, L. Y., Pratt, L. Y., Hanson, S. J., Giles, C. L., and Cowan, J. D.: Discriminability-Based Transfer between Neural Networks, in: Advances in Neural Information Processing Systems 5, pp. 204–211, Morgan Kaufmann, 1993.

Quinlan, J.: Combining Instance-Based and Model-Based Learning, in: Machine Learning Proceedings 1993, pp. 236–243, Elsevier, https://doi.org/10.1016/B978-1-55860-307-3.50037-X, https://linkinghub.elsevier.com/retrieve/pii/B978155860307350037X, 1993.

Quinlan, J. R.: Learning with Continuous Classes, in: 5th Australian Joint Conference on Artificial Intelligence, vol. 92, pp. 343–348, World Scientific, 1992.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2019.

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, http://www.sciencedirect.com/science/article/pii/S0016706112004314, 2013.

Rehbein, K., Sprecher, C., and Keller, A.: Übersicht Stand Bodenkartierung in Der Schweiz. Ergänzung Des Bodenkartierungskataloges Schweiz Um Bodeninformationen Aus Meliorationsprojekten, Servicestelle NABODAT, Agroscope, Zürich, 2020.

Rousseeuw, P. J.: Least Median of Squares Regression, Journal of the American Statistical Association, 79, 871–880, https://doi.org/10.1080/01621459.1984.10477105, 1984.

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, publisher: American Chemical Society, 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, http://www.sciencedirect.com/science/article/pii/S0016706119304537, 2019.

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, http://linkinghub.elsevier.com/retrieve/pii/S0169743916300351, 2016.

Skjemstad, J. and Dalal, R.: Spectroscopic and Chemical Differences in Organic Matter of Two Vertisols Subjected to Long Periods of Cultivation, Soil Research, 25, 323, https://doi.org/10.1071/SR9870323, 1987.

Solomatine, D.: Combining Machine Learning and Domain Knowledge in Modular Modelling, in: Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications, edited by Abrahart, R. J., See, L. M., and Solomatine, D. P., Water Science and Technology Library, pp. 333–345, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-79881-1_24, 2008.

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, 2014.

Stanfill, C. and Waltz, D.: Toward Memory-Based Reasoning, Communications of the ACM, 29, 1213–1228, https://doi.org/10.1145/7902.7906, http://portal.acm.org/citation.cfm?doid=7902.7906, 1986.

Stenberg, B. and Rossel, R. V.: Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing, in: Proximal Soil Sensing, edited by Viscarra Rossel, R. A., McBratney, A. B., and Minasny, B., Progress in Soil Science, pp. 29–47, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-90-481-8859-8_3, 2010.

Stevens, A. and Ramirez-Lopez, L.: An introduction to the prospectr package, r package version 0.1.3, 2013.

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, 2013.

Thrun, S. and Pratt, L., eds.: Learning to Learn, Springer US, Boston, MA, https://doi.org/10.1007/978-1-4615-5529-2, 1998.

Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B., and Zalidis, G. C.: Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network, Geoderma, 367, 114 208, https://doi.org/10.1016/j.geoderma.2020.114208, http://www.sciencedirect.com/science/article/pii/S0016706119308870, 2020.

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, http://www.sciencedirect.com/science/article/pii/S0016706118307006, 2019.

Varmuza, K. and Filzmoser, P.: Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, 0 edn., https://doi.org/10.1201/9781420059496, https://www.taylorfrancis.com/books/9781420059496, 2016.

Viscarra Rossel, R., Walvoort, D., McBratney, A., Janik, L., and Skjemstad, J.: Visible, near Infrared, Mid Infrared or Combined Diffuse Reflectance Spectroscopy for Simultaneous Assessment of Various Soil Properties, Geoderma, 131, 59–75, https://doi.org/10.1016/j.geoderma.2005.03.007, 2006.

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, 2016.

Viscarra Rossel, R. A. and McBratney, A. B.: Soil chemical analytical accuracy and costs: implications from precision agriculture, Australian Journal of Experimental Agriculture, 38, 765, https://doi.org/10.1071/EA97158, http://www.publish.csiro.au/?paper=EA97158, 1998.

Viscarra Rossel, R. A. and Webster, R.: Predicting Soil Properties from the Australian Soil Visible-near Infrared Spectroscopic Database, European Journal of Soil Science, 63, 848–860, https://doi.org/10.1111/j.1365-2389.2012.01495.x, 2012.

Viscarra Rossel, R. A., Lobsey, C. R., Sharman, C., Flick, P., and McLachlan, G.: Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition, Environmental Science & Technology, 51, 5630–5641, https://doi.org/10.1021/acs.est.7b00889, 2017.

Wang, Y. and Witten, I. H.: Induction of model trees for predicting continuous classes, Working Paper 96/23, University of Waikato, Department of Computer Science, Hamilton, New Zealand, https://researchcommons.waikato.ac.nz/handle/10289/1183, accepted: 2008-10-29T02:09:15Z ISSN: 1170-487X, 1996.

Wickham, H.: tidyverse: Easily Install and Load the 'Tidyverse', https://CRAN.R-project.org/package=tidyverse, r package version 1.3.0, 2019.

Wold, S., Martens, H., and Wold, H.: The Multivariate Calibration Problem in Chemistry Solved by the PLS Method, in: Matrix Pencils, edited by Kågström, B. and Ruhe, A., vol. 973, pp. 286–293, Springer Berlin Heidelberg, https://doi.org/10.1007/BFb0062108, http://link.springer.com/10.1007/BFb0062108, 1983.

Wolpert, D. and Macready, W.: No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation, 1, 67–82, https://doi.org/10.1109/4235.585893, 1997.

Wolpert, D. H.: The Lack of A Priori Distinctions Between Learning Algorithms, Neural Computation, 8, 1341–1390, https://doi.org/10.1162/neco.1996.8.7.1341, 1996.