Geoderma 365 (2020) 114227

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy

Leonardo Deiss[a,*,1], Andrew J. Margenot[b,2], Steve W. Culman[c,3], M. Scott Demyan[d,4]

[a] School of Environment and Natural Resources, The Ohio State University, 414A, Kottman Hall, 2021 Coffey Road, Columbus, OH 43210, USA
[b] Crop Sciences Department, University of Illinois Urbana-Champaign, 1201 S Dorner Dr, Urbana, IL 61801, USA
[c] School of Environment and Natural Resources, The Ohio State University, 1680, Madison Ave, Wooster, OH 44691, USA
[d] School of Environment and Natural Resources, The Ohio State University, 408B, Kottman Hall, 2021 Coffey Road, Columbus, OH 43210, USA

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Machine-learning
Kernel
Error-grid
FTIR
RMSE

ABSTRACT

Estimating soil properties in diffuse reflectance infrared Fourier transform spectroscopy in the mid-infrared region (mid-DRIFTS) uses statistical modeling (chemometrics) to predict soil properties from spectra. Modeling approaches can have major impacts on prediction accuracy. However, the impact of selecting best parameters for an algorithm (tuning), to optimize non-linear models for predicting soil properties, is relatively unexplored in the domain of soil sciences. This study aimed to evaluate the predictive performance of linear (partial least squares, PLS) and non-linear (support vector machines, SVM) multivariate regression models in estimating soil physical, chemical, and biological properties with mid-DRIFTS. We evaluated the impact of optimizing two hyperparameters (*epsilon* and *cost*) based on the noise tolerance in the ε-insensitive loss function of SVM models using two contrasting and diverse sets of soils, one from northern Tanzania (n = 533) and another one from USA Midwest (n = 400). Regression models were trained on calibration sets (75%) and tested on independent validation sets (25%) separately for each dataset. Support vector machines outperformed PLS models for all tested soil properties (clay, sand, pH, total organic carbon, and permanganate oxidizable carbon) in both datasets. Tuning hyperparameters *epsilon* and *cost* maintained or improved prediction accuracy of SVM models based on root mean squared errors of independent validation sets. Support vector machines tuned hyperparameters differed among soil properties and also for the same soil property in distinct datasets, suggesting the need for parameterizing non-linear models for specific soil properties and datasets. Optimizing SVM regression models in mid-DRIFTS improves prediction accuracy of soil properties and therefore will likely enable obtaining more robust predictive outcomes even in datasets with diverse land uses, parent materials, and/or soil orders. We recommend that tuning should be included as a routine step when using SVM for estimating soil properties.

1. Introduction

Multivariate modeling has mainstreamed diffuse reflectance infrared Fourier transform spectroscopy in the mid-infrared region (mid-DRIFTS), transforming soil sciences by enabling high-throughput predictions of soil properties. The mid-DRIFTS technique, also known as middle-infrared (MIR) spectroscopy or Fourier transform infrared (FTIR) spectroscopy, differs from traditional laboratory approaches to soil analysis (e.g., wet chemistry) in that outputs are predictions or estimates derived from the statistical modeling of the complex relationships between a reference soil property and the mid-infrared spectrum of the same soil. An absorbance spectrum exhibits peaks that represent absorption of infrared electromagnetic energy at frequencies (cm^{-1}) specific to the type and vibrational mode(s) of polar bonds of organic and inorganic functional groups (Parikh et al, 2014; Nocita et al, 2015). Most soil properties cannot be directly estimated from specific peak measurements in mid-DRIFTS of neat soil samples (Niemeyer et al., 1992) due to overlapping and overtone vibrations that occur in mid-infrared frequencies (4000–400 cm^{-1}) (Soriano-Disla et al. 2014) or simply by the lack of peaks specific to those soil

* Corresponding author.
E-mail addresses: deiss.8@osu.edu (L. Deiss), margenot@illinois.edu (A.J. Margenot), culman.2@osu.edu (S.W. Culman), demyan.4@osu.edu (M.S. Demyan).
[1] ORCID: 0000-0003-2001-9238.
[2] ORCID: 0000-0003-0185-8650.
[3] ORCID: 0000-0003-3985-257X.
[4] ORCID: 0000-0001-6198-3774.

https://doi.org/10.1016/j.geoderma.2020.114227
Received 13 September 2019; Received in revised form 23 January 2020; Accepted 23 January 2020
Available online 05 February 2020

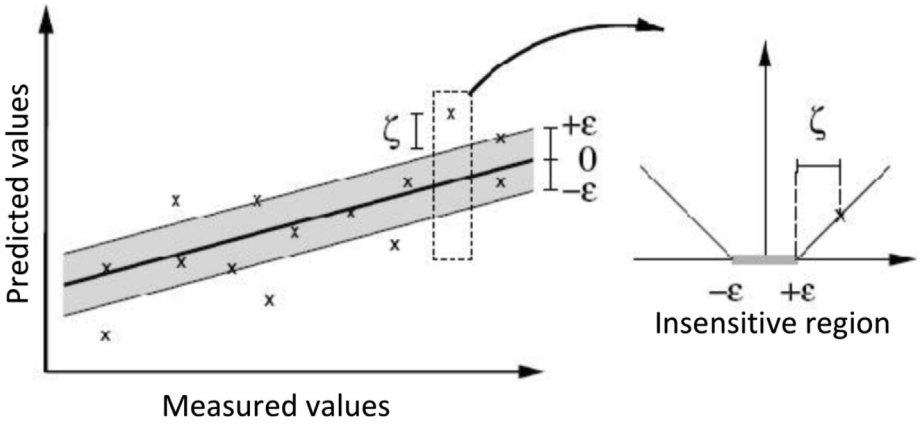

Fig. 1. Theoretical representation of how support vector machines models address the loss function using an ε-intensive band. The soft margin loss setting corresponds to a linear kernel of SVM, where the *cost* (*C*) size of errors (ζ) is measured in a high-dimensional space, and only errors larger than a cut-off (ε) are taken into account. Image adapted from Schölkopf and Smola (2002) with permission from The MIT Press.

properties. Soil properties can be predicted, however, by multivariate regression models that extract and model relevant information from the spectra. These predictions rely on a spectral library that contains measured soils data obtained from traditional analytic methods, but after developing a model on a training dataset, and validating it with acceptable errors on an independent test set, one can use the trained model to perform predictions on new "unknown" soil samples. Given the complexity of a soil mid-DRIFTS spectrum, which generally contain more than a thousand spectral variables (e.g., 1650 variables in a 2 cm^{-1} spectrum ranging 4000–700 cm^{-1}, without zero filing/interpolating variables), and whose individual peaks are not necessarily directly associated with the soil property of interest, multivariate models are often used to analyze soil spectra and generate quantitative predictions of soil properties. The chemometrics component of the mid-DRIFTS measurement process is indispensable and modeling approaches can strongly affect the predictive outcomes.

In predictive applications of soil mid-DRIFTS, chemometrics is composed of two main steps: spectral treatments and multivariate regression modeling. Mathematical spectral treatments are used to enhance spectral features and increase ability of models to extract vibrational information from the spectra (see e.g. Stenberg et al., 2010; Gholizadeh et al., 2013 for more details on spectral treatments). Whereas multivariate models use pre-treated spectral data to develop calibrations based on known values of a given soil property and specific spectral features. There are two classes of multivariate regression models: linear and non-linear (Wehrens, 2011), and several model types have been used to calibrate spectral data with measured soil data. For example, partial least squares (PLS) is a widely used linear multivariate regression model, a class of models that also includes multiple linear regression and principal components regression. On the other hand, support vector machines (SVM), random forests, and neural networks are examples of non-linear multivariate regression models. In soil sciences, these linear and/or non-linear models have been compared for predictive accuracy (e.g. Souza et al., 2012; Kang et al., 2017; Jia et al., 2017; Campbell et al., 2018), but in general non-linear models, especially SVM (Gholizadeh et al., 2013), have been underused. Moreover, the previous limited comparisons of linear and non-linear multivariate regression models in mid-DRIFTS have found variable results depending on the soil property of interest and the soil sample set. Though such variability in accuracy can in part reflect intrinsic soil characteristics, it may also be due to differences in modeling approaches between studies. Modeling choices of spectral treatments and model class/parameterization can significantly compromise or even degrade predictive applications of mid-DRIFTS of soils.

Support vector machines regression is a supervised, nonparametric, statistical learning technique (Vapnik, 1995), and it generally has adequate balance between predictive accuracy and the ability to generalize trained models to unseen data (Gholizadeh et al., 2013). Advantages of SVM models are their ability to handle high-dimensional multivariate spaces (Karatzoglou et al., 2006) and to deal with noisy patterns and multi-modal class distributions of soil properties (Gholizadeh et al., 2013). However, challenging analytical approaches are that SVM models have different algorithms and optimizing (tuning) parameters that can be specifically targeted to improve prediction outcomes. This latter part has been less explored in soil sciences, and information regarding specific SVM parameters to be used in these models to optimize predictive outcomes for soil properties is lacking. Support vector machines regression models have two main features that can be optimized. First is the selection of the kernel function (algorithm), and second the noise tolerance in the epsilon (ε)-insensitive loss function for each kernel. Kernel functions return the inner product between two points in a suitable feature space, thus defining a notion of similarity in high-dimensional spaces (Karatzoglou et al., 2006). In SVM regression, there are four main families of kernels: linear, polynomial, radial, and sigmoid, and each kernel has its own optimization parameters, and potentially scenarios of suitability. A common optimization parameter among most SVM kernels is the noise tolerance in the ε-insensitive loss function.

A theoretical representation of how SVM models deal with the ε-insensitive loss function is presented in Fig. 1. Generally, a typical parameterization of the SVM function is to use the ε-insensitive error function in which an $\varepsilon = 0.1$ corresponds to a value of 1 for the penalization or cost parameter (*C*) (Wehrens, 2011). However, this parameter ε can be optimized based on the trade-off between the size of ε (insensitivity zone) and *C*. Reducing the insensitivity zone will generally increase the size of *C*, i.e. the distance between points outside of the insensitive zone to the limit of the insensitive zone. These optimum parameters can be searched in a user-defined hyperparameteric range using a cross-validated error grid (e.g., root mean squared error, RMSE) to find the greatest prediction accuracy of a soil property of interest. Information about the coefficients ε and *C* is virtually absent in SVM parameter optimization in mid-DRIFTS of soils. As parameter optimization should lead to more accurate prediction outcomes, finding a model composition that minimizes prediction errors is an important step to increase robustness of soil analysis with mid-DRIFTS.

Existing implementations of SVM regression models generally treat its parameters as user-defined inputs, but there is lack of information about specific values to use when predicting soil properties with mid-DRIFTS. Selecting specific kernel type and function parameters is usually based on application-domain knowledge. In the case of soil spectra, user-defined inputs might be specific for each soil property in a given spectral library and differ among sample sets. Generating information about the optimization parameters allows users to compare and possibly improve prediction performances using robust modeling approaches. Once a valid kernel function and its optimization parameters have been selected, one can develop further predictions with minimal additional computational cost. Therefore, the objective of this study is to compare predictive performance of linear (PLS) and non-

Table 1
Summary statistics of measured soil properties of the USA Midwest and Tanzania soil sets. Permanganate oxidizable carbon.

	Clay %	Sand %	pH	TOC %	POXC mg kg^{-1}
USA Midwest					
n	400	400	400	399	400
Min.	1	0	2.3	0.1	3
1st Qu.	17	8	4.9	0.3	102
Median	26	18	5.9	0.5	154
Mean	29	26	5.9	0.9	286
3rd Qu.	39	35	7.0	1.1	327
Max.	85	98	8.0	9.1	1412
Tanzania					
n	533	533	335	533	532
Min.	21	1	4.5	0.6	7
1st Qu.	46	15	6.0	1.8	331
Median	55	20	6.6	2.4	507
Mean	54	22	6.4	2.4	522
3rd Qu.	62	26	6.9	2.9	684
Max.	91	58	7.8	6.1	1404

pH: 1:1 v/v soil:water. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

linear (SVM) multivariate regression models in mid-DRIFTS of soils, and to evaluate how SVM model parameters ε and C affect prediction accuracy of soil physical, chemical, and biological properties (clay, sand, pH, total organic carbon (TOC), and permanganate oxidizable C (POXC)).

2. Methods

2.1. Soils and study areas

Two geographically distinct and edaphically diverse soil sample sets were used for a comparative evaluation of parameter optimization of SVM relative to PLS, from northern Tanzania (n = 533) and from USA Midwest (n = 400) (Table 1).

In northeastern Tanzania, soils were sampled across a mountainous landscape dominated by smallholder agriculture in Lushoto District, Tanga Province. The study site is located in the Western Usambara Mountains, a deeply dissected plateau that rises in a steep escarpment from the surrounding Maasai Plains (Massawe et al., 2017). Soils are developed on Precambian metamorphic parent material (intermediate gneiss) (Appel et al., 1998). Depending on landscape position, Ultisols generally occupy midslope and upperslope positions, while toeslope and valley positions are a complex of Mollisols, Alfisols, Ultisols, Inceptisols, and Entisols (Massawe et al., 2017). A total of 67 fields from 23 farms were identified based on landscape position and management intensity representative of East African highland smallholder systems (Massawe et al., 2017; Winowiecki et al., 2016) to furnish a diverse set of soils and soil properties. In February 2014, each field was sampled with an auger at five within field locations at 0–20 cm, and three randomly selected locations from those five were further sampled at 20–40 cm depth, for a total of n = 533 soil samples.

The 400 USA soil samples were selected from the National Cooperative Soil Survey (NCSS, https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/survey/) distributed over four different physiographic regions of the USA Midwest: Glaciated Allegheny Plateau, Unglaciated Allegheny Plateau, Till Plains, and Huron-Erie Lake Plains. For each region, 100 samples were systematically selected from archived genetic horizons sampled by NCSS (1950–2012) to represent the full range and distribution of the entire dataset within each physiographic region. Legacy data is digitally available at https://ncsslabdatamart.sc.egov.usda.gov/. Samples were assigned to a physiographic region based on their county location, and samples from counties containing two or more physiographic regions were excluded. Additionally, only samples after 1966 were considered, as prior to that year wet combustion (oxidation via the Walkley Black method) was used to determine TOC.

2.2. Soil analysis

Soils from both regions (Tanzania and the USA) were analyzed for soil particle size distribution (soil texture) using the pipet method (method 3A1, Burt, 2011). The soil texture variables used in this study were total sand (< 2.0 mm and > 0.05 mm) and total clay (< 0.002 mm). Total organic carbon was analyzed in Tanzania soils using dry combustion-chromatography; and in the USA soils using dry combustion-Dumas (method 6A2a, Burt, 2011). In soils without carbonates, total carbon was taken as total organic carbon, while in soils with carbonates, inorganic carbon was determined separately by the gasometric method (Dreimanis, 1962) and inorganic carbon content subtracted from total carbon to yield total organic carbon. Carbonates were not detected in Tanzania soils. Soil pH was measured in Tanzania using a 1:2 soil:water mixture (v:v); and in the USA using a 1:1 soil:-water mixture (v:v) (method 4C1a2a, Burt, 2011). In both soil sets, permanganate oxidizable carbon (POXC, mg kg^{-1} soil) was measured based on the methods of Weil et al. (2003) adapted by Culman et al. (2012).

2.3. Sample preparation and instrument set-up for mid-DRIFTS

2.3.1 The Tanzania samples

Soil samples were air-dried and initially sieved to < 2 mm for standard laboratory analysis, while all samples analyzed using mid-DRIFTS were ground to < 100 μm with an agate mortar and pestle, according to procedures described in Terhoeven-Urselmans et al. (2010). Soil samples were loaded in four replicate wells on aluminum microtiter plates (A752-96, Bruker Optics, Karlsruhe) using a microspatula to fill the 6-mm-diameter wells and level the soil. Soil mid-DRIFT spectra were obtained using a FT- IR Tensor 27 with high-throughput screening extension unit with robotic arm ([Twister Microplate Handler], Bruker Optics, Karlsruhe, Germany; illustrated in Shepherd and Walsh, 2007). The detector was a liquid N_2-cooled MCT detector. Spectra were collected across 4000–602 cm^{-1} with a resolution of 4 cm^{-1}. Background measurements of the first empty well were taken before each single measurement to account for changes in temperature and air humidity. Each one of the four replicate wells had 32 co-added scans, and the four spectra were averaged to account for within-sample variability and differences in particle size and packing density (Terhoeven-Urselmans, et al., 2010).

2.3.2 The USA samples

The samples from the soil survey were originally crushed and sieved to < 2.0 mm and stored in an air-dried state. No further grinding was performed for mid-DRIFTS (Deiss et al., 2019a). Before acquiring spectra, soils were dried for > 48 h at 40 °C and at 12–14% relative humidity. To analyze samples in the mid-DRIFTS instrument, 24-well anodized aluminum plates were used. These plates hold 24 removable polystyrene sample cups with a top circular opening area of 10 mm diameter and 5.5 mL volume. The sample cups were loaded by initially over-filling the cups with soil, then tapping the cup side gently thrice to settle the soil into the cup, and finally smoothing the surface by scraping excess soil with the narrow edge of a stainless-steel spatula. The soil was not packed or compressed into the well other than by tapping and scrapping to avoid artifacts of matrix density (Terhoeven-Urselmans, et al., 2010).

Spectra from USA soils were obtained using an X,Y Autosampler (Pike Technologies Inc., Madison, WI) coupled with a Nicolet iS50 spectrometer equipped with a diffuse reflectance accessory (Thermo Fisher Scientific Inc., Waltham, MA). Potassium bromide (KBr) was used for background spectrum collected at the beginning of each plate

reading (i.e., every 23 samples). All measurements were conducted from 4000 to 400 cm^{-1}, 4 cm^{-1} wavenumber resolution and with 24 co-added scans in absorbance mode. We further reduced the spectral data to 4000–700 cm^{-1} to eliminate increased noise at the upfield spectral boundary to conduct spectral analysis and predictions. For each soil sample, four soil subsamples were measured with one spectral reading per well (24 co-added scans each) to generate the spectral replicates that were further averaged prior to qualitative analysis and predictions. The spectral readings were randomly located within a 3 mm diameter in the central position of each well configured in AutoPro™ software (Pike Technologies Inc., Madison, WI).

2.4. Spectra characterization

Characterization of absorbance (log R^{-1}, where R is reflectance) spectra was summarized using principal components analysis (PCA) with spectra mean-centered by subtracting wavenumber-specific absorbance means (overall spectra) from each spectrum wavenumber-specific absorbance (centering was done with R Package 'base', R Core Team, 2016). We used the iterative NIPALS algorithm (Martens and Naes, 1989) to derive the principal components (R package 'chemometrics', Varmuza and Filzmoser, 2009). The first two principal component scores and loadings were plotted to evaluate soil datasets spectra PCA dispersion and wavenumber-specific PCA loadings distribution.

2.5. Spectral treatment and selection

Several spectral treatments were evaluated for ability to extract vibrational information from the spectra, and increase model robustness, accuracy, repeatability, and reproducibility (Stevens and Ramirez-Lopez, 2015). Tested treatments were Savitzky–Golay smoothing and derivative, GapSegment derivative, continuum-removal, detrend normalization, standard normal variate, block scaling, and sum of squares block weighting. Standard normal variate transformation (Fearn, 2008) and detrend normalization were also tested in combination with filtering (applied after Savitzky–Golay and Gap-Segment) (R package 'prospectr', Stevens and Ramirez-Lopez, 2015). Selected spectral treatments specific for each soil property and soil set are described in Tables 2 and 3.

Models were trained on a representative calibration set (75% of the dataset) selected using the Kennard–Stone sampling algorithm (Kennard and Stone, 1969), specifically for each spectral treatment, to explain ≥95% of the total variance and validated on the remaining samples (25% of the dataset) (R package 'prospectr', Stevens and Ramirez-Lopez, 2015). In the USA dataset, this selection process was separately performed for each one of the four physiographic locations (n = 100 each) for a final calibration set of n = 300 and validation set of n = 100 (except TOC, calibration = 300 and validation n = 99). In the Tanzania dataset, this selection process was done across all samples maintaining the proportion 75% calibration set to 25% test set for all soil properties.

Prior to modeling, spectral outliers were detected using absorbance spectra considering orthogonal distance and score distance. Orthogonal distance was between the true position of each data point and its projection in space of the first few principal components to explain ≥80% of the total variance. Score distance was the projection of a sample to the center of all sample projections (Wehrens, 2011). The final dataset was constrained to a sample set with orthogonal distance < 25 and score distance < 6 for the USA dataset, and orthogonal distance < 4 and score distance < 6 for the Tanzania dataset. No outliers were excluded within these orthogonal distance and score distance ranges.

2.6. Prediction model calibration and independent validation

We trained PLS and SVM models with different algorithms on calibration sets and these were subsequently tested on independent validation sets. For PLS, three algorithms were tested, including kernel, SIMPLS, and classical orthogonal scores (R package 'pls', Mevik and Wehrens, 2007). The number of latent vectors in PLS was determined via 10-fold cross-validation (R package 'chemometrics', Varmuza and Filzmoser, 2009). For SVM, four kernels (classes of algorithms in SVM) were tested, including linear kernel, Gaussian Radial Basis Function (RBF) kernel, polynomial kernel (second and third degrees), and hyperbolic tangent kernel (sigmoid) (R package 'e1071', Meyer et al., 2015). A common configuration tested in all PLS algorithms was with or without a scaling function. Pre-treated spectra were scaled or not scaled for PLS by dividing centered wavenumber-specific absorbances by their standard deviations (Mevik and Wehrens, 2007; Varmuza and Filzmoser, 2009), whereas for SVM both pre-treated spectra and predictor were always scaled to zero mean and unit variance prior to calibration (Meyer et al., 2015).

Best combination of spectral treatment and multivariate regression model (PLS and SVM) were selected for each soil property and dataset based on sequential criteria looking first at the lowest root mean squared error ($RMSE_v$), then greatest residual prediction deviation

Table 2
Partial least squares (PLS) spectral treatments and model configurations used to predict soil variables in datasets from USA and Tanzania (TZ) in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS).

Soil property	Locate	Spectral treatment[a]	Arguments[b]	Multivariate regression model details[c]		
				Algorithm	Scaling	Latent variables (LV)
Clay (%)	USA	S-G/DT	DO 0, PO 1, SS 11	classical	scaled	7 LV
	TZ	G-S/SNV	DO 2, FL 11, SS 1	classical	scaled	4 LV
Sand (%)	USA	S-G/SNV	DO 0, PO 2, SS 11	classical	scaled	8 LV
	TZ	G-S/DT	DO 1, FL 11, SS 10	classical	scaled	6 LV
pH	USA	S-G/SNV	DO 0, PO 4, SS 11	classical	scaled	7 LV
	TZ	S-G/SNV	DO 1, PO 3, SS 11	simpls	scaled	5 LV
TOC (%)[d]	USA	Absorbance	(log reflectance^{-1})	simpls	non-scaled	6 LV
	TZ	S-G/DT	DO 2, PO 4, SS 10	kernelpls	scaled	2 LV
POXC (mg kg^{-1})[d]	USA	Movav	FL 11	classical	scaled	5 LV
	TZ	D	DO 1	simpls	scaled	1 LV

[a] Savitzky-Golay (S-G), Gap-Segment (G-S), Detrend (DT), Standard Normal Variate (SNV), and Moving Average filter (Movav). Absorbance (log reflectance^{-1}) was the spectral basis for all other pretreatments.
[b] Derivative orders (DO), segment sizes (SS), polynomial orders (PO), and filter length (FL).
[c] For PLS three algorithms were tested: the kernel, SIMPLS and the classical orthogonal scores.
[d] Modeling was conducted with logarithmic transformed data. Spectra was scaled or not by dividing centered wavenumber-specific absorbances by their standard deviations. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

Table 3
Classical support vector machines (SVM) and tuned SVM (tSVM) spectral treatments and model configurations used to predict soil properties in datasets from USA and Tanzania (TZ) in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS).

Soil property	Locate	Spectral treatment[a]	Arguments[b]	Model	Multivariate regression model details [c]		
					cost (*C*)	*epsilon* (ε)	N vectors
Clay (%)	USA	DT	–	SVM	1.00	0.1	230
				tSVM	4.78	0.001	298
	TZ	S-G/SNV	DO 1, PO 1, SS 11	SVM	1.00	0.1	309
				tSVM	2.14	0.001	400
Sand (%)	USA	S-G/SNV	DO 0, PO 4, SS 11	SVM	1.00	0.1	231
				tSVM	17.15	0.001	300
	TZ	G-S/DT	DO 1, FL 11, SS 10	SVM	1.00	0.1	293
				tSVM	4.29	0.001	399
pH	USA	S-G/SNV	DO 0, PO 4, SS 11	SVM	1.00	0.1	250
				tSVM	4.29	0.201	182
	TZ	G-S/DT	DO 3, FL 11, SS 1	SVM	1.00	0.1	188
				tSVM	2.14	0.001	250
TOC (%)	USA	S-G/SNV	DO 0, PO 1, SS 11	SVM	1.00	0.1	157
				tSVM	8.57	0.001	297
	TZ	G-S/DT	DO 1, FL 11, SS 1	SVM	1.00	0.1	333
				tSVM	2.14	0.201	281
POXC (mg kg^{-1})	USA	S-G/DT	DO 0, PO 1, SS 11	SVM	1.00	0.1	188
				tSVM	4.28	0.001	299
	TZ	G-S/DT	DO 2, FL 11, SS 5	SVM	1.00	0.1	313
				tSVM	1.07	0.201	242

[a] Savitzky-Golay (S-G), Gap-Segment (G-S), Derivative (D), Detrend (DT), and Standard Normal Variate (SNV).
[b] Absorbance was the spectral basis for all other pretreatments. Derivative orders (DO), segment sizes (SS), polynomial orders (PO), and filter length (FL).
[c] Four SVM kernels were tested: linear, polynomial (second and third degrees), radial basis and sigmoid. N vectors: number of support vectors. Spectra and predictor were scaled to zero mean and unit variance prior to calibration. For all SVM and tSVM, kernel: radial. Gamma: 0.000587 (clay), 0.000594 (sand), 0.000587 (pH), 0.000146 (TOC), and 0.000596 (POXC). Range of tested parameters: *C* from 0 to 25 for both datasets. ε was from 0.001 to 0.5 in the Tanzania dataset and from 0.001 to 1.0 in the USA dataset. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

(RPD_v), and then greatest coefficient of determination (R^2_v) of the independent validation datasets. The RMSE is the difference between observed values and the predicted values. The RPD is the standard deviation of observed values divided by the RMSE. The RPD takes both the prediction error and the variation of observed values into account, providing a metric of model validity that is more objective for comparisons across evaluated properties and studies. The R^2 is a measure of how well observed outcomes are reproduced by the model, based on the proportion of total variation explained by the model. The R^2 also allow comparison across evaluated properties and studies, but is highly dependent on a property's range of values.

To determine the wavenumber importance for each soil property (i.e., main selected spectral variables), specific methods were used for each multivariate regression model (PLS or SVM). For PLS, loadings vectors of the first two latent variables were extracted from the PLS models and plotted against the wavenumbers. Interpretation of this method is that the more intense negative or positive loadings at specific wavenumbers indicate more important wavenumbers for prediction development. In the support vector machines (SVM) models, a recursive feature elimination algorithm was used (R package 'caret', Kuhn, 2018). This approach implements backward selection of predictors (wavenumbers) based on predictor importance ranking from the first to the least important wavenumbers. The recursive feature elimination was processed using a 10-fold cross-validation with a 75% calibration set to 25% leave-group out cross-validation. For both PLS and SVM, the same spectral pretreatment method selected to develop the prediction model of each soil property was used to determine the wavenumber importance.

2.7. Tuning support vector machines models

Support vector machines multivariate regression models performance (accuracy) depends on selecting the kernel and setting the parameters *C* and ε. To identify appropriate parameters, we first selected the best kernel based on prediction accuracy using typically implemented parameters in the ε-insensitive error function (with $\varepsilon = 0.1$, a value of 1 for the penalization factor C, these are the default parameters implemented in R package 'e1071'), and then tested different parameters *C* and ε for the selected best performing kernel. For all soil properties in both datasets, the Gaussian Radial Basis Function (radial) was selected based on the best model performance (i.e., < RMSE, > RPD, and > R^2). After selecting the kernel, the best combination of *C* and ε was searched using an error grid set on a hyperparameter range (*C* and ε) (R package 'e1071', Meyer et al., 2015). We conducted preliminary tests to set the final range of tested parameters for each dataset, and selected ranges based on model performance. We tested *C* up to 100 and ε up to 10. The final range of tested parameters was set for *C* from 0 to 32 and ε from 0.001 to 1.0 for the USA dataset and *C* from 0 to 32 and ε from 0.001 to 0.5 for the Tanzania dataset. The grid search was conducted by a 10-fold cross validation, and hyperparameters were selected based on the best model performance (lowest $RMSE_{CV}$).

2.8. Data processing and statistical analyses

Data was processed and analyzed using R version 3.3.3 (R Foundation for Statistical Computing, Vienna, Austria) using the packages 'chemometrics' (Varmuza and Filzmoser, 2009), 'ChemometricsWithR' (Wehrens, 2011), 'e1071' (Meyer et al., 2015), 'pls' (Mevik and Wehrens, 2007), 'prospectr' (Stevens and Ramirez-Lopez, 2015), and 'stats' (R Core Team, 2016).

3. Results

The two soil datasets evaluated in this study (USA and Tanzania) entailed a wide range of soil physical, chemical, and biological properties (Table 1). The USA dataset had wider ranges for clay, sand, pH, and TOC than Tanzania. Permanganate oxidizable carbon had more similar ranges between the two datasets, but the distribution of values across ranges differed as it can be observed by the quartiles, median,

and mean POXC values (Table 1). The USA Midwest soil dataset had a larger range of spectral variability than Tanzania soil dataset, as illustrated by the dispersion of PCA scores (Supplementary Fig. 1). The PCA loading vectors (PC1 and PC2) had a distinct composition of spectral features between the two datasets indicating that dataset-specific key wavenumbers were explaining most of the spectral variability (Supplementary Fig. 2).

The main variation of user defined parameters in selected PLS models across soil properties and datasets was the number of selected latent variables, and these varied from one latent variable for POXC in the Tanzania dataset to eight latent variables for sand in the USA dataset (Table 2). In general, most of the best performing PLS models were obtained with the *classical* algorithm and scaled spectra. Exceptions in which other algorithms achieved best predictions were for TOC using the *kernelpls* algorithm and for POXC using the SIMPLS algorithm in the Tanzania dataset, and for TOC using the SIMPLS algorithm and no spectra scaling in the USA dataset. Non-tuned SVM models used fixed ε and C parameters of 0.1 and 1.0 respectively. However, there was variation across soil properties on the number of support vectors, which varied from 157 support vectors for TOC in the USA dataset, to 333 support vectors for TOC in the Tanzania dataset.

The SVM models were optimized based on the parameters ε and C using an error grid set on a specified hyperparameters range, as exemplified for clay concentration from Tanzania soils (Fig. 2). For the USA dataset, the selected parameter of C ranged from 4.28 to 17.15 across all measured soil properties and the selected parameter of ε was generally 0.001, except for soil pH (ε = 0.201) (Table 3 and Supplementary Fig. 3). The Tanzania dataset had C between 1.07 and 4.29 and ε ranging from 0.001 (clay and sand) to 0.201 (pH and POXC) (Table 3 and Supplementary Fig. 4). In both datasets, tuning SVM increased the number of support vectors, except for the tuned models with larger ε (i.e., > 0.2) for which a lower number of support vectors was obtained when compared to non-tuned SVM models.

The predictive response of SVM outperformed PLS for all soil properties in both datasets, and the magnitude of improvement depended on the soil property in each data set (Figs. 3 and 4, Tables 4 and 5). The SVM optimization of the parameters ε and C, searched using a cross-validated error grid ($RMSE_{CV}$) set on a specified hyperparameters range (Fig. 2, and Supplementary Figs. 3 and 4), maintained or improved prediction accuracies when compared to non-tuned SVM models, based on $RMSE_V$, RPD_V, and R^2_V statistical coefficients of independent validation sets (Figs. 3 and 4, Tables 4 and 5). As hypothesized, these optimized SVM models always improved model calibration statistics.

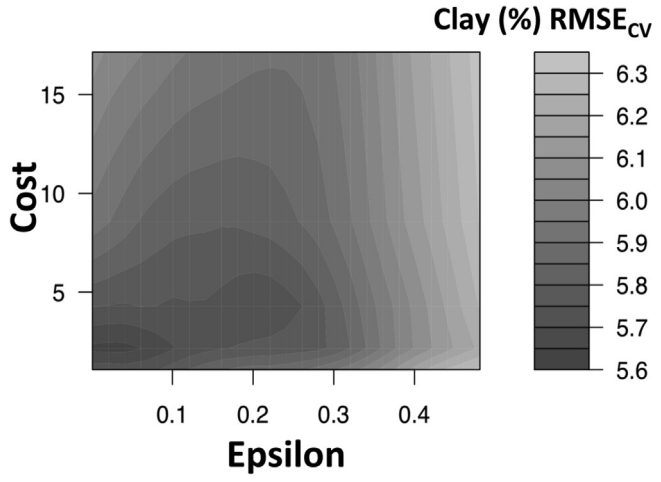

Fig. 2. Hyperparameters search grid to optimize support vector machines regression models predicting soil clay concentration in Tanzania soils from diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS). The other soil properties search grids are in Supplementary Figs. 3 and 4.

The wavenumber importance was measured by different methods for PLS or SVM (Fig. 5). Several similarities in wavenumber importance existed between PLS and SVM, but each model exhibited distinct use of wavenumbers depending on the soil property and dataset. For example, in the Midwest USA dataset, prediction of clay content for both multivariate models (PLS and SVM) drew upon, among other regions or the MIR, the downfield of the MIR (~4000–3777 cm^{-1}). To our knowledge, this region does not express absorbance features from soil functional groups, and the importance could be related with the overall spectra reflectiveness (absorbance values overall wavenumbers). The importance of this region in both models was evidenced by (i) the intense negative values of PLS latent variable (LV) loadings in LV1 and (ii) the recursively selected important wavenumbers of SVM. The region at 3700 cm^{-1} to 3200 cm^{-1} of the same spectrum was a positive loading in the PLS LV1, and the same region is expressed in the SVM models as important wavenumbers (darker tones). This spectral region corresponds to the functional group O–H of hydroxyl stretching (kaolinite and others) (3723–3686 cm^{-1}, Russell, 1987) and Si–O functional group of 2:1 layer aluminosilicates (3686–3565 cm^{-1}, Nguyen et al., 1991). Another similar behavior between the PLS LV1 loadings and SVM important wavenumbers for clay in USA soils can be observed for wavenumbers across 1400 cm^{-1} to 1200 cm^{-1}, a region that contains peaks of symmetric –COO– stretch and/or –CH bending of aliphatics. An example of a poorly defined relationship between important regions of PLS and SVM can be observed for TOC in the Tanzania dataset. The frequencies between ~2600 cm^{-1} and 2400 cm^{-1} were considered important for SVM but not so evidently for PLS, which had one of the noisiest loading vectors distributions among all soil variables. This region corresponds to the functional group CO_3 of calcite (peaks ranging from 2650 to 2420 cm^{-1}, Nguyen et al., 1991).

4. Discussion

Predicting soil properties with mid-DRIFTS has clearly demonstrated potential, but many methodological decisions will impact predictive performance. Regardless of sample preparation and spectra acquisition (Deiss et al., 2019b), chemometric modeling can have major impacts on prediction accuracy. Soil measurements in mid-DRIFTS rely on reference data from traditional analytic methods to calibrate models for a specific set of soils, but can later facilitate estimation of soil properties due to the reduced time, labor and costs associated with the technique (Soriano-Disla et al., 2014; Nocita et al., 2015). Once reference measured data is used to calibrate mid-DRIFTS models, the soil measurements must be accurate, precise, and reproducible following rigorous laboratory standards, so that calibrated models can be reliable. After training and validating a model, predictions can be performed on new soil samples using only the spectra, but these new samples must be spectrally similar to the spectral library used during modeling. Important characteristics to be taken into consideration are edaphic properties that can affect spectral features such as soil order, soil mineralogy, soil property distribution (e.g., ranges and quartiles), soil depth, and any other edaphic source of variation in a dataset. Potential outliers can be detected based on spectral properties, for example, by specifying a threshold on multidimensional spectral dispersions based on Euclidean/Mahalanobis distances of a spectral matrix (e.g., Mirzaeitalarposhti et al., 2017).

In chemometrics, prediction accuracy will mostly depend on soil properties that affect reflectance characteristics, and a combination of spectral treatments and multivariate modeling. Spectral quality attributes such as spatial resolution, signal-to-noise ratio, and presence of spectral artifacts (Kimber and Kazarian, 2017) may also affect the data analytical process. Our results showed that SVM outperformed PLS for all predictions, and tuning SVM models maintained or improved accuracy in relation to non-tuned SVM models (Figs. 3 and 4 and Tables 4 and 5). Each soil property predicted from both datasets had specific

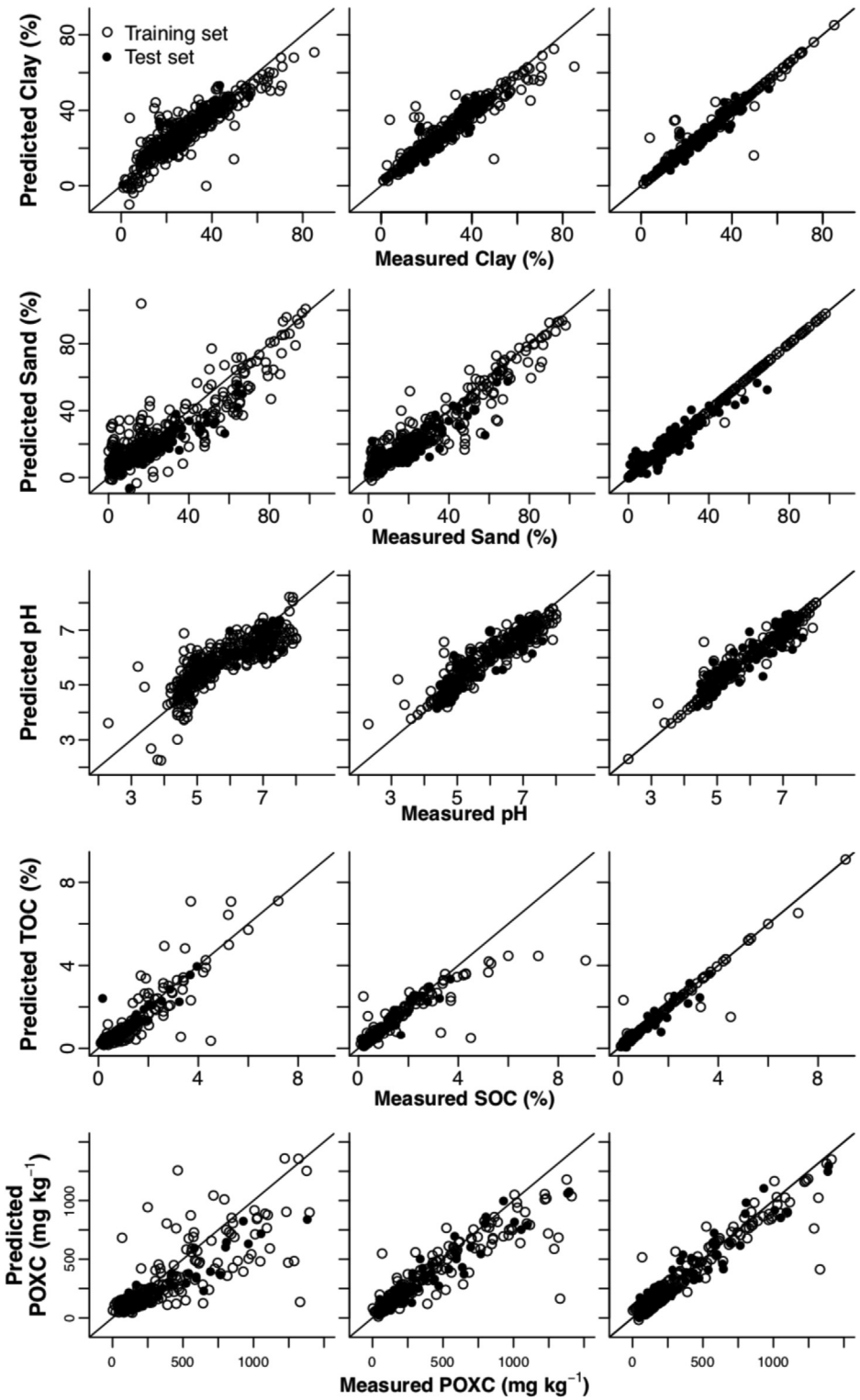

Fig. 3. USA Midwest soils scatter plots of measured versus predicted values using linear (partial least squares: PLS) and non-linear (SVM: support vector machines and tSVM: tuned SVM) multivariate regression models in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS). Regression models were trained on calibration sets (75% of dataset) and tested on independent validation sets (25% of dataset). TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

spectral treatments, SVM model configurations, and tuning parameters (Table 3, Fig. 2, and Supplementary Figs. 3 and 4), suggesting model optimization is a soil property- and dataset-specific process that can improve prediction accuracy of mid-DRIFTS.

During the SVM parameter optimization process, there was a trade-off between the size of ε (insensitivity zone) and the penalty parameter C (Table 3). Reducing the insensitivity zone increased the size of C for most soil properties, and generally increased the number of support vectors. This trade-off was expected based on the ε-insensitive loss function behavior (Smola and Schölkopf, 2004). Values are expected to be more distant from and/or out of the insensitive zone (greater C) by decreasing the size of the sensitive zone (smaller ε) (Fig. 1). These

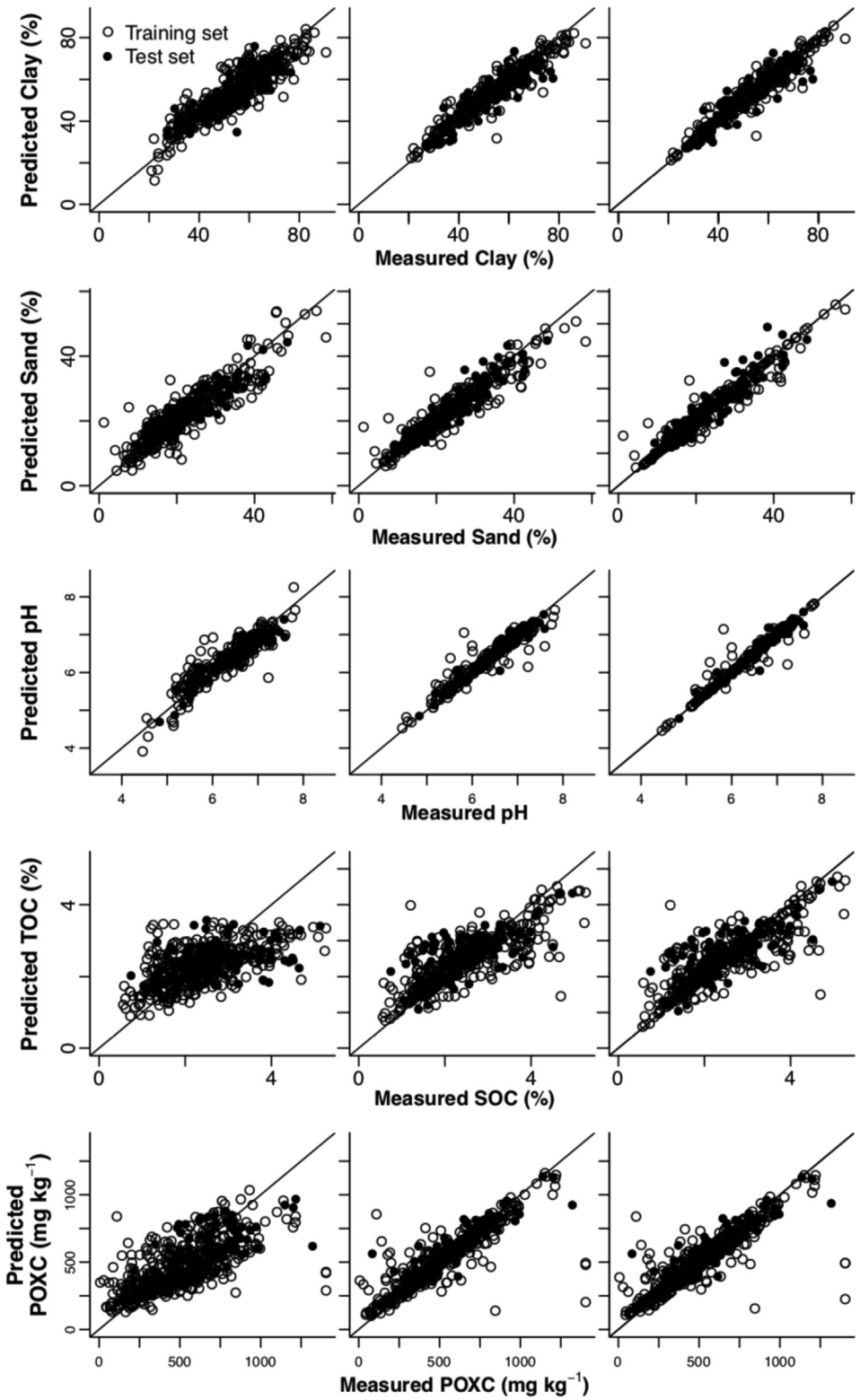

Fig. 4. Tanzania soils scatter plots of measured versus predicted values using linear (partial least squares: PLS) and non-linear (SVM: support vector machines and tSVM: tuned SVM) multivariate regression models in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS). Regression models were trained on calibration sets (75% of dataset) and tested on independent validation sets (25% of dataset). TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

optimum parameters (ε and C) were searched using a cross-validated error grid ($RMSE_{CV}$) set on a specified hyperparameters range (Fig. 2, and Supplementary Figs. 3 and 4), and this may be an effective approach to identify optimum parameters for SVM to develop predictions of soil properties using mid-DRIFTS, once prediction accuracy was often improved (Figs. 3 and 4 and Tables 4 and 5). When the ε was increased in relation to non-tuned SVM (i.e., pH in the USA dataset and TOC in the Tanzania dataset) (Table 3), the numbers of support vectors were reduced revealing another potential trade off of SVM optimization. The number of support vectors indicates the number of training samples to

Table 4
Prediction performance of USA Midwest soils using linear (partial least squares: PLS) and non-linear (SVM: support vector machines and tSVM: tuned SVM) multivariate regression models in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS).

Soil property	Model	Calibration set (75% of dataset)			Validation set (25% of dataset)		
		RMSE	RPD	R^2	$RMSE_V$	RPD_V	R^2_V
Clay (%) (n = 400)	PLS	7.27	1.81	0.81	4.58	2.38	0.82
	SVM	5.36	2.72	0.89	3.81	2.77	0.87
	tSVM	3.21	5.04	0.96	3.22	3.31	0.91
Sand (%) (n = 400)	PLS	9.66	1.87	0.85	7.70	2.02	0.75
	SVM	7.97	2.82	0.90	7.45	1.59	0.76
	tSVM	1.00	25.0	0.99	5.24	2.54	0.87
pH (n = 400)	PLS	0.54	1.77	0.77	0.44	2.27	0.80
	SVM	0.40	2.44	0.87	0.40	2.33	0.84
	tSVM	0.28	3.62	0.94	0.36	2.53	087
TOC (%) (n = 399)	PLS[a]	0.50	2.49	0.82	0.29	2.42	0.85
	SVM	0.51	1.76	0.81	0.18	3.69	0.93
	tSVM	0.22	5.03	0.96	0.19	3.60	0.93
POXC (mg kg^{-1}) (n = 400)	PLS[a]	207	1.47	0.56	123	1.25	0.74
	SVM	122	2.10	0.84	69	3.64	0.93
	tSVM	89	3.32	0.92	74	3.79	0.92

[a] Statistical coefficients were determined on logarithmic back-transformed data. RMSE: root mean squared error, RPD: residual prediction deviation, and R^2: coefficient of determination. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

encode the calibration set, and reducing the number of support vectors minimize chances of model over-fitting. Moreover, increasing the number of support vectors can add significant amount of time while modeling and/or running predictions. However, this additional time can be beneficial if the prediction accuracy is improved.

Setting guidelines on SVM regression parameterization is important because these parameters are user-defined inputs and there is lack of information about specific values to use when predicting soil properties with mid-DRIFTS. To our knowledge, this is the first use of SVM regression tuning to enhance prediction of soil properties from mid-DRIFTS. Grid search for hyperparameters has been used in near-infrared spectroscopy and other soil applications non-related to spectroscopy. Predicting TOC with near-infrared spectroscopy and SVM, Chen et al. (2015) tested the effect of tuning other SVM parameters (*gamma* and *sigma*), different than those included in our study (ε and C). Optimizing SVM using grid search for *gamma* and *sigma* improved accuracy of TOC predictions (6.6% reduction of $RMSE_V$) (Chen et al., 2015). Grid search has also been used in other soil applications beyond infrared spectroscopy such as prediction of soil pore-water pressure, soil heavy metal concentrations, and soil water retention potential (Mirhosseini, 2017; Babangida et al., 2016; Wu et al., 2016; Khlosi et al., 2016).

Optimizing chemometrics in mid-DRIFTS allows better extracting information from spectra to more accurately predict soil physical, chemical, and biological properties. However, modeling is susceptible to generating meaningless outputs and non-linear models can be easily over-fitted. Chemometrics is generally a naive modeling approach because the procedure does not systematically consider specific peaks in the input spectra if users do not assign weights to wavenumbers or truncate spectra. These models identify frequencies (wavenumbers) of the spectrum that are more related to the variation of a certain soil property, regardless of the kind of soil organic or inorganic functional group corresponding to the wavenumbers, and use these wavenumbers to predict the soil property. Adding to that, some properties can be totally or partially predicted in multivariate models because of their correlation or covariation with other soil properties (Chang et al., 2001; Stenberg et al., 2010, Reeves, 2010). We found that PLS and SVM did not necessarily use the same wavenumbers to develop predictions for each soil property, and there were occasions where these most important wavenumbers did not directly relate to the property of interest (Fig. 5). For example, both PLS and SVM predictions drew upon the downfield region of the MIR (~4000–3777 cm^{-1}), but this region does not have specified peaks defined by soil functional groups (e.g.,

Table 5
Prediction performance of Tanzania soils using linear (partial least squares: PLS) and non-linear (SVM: support vector machines and tSVM: tuned SVM) multivariate regression models in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS).

Soil property	Model	Calibration set (75% of dataset)			Validation set (25% of dataset)		
		RMSE	RPD	R^2	$RMSE_V$	RPD_V	R^2_V
Clay (%) (n = 533)	PLS	5.26	1.97	0.84	5.18	2.13	0.78
	SVM	4.02	2.98	0.90	4.58	2.13	0.81
	tSVM	3.24	3.88	0.94	4.66	2.15	0.81
Sand (%) (n = 533)	PLS	3.95	1.95	0.81	3.4	2.31	0.81
	SVM	2.87	2.83	0.90	2.73	2.70	0.88
	tSVM	2.00	4.36	0.95	2.83	2.79	0.87
pH (n = 335)	PLS	0.28	1.68	0.82	0.21	2.80	0.87
	SVM	0.20	3.04	0.91	0.13	4.10	0.94
	tSVM	0.16	4.14	0.94	0.13	4.41	0.95
TOC (%) (n = 533)	PLS[a]	0.72	0.74	0.35	0.86	0.53	0.13
	SVM	0.56	1.25	0.64	0.59	1.03	0.54
	tSVM	0.50	1.52	0.71	0.56	1.13	0.57
POXC (mg kg^{-1}) (n = 532)	PLS[a]	494	0.38	0.40	165	1.10	0.46
	SVM	133	1.62	0.73	88	2.19	0.84
	tSVM	132	1.61	0.73	89	2.18	0.84

[a] Statistical coefficients were determined on logarithmic back-transformed data. RMSE: root mean squared error, RPD: residual prediction deviation, and R^2: coefficient of determination. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon.

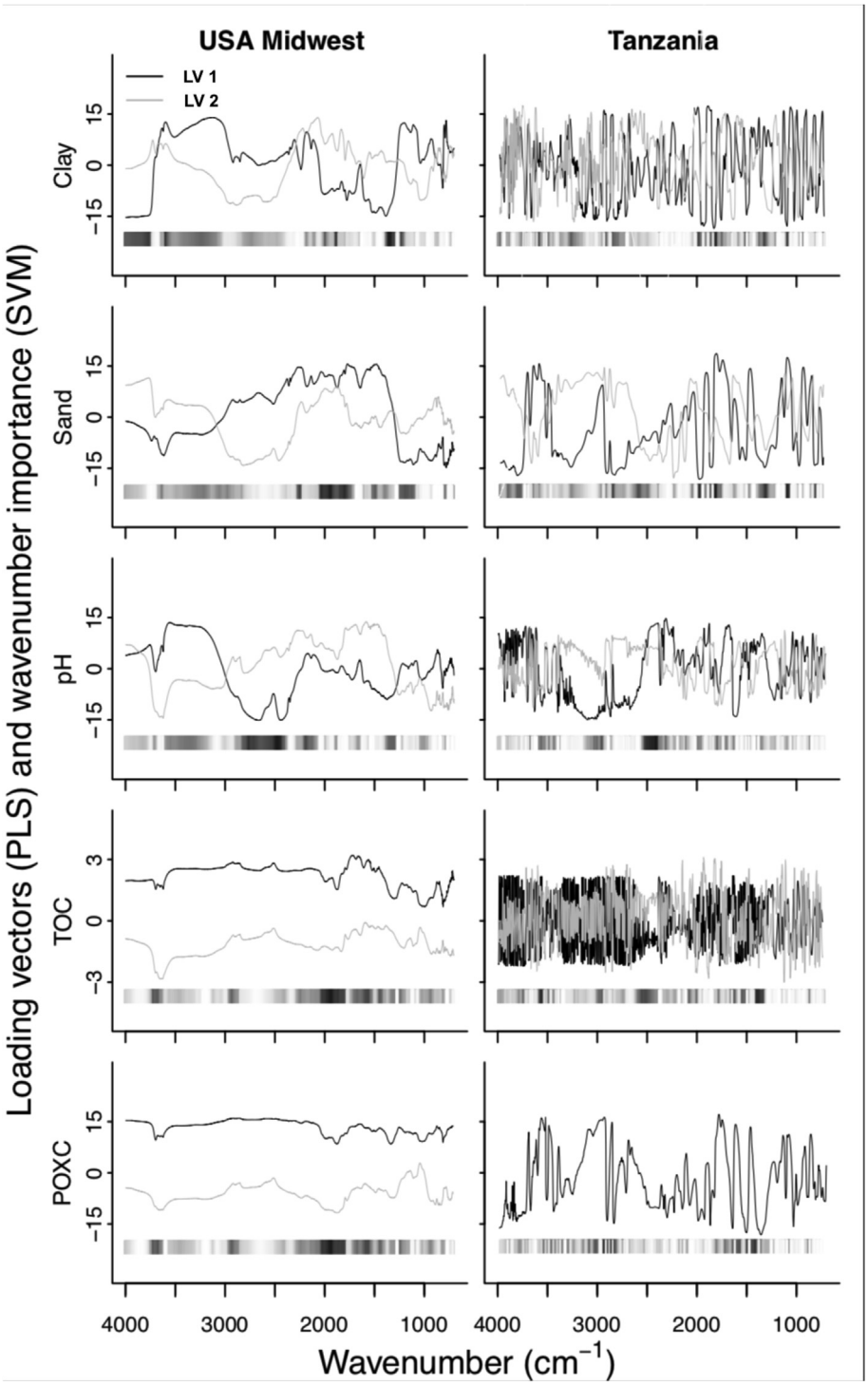

Fig. 5. Wavenumber importance of support vector machines (SVM) regression models determined by recursive feature selection (grey scale) and the first two latent variables loading vectors (LV1 and LV2) of partial least squares (PLS) regression models in diffuse reflectance infrared Fourier transform spectroscopy (mid-DRIFTS). Soils were from USA Midwest and Tanzania. TOC: Total organic carbon. POXC: Permanganate oxidizable carbon. One latent variable was used for POXC in the Tanzania dataset and thus they have only the LV1 line. In the Tanzania dataset, TOC PLS models used a different algorithm (kernelpls) than the other soil properties/datasets (classical or simpls algorithms), and the loadings were magnified 5 times so they were in a comparable scale to the other loading vectors.

Kodama, 1985; Parikh et al., 2014). Moreover, the relatively high number of support vectors in relation to the size of the calibration set (Table 3) and sand calibration statistical outcomes in the USA dataset (Table 4) indicates that the datasets can still be broadened to cover a greater extent of the soil spectra variability to reduce over-fitting and increase robustness of those tuned models to develop predictions to new samples. For these reasons, independent validation sets are indispensable; once kernel models are sensitive to over-fitting (Ali et al., 2015) and potentially do not translate predictive prediction accuracy from the calibration sets to independent test sets or new samples.

Several multivariate regression models have been used to predict soil properties in mid-DRIFTS. However, there is no consensus on what model class improves accuracy when measuring soil properties. At same time, this does not imply that a single modeling approach will work in every particular case or even as a universal model class, but finding ways to chemometrically deal with complex soil spectra may allow improving prediction robustness in mid-DRIFTS of soils. Partial least squares models are generally easier to derive and interpret (e.g., PLS LV loadings), and are insensitive to colinearity (Haaland and Thomas, 1988; Gholizadeh et al., 2013). On the other hand, non-linear multivariate regression methods can be complex to interpret (Soriano-Disla et al. 2014) and they are not always available in commercially available spectral processing software. Previous multivariate regression models comparisons in mid-DRIFTS of soils have shown variable responses for

different datasets and soil properties. Comparing multivariate regression models in soils from the Ribeirão Inhaúma basin, Brazil (n = 184), Campbell et al. (2018) obtained greater prediction accuracies with PLS for TOC and Mehlich-1 extractable phosphorus whereas SVM performed better for clay. Jia et al. (2017) obtained greater accuracy with SVM when compared to PLS predicting TOC in soils from an alpine landscape on the Qinghai–Tibet Plateau (n = 330). In oak forest soils across East China (n = 140), Kang et al. (2017) found that SVM has similar or better prediction than PLS for several organic carbon compounds. Comparing two non-linear models in representative soil profiles from Brazil (n = 1117 from 367 soil profiles), Souza et al. (2012) obtained greater prediction performances for total organic matter using SVM than neural networks. In most cases, non-linear models such as SVM or neural networks performed better than or comparable to linear models including PLS. Less commonly, better prediction performances were obtained with linear than non-linear models. Finding non-linearities in the relationship between spectral characteristics and soil properties distributions is expected for heterogeneous, edaphically diverse soil sets (e.g., Calderón et al., 2017). There is a wide variability of soil reflectance patterns in mid-infrared frequencies of the spectrum and that could be benefiting non-linear models in these predictions.

Support vector machines regression models have been underused compared to PLS models (Viscarra-Rossel et al., 2006; Gholizadeh et al., 2013) but the emergence of SVM tuning stands to increase the utility of this chemometric approach. Tuning SVM models will add more computational demand and time to the modeling process depending on dataset characteristics (e.g., spectral resolution and number of samples), number of support vectors, and/or cross-validation configurations. Support vector machines regression model tuning should be done for each soil property after selecting an optimal spectral treatment and a SVM kernel. While this may require a greater initial investment in model development, once optimal parameters have been found similar time and computational demand can be expected as in non-tuned SVM models, presenting a worthwhile investment that can improve prediction accuracy. These parameters are user defined inputs and can be specified, *a priori*, when developing predictions for new samples. Tuning SVM can also be done for other parameters such as *gamma* and *sigma* (e.g., Chen et al., 2015), and more research is needed to fully vet SVM models parameterization for mid-DRIFTS soil analysis.

5. Conclusion

Optimizing chemometric models by curating each prediction based on the combination of spectral treatments, model selection and configurations, and tuning parameters may improve prediction accuracy of mid-DRIFTS to predict soil properties. Non-linear models (support vector machines) outperformed linear models (partial least squares) for all tested soil properties (sand, clay, pH, TOC, POXC) in soils from both Tanzania and USA Midwest. Specific spectral treatments were used for each prediction, and the Gaussian Radial Basis Function (radial) was the most accurate kernel in support vector machines regression models. Tuning support vector machines models based on the parameters C and ε maintained or improved accuracy in relation to non-tuned support vector machines models. Therefore, tuned support vector machines regression models may be used as a way to derive predictions from the complex relationships between a soil property and the mid-infrared spectrum. Vetting modeling strategies in mid-DRIFTS allows better using information from spectra to more accurately predict soil physical, chemical, and biological properties.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Foundation for Food and Agricultural Research and the School of Environment and Natural Resources at Ohio State University. We sincerely thank the anonymous reviewers for the comprehensive evaluations of our manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2020.114227.

References

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7, 221–236. https://doi.org/10.3390/rs71215841.

Appel, P., Moller, A., Schenk, V., 1998. High-pressure granulite facies metamorphism in the Pan-African belt of eastern Tanzania: P-T–t evidence against granulite formation by continent collision. J. Metamorph. Geol. 16, 491–509. https://doi.org/10.1111/j.1525-1314.1998.00150.x.

Babangida, N.M., Mustafa, M.R.U., Yusuf, K.W., Isa, M.H., 2016. Prediction of pore-water pressure response to rainfall using support vector regression. Hydrogeol. J. 24, 1821–1833. https://doi.org/10.1007/s10040-016-1429-4.

Burt, R., 2011. Soil Survey Laboratory Information Manual. Soil Survey Investigations Report No. 45 (Version 2.0). United States Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center, Lincoln. 530p.

Calderón, F.J., Culman, S., Six, J., Franzluebbers, A.J., Schipanski, M., Beniston, J., Grandy, S., Kong, A.Y.Y., 2017. Quantification of soil permanganate oxidizable C (POXC) using infrared spectroscopy. Soil Sci. Soc. Am. J. 81, 277–288. https://doi.org/10.2136/sssaj2016.07.0216.

Campbell, P.M.M., Fernandes-Filho, E.I., Francelino, M.R., Demattê, J.A.M., Pereira, M.G., Guimarães, C.C.B., Pinto, L.A.S.R., 2018. Digital soil mapping of soil properties in the “Mar de Morros” environment using spectral data. Rev. Bras. Cienc. Solo 42, e0170413. https://doi.org/10.1590/18069657rbcs20170413.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490. https://doi.org/10.2136/sssaj2001.652480x.

Chen, H.-Z., Shi, K., Cai, K., Xu, L.-L., Feng, Q.-X., 2015. Investigation of sample partitioning in quantitative near-infrared analysis of soil organic carbon based on parametric LS-SVR modeling. RSC Adv. 5, 80612–80619. https://doi.org/10.1039/C5RA12468A.

Culman, S.W., Freeman, M., Snapp, S.S., 2012. Procedure for the determination of permanganate oxidizable carbon. In: Kellogg Biological Station-Long Term Ecological Research Protocols. Hickory Corners, MI.

Deiss, L., Demyan, M.S., Culman, S.W., 2019a. Grinding and sample replication do not improve mid-DRIFTS predictions of soil properties. Soil Sci. Soc. Am. J. https://doi.org/10.2136/sssaj2019.05.0147. (IN PRESS).

Deiss, L., Margenot, A.J., Demyan, M.S., Culman, S.W., 2019b. Optimizing acquisition parameters in diffuse reflectance infrared Fourier transform spectroscopy of soils. Soil Sci. Soc. Am. J. https://doi.org/10.2136/sssaj2019.05.0148. (IN PRESS).

Dreimanis, A., 1962. Quantitative gasometric determination of calcite and dolomite by using Chittick apparatus. J. Sediment. Res. 32, 520–529. https://doi.org/10.1306/74D70D08-2B21-11D7-8648000102C1865D.

Fearn, T., 2008. The interaction between standard normal variate and derivatives. NIR news 19, 16–17. https://doi.org/10.1255/nirn.1098.

Gholizadeh, A., Luboš, B., Saberioon, M., Vašát, R., 2013. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. Appl. Spectrosc. 67, 1349–1362. https://doi.org/10.1366/13-07288.

Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60, 1193–1202.

Jia, X., Chen, S., Yang, Y., Zhou, L., Yu, W., Shi, Z., 2017. Organic carbon prediction in soil cores using VNIR and MIR techniques in an alpine landscape. Sci. Rep. 7, 1–9. https://doi.org/10.1038/s41598-017-02061-z.

Kang, H., Gao, H., Yu, W., 2017. Evaluation of spectral pretreatments, spectral range, and regression methods for quantitative spectroscopic analysis of soil organic carbon composition. Spectrosc. Lett. 50, 143–149. https://doi.org/10.1080/00387010.2017.1297956.

Karatzoglou, A., Meyer, D., Hornik, K., 2006. Support vector machines in R. J. Stat. Softw. 15, 1–28. https://doi.org/10.18637/jss.v015.i09.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148. https://doi.org/10.1080/00401706.1969.10490666.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W.M., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur. J. Soil Sci. 67, 276–284. https://doi.org/10.1111/ejss.12345.

Kimber, J.A., Kazarian, S.G., 2017. Spectroscopic imaging of biomaterials and biological systems with FTIR microscopy or with quantum cascade lasers. Anal. Bioanal. Chem. 409, 5813–5820. https://doi.org/10.1007/s00216-017-0574-5.

Kodama, H., 1985. Infrared Spectra of Minerals. Reference Guide to Identification and

Characterization of Minerals for the Study of Soils. Technical Bulletin 1985-1E, Research Branch, Agriculture Canada, Ottawa.
Kuhn, M., 2018. The Package 'caret': Reference manual. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/caret/caret.pdf (accessed 14 January 2019).
Martens, H., Naes, T., 1989. Multivariate Calibration. John Wiley and Sons, Chichester https://doi.org/10.1002/cem.1180040607.
Massawe, B.H.J., Winowiecki, L., Meliyo, J.L., Mbogoni, J.D.J., Msanya, B.M., Kimaro, D., Deckers, J., Gulinck, H., Lyamchai, C., Sayula, G., Msoka, E., Vagen, T.-G., Brush, G., Jelinski, N.A., 2017. Assessing drivers of soil properties and classification in the West Usambara mountains, Tanzania. Geoderma Reg. 11, 141–154. https://doi.org/10.1016/j.geodrs.2017.10.002.
Mevik, B.-H., Wehrens, R., 2007. The pls package: principal component and partial least squares regression in R. J. Stat. Softw. 18, 1–23. https://doi.org/10.18637/jss.v018.i02.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2015. e1071: Misc functions of the Department of Statistics. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=e1071 (accessed 21 June 2017).
Mirhosseini, R.T., 2017. Seismic response of soil-structure interaction using the support vector regression. Struct. Eng. Mech. 63, 115–124. https://doi.org/10.12989/sem.2017.63.1.115.
Mirzaeitalarposhti, R., Demyan, S.M., Rasche, F., Cadisch, G., Müller, T., 2017. Mid-infrared spectroscopy to support regional-scale digital soil mapping on selected croplands of South-West Germany. Catena 149, 283–293. https://doi.org/10.1016/j.catena.2016.10.001.
Niemeyer, J., Chen, Y., Bollag, J.-M., 1992. Characterization of humic acid, composts, and peat by diffuse reflectance Fourier-transform infrared spectroscopy. Soil Sci. Soc. Am. J. 56, 135–140. https://doi.org/10.2136/sssaj1992.03615995005600010021x.
Nguyen, T., Janik, L.J., Raupach, M., 1991. Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. Soil Res. 29, 49–67. https://doi.org/10.1071/SR9910049.
Nocita, M., Stevens, A., Wesemael, B. Van, Aitkenhead, M., Bachmann, M., Barth, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-lopez, L., Robertson, J., Sakai, H., Soriano-disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. Adv. Agron. 132, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002.
Parikh, S.J., Goyne, K.W., Margenot, A.J., Mukome, F.N.D., Calderón, F.J., 2014. Soil chemical insights provided through vibrational spectroscopy. Adv. Agron. 126, 1–148. https://doi.org/10.1016/B978-0-12-800132-5.00001-8.
R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/ (accessed 21 June 2017).
Reeves, J.B., 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma 158, 3–14. https://doi.org/10.1016/j.geoderma.2009.04.005.
Russell, J.D., 1987. Infrared spectroscopy of inorganic compounds. In: Willis, H. (Ed.), Laboratory Methods in Infrared Spectroscopy. Wiley, New York.
Schölkopf, B., Smola, A.J., 2002. Learning with Kernels. MIT Press.
Shepherd, K.D., Walsh, M.G., 2007. Infrared spectroscopy: enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. J. Near Infrared Spectrosc. 15, 1–19. https://doi.org/10.1255/jnirs.716.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. https://doi.org/10.1023/B:Stco. 0000035301.49549.88.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081.
Souza, D.M., Madari, B.E., Guimarães, F.F., 2012. Aplicação de técnicas multivariadas e inteligência artificial na análise de espectros de infravermelho para determinação de matéria orgânica em amostras de solos. Quim. Nova 35, 1738–1745. https://doi.org/10.1590/S0100-40422012000900007.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7.
Stevens, A., Ramirez-Lopez, L., 2015. An introduction to the prospectr package. GitHub. http://antoinestevens.github.io/prospectr/ (accessed 21 June 2017).
Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., Shepherd, K.D., 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Sci. Soc. Am. J. 74, 1792–1799. https://doi.org/10.2136/sssaj2009.0218.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, New York.
Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton https://doi.org/10.1201/9781420059496.
Viscarra-Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.
Wehrens, R., 2011. Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences. Springer, Berlin https://doi.org/10.1007/978-3-642-17841-2.
Weil, R.R., Islam, K.R., Stine, M.A., Gruver, J.B., Samson-Liebig, S.E., 2003. Estimating active carbon for soil quality assessment: a simplified method for laboratory and field use. Am. J. Alternative. Agric. 18, 3–17. https://doi.org/10.1079/AJAA200228.
Winowiecki, L., Vågen, T.-G., Massawe, B., Jelinski, N.A., Lyamchai, C., Sayula, G., Msoka, E., 2016. Landscape-scale variability of soil health indicators: effects of cultivation on soil organic carbon in the Usambara Mountains of Tanzania. Nutr. Cycl. Agroecosystems 105, 263–274. https://doi.org/10.1007/s10705-015-9750-1.
Wu, J., Teng, Y., Chen, H., Li, J., 2016. Machine-learning models for on-site estimation of background concentrations of arsenic in soils using soil formation factors. J. Soils Sediment. 16, 1787–1797. https://doi.org/10.1007/s11368-016-1374-9.