Plant Archives Vol. 26, No. 1, 2026 pp. 77-87 e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org

DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2026.v26.no.1.010

CHEMOMETRICS: A PREDICTIVE TOOL FOR SOIL HEALTH ASSESSMENT

Omji Nath M.S., Beena V. I., Nayana M. V. and Nithin S.

Department of Soil Science and Agricultural Chemistry, College of Agriculture Vellanikkara, Kerala Agricultural University, Thrissur, Kerala, India. Post code: 680656.

*Corresponding author E-mail: omjinath@gmail.com

(Date of Receiving-19-11-2025; Date of Revision-27-12-2025; Date of Acceptance-06-02-2026)

ABSTRACT

Healthy soil is essential for sustainable agriculture, but traditional laboratory methods for assessing soil quality have significant drawbacks. These methods often use harmful chemicals, require expensive equipment, and take considerable time to complete. Chemometrics provides a practical solution by combining mathematics, statistics, and computer science to interpret spectral data from soil samples. This review explores how chemometrics can be used to assess soil health through various spectroscopic techniques and predictive modelling approaches. Several spectral methods are examined, including visible near-infrared (Vis-NIR), mid-infrared (MIR), Fourier Transform Infrared (FTIR), Attenuated Total Reflectance (ATR), Raman spectroscopy, Laser Induced Breakdown Spectroscopy (LIBS), and Energy Dispersive X-Ray Fluorescence (EDXRF). These techniques allow scientists to quickly analyse soil properties such as organic carbon, nitrogen content, pH, cation exchange capacity, and micronutrients without destroying the sample. Before building prediction models, spectral data must be cleaned using preprocessing steps like smoothing, derivatives, and scatter corrections to remove unwanted noise. Among the statistical tools available, Partial Least Square Regression (PLSR) is most commonly used for developing calibration models, though Random Forest, Support Vector Machine Regression, and Cubist models also show promising results. Model accuracy is evaluated using parameters such as R^2, RMSE, and RPD. The main benefits of chemometrics include speed, low running costs, minimal sample preparation, and reduced environmental impact. However, challenges remain, including soil variability, high initial setup costs, and difficulty in transferring models between locations. Future work should aim to standardize methods and expand predictions to include soil biological properties.

***Key words*:** Spectroscopy, Partial Least Square Regression (PLSR), Soil health, Data preprocessing, Regression, Machine learning.

Introduction

Soil health refers to the ongoing ability of soil to serve as a crucial living environment that supports plants, animals, and humans. It establishes a link between agricultural practices, soil science, policy-making, stakeholder requirements, and sustainable supply-chain management. While traditional soil assessments primarily concentrate on crop yield, contemporary perspectives recognize the broader implications of soil health, encompassing its impact on water quality, climate change, and human well-being. Despite an increasing awareness of the significance of soil biodiversity, the assessment of soil health continues to rely heavily on chemical indicators due to lack of comprehensive functional understanding and effective measurement methods (Lehmann *et al.*, 2020; Patra *et al.*, 2015). Traditional laboratory techniques involve the use of chemical reagents, which can be environmentally unfriendly, necessitate sophisticated equipment, and involve time-consuming and costly protocols. Certain chemicals, such as chromate salts (e.g., potassium dichromate used in estimating soil organic carbon), are carcinogenic. Consequently, there is a need to adopt alternatives that eschew harmful chemical reagents and swiftly predict soil properties through spectral data analysis. Near-infrared (NIR) and mid-infrared (MIR) spectroscopy are increasingly used in soil

science for measuring various soil attributes mainly related to chemical composition (e.g., various forms of carbon, N, P, K contents, cation-exchange capacity (CEC) and pH) but also to some extent, related to physical parameters (e.g., texture, structure, porosity or bulk density) (Rossel *et al.*, 2006; Cecillon *et al.*, 2009). Various statistical tools like Partial Least Square Regression (PLSR), random forest, etc., have hence been found effective in developing models for the prediction of soil properties (Bellon-Maurel *et al.*, 2010). Chemometrics is a scientific discipline between measurement-oriented chemistry and applied statistics. Analytical chemistry is the most significant discipline where chemometrics plays an important role (Adams, 1990; Wilkins, 1990; Voncina, 2009).

Chemometrics

Chemometrics is the branch of analytical chemistry that employs mathematical, statistical, and computer tools to unveil concealed insights within chemical analyses (Brereton, 2003). It represents a fusion of chemistry, mathematics, statistics, and computer applications (Figure 1). The term "chemometrics" was introduced by the Swedish scientist Svante Wold in 1971, serving as the English equivalent of the Swedish word '*kemometri*,' where 'kemo' refers to chemistry and 'metri' denotes measure (Bystrzanowska and Tobiszewski, 2020).

Application areas

Chemometrics finds application in diverse domains where chemical analyses are essential. It plays a role in environmental monitoring, contributing to the analysis of soil, water, and plant samples. Additionally, it is utilized in the examination of various food items such as grains, meat, fruits, dairy products, oils, honey, and wine. Beyond that, chemometrics extends its utility to fields like pharmaceutical analysis and forensic science also (Santos *et al.*, 2019).

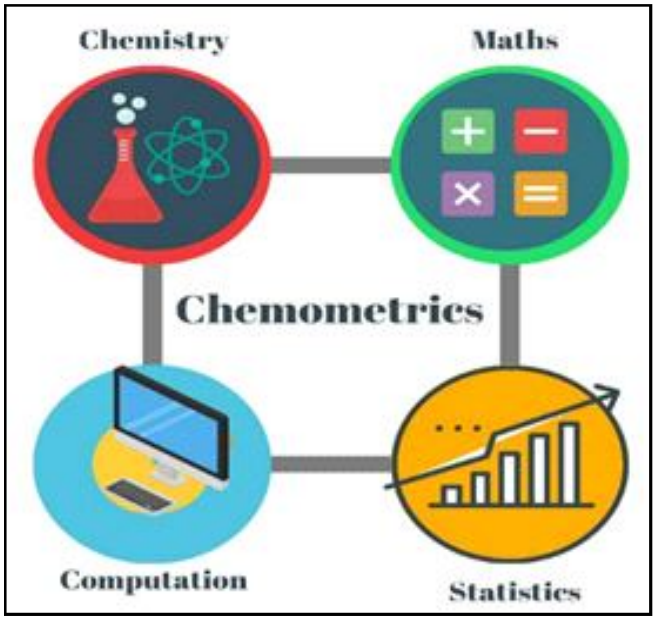

Fig. 1: Disciplines involved in chemometrics.

NIR and MIR spectroscopies play a crucial role as distinct analytical techniques widely employed for the differentiation and validation of herbal medicines. When integrated with chemometrics that involve pattern recognition, the spectra obtained from NIR and MIR can be processed to facilitate a more straightforward interpretation, assisting in decisions related to practices of adulteration. The amalgamation of NIR and MIR spectra with chemometrics offers efficient and swift methods for distinguishing and authenticating herbal medicines (Rohman *et al.*, 2019).

In the classification of *Dendrobium officinale*, a tonic herb commonly used in traditional Chinese medicine, Fourier Transform MIR (FT-MIR) with attenuated total reflectance (ATR) has been applied. The variables for constructing the classification model encompassed MIR spectra within the wavenumbers of 4,000–550 cm^{-1}. The Random Forest model exhibited the ability to differentiate *D. officinale* across various harvesting periods, achieving accuracy levels of 94.44% and 97.92% in the calibration and validation sets, respectively (Wang *et al.*, 2018).

Steps in chemometrics

Chemometrics involves utilizing chemical analytical data alongside diverse machine learning tools, enabling the prediction of soil properties from spectral data once

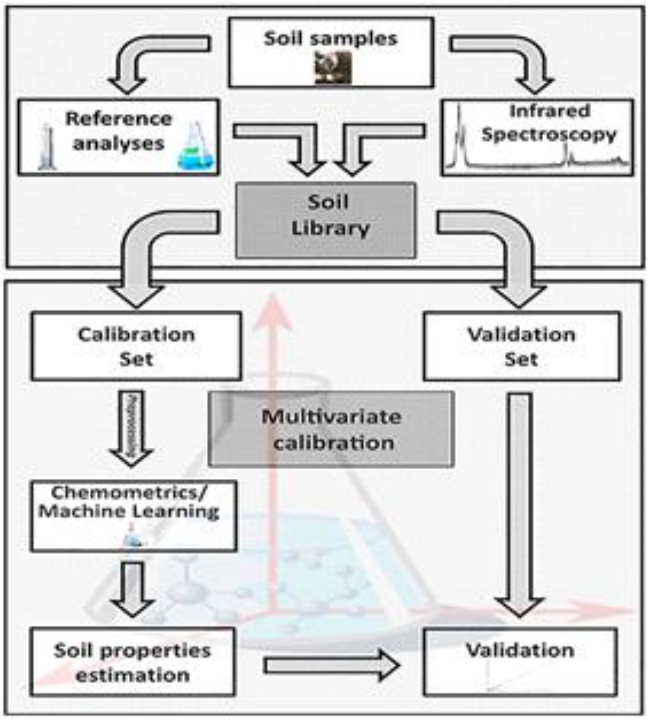

Fig. 2: Steps in chemometrics.

the tools are appropriately trained. The initial stages of chemometrics closely resemble those of conventional chemical analysis, commencing with the collection of soil samples. These samples undergo analysis for soil properties through traditional techniques following necessary processing. Simultaneously, spectral data for the same samples are acquired using various methods, under controlled conditions of constant temperature in a dark room after shade drying. This process results in the creation of a soil library containing both spectral and soil property data. Of the total samples collected, 70 percent are designated as the calibration set, while the remaining 30 percent forms the validation set. The calibration set data is employed to train models, and the validation set is instrumental in evaluating the accuracy and predictive performance of the model (Barra *et al.*, 2021).

Collection of soil samples

A necessary quantity of soil samples is collected and subjected to processing for both chemical analysis and the acquisition of spectral data.

Soil chemical analysis

The processed soil samples are analysed for various soil properties by routine chemical analytical techniques.

Spectral data acquisition

The spectral data of the soil is recorded through various techniques, predominantly spectroscopic methods operating across different ranges of electromagnetic radiation.

Infrared (IR) spectroscopy

IR spectroscopy is a method that examines molecular vibrations, relying on the principle of simple harmonic oscillation. For instance, the shared bonds between two atoms form a system of simple harmonic oscillators. These bonds exhibit resonant characteristics determined by the "spring" constant describing the force between them and the atomic weight at each bond's end. Consequently, the oscillation of a specific chemical bond aligns with a distinct frequency and energy intensity. The frequencies of oscillation between bonds correspond to particular wavenumbers (cm^{-1}), exhibiting a positive correlation. The incident radiation is directed onto a sample at similar vibrational frequencies shared between the bonds. Some of this radiation becomes absorbed, while others are reflected. A spectrophotometer records the reflectance, creating a reflectance spectrum that displays the energy magnitude captured as a function of wavelength. Identification of a molecule is possible by comparing its absorption peak to a database of spectra (Ferrari *et al.*, 2004; Burton *et al.*, 2020). Compared to conventional laboratory methods, visible–near-infrared reflectance (vis–NIR) spectroscopy is a more practical and cost-effective approach for estimating soil physical and chemical properties (Liu *et al.*, 2020).

The diffuse reflectance spectra of soil samples in the visible near infrared (VNIR) and the mid-infrared (MIR) wavelength ranges and PLSR was used to develop prediction models for soil organic carbon. On comparison using the coefficient of determination (R^2), ratio of performance to deviation (RPD) and the root mean square error (RMSE), VNIR spectra offer better accuracy with an $R^2 = 0.90$, RPD = 3.12 and RMSE = 0.07 compared to the MIR spectra with an $R^2 = 0.85$, RPD = 3.09 and RMSE = 0.08 (Olatunde, 2021).

Fourier Transform Infrared (FTIR) spectroscopy

A technique for simultaneously measuring all infrared frequencies, as opposed to individually, was developed using a straightforward optical device known as an interferometer. This device generates a distinctive signal incorporating all infrared frequencies.

Interferometers utilize a beam splitter that takes the incoming infrared beam and splits it into two optical beams. One beam reflects off a stationary flat mirror, while the other beam reflects off another mirror that typically moves a few millimeters away from the beam splitter. The two reflected beams are then recombined when they converge back at the beam splitter, resulting in a signal known as an interferogram. This allows for the simultaneous measurement of all frequencies. To obtain the frequency spectrum necessary for sample identification, the interferogram is decoded using a mathematical technique called Fourier transformation, which is performed by computers (Ismail *et al.*, 1997).

The Fourier Transform (FT) is a mathematical operation that converts a function into a representation describing the frequencies present in the original function. The output of this transform is a complex-valued function of frequency. The term "Fourier transform" encompasses both this complex-valued function and the mathematical operation itself (Sawant *et al.*, 2011).

Diffuse Reflectance Infrared Fourier Transform (DRIFT) spectroscopy – PLSR calibration model was found acceptable with coefficient of determination (R^2) of 0.49 and RPD of 1.4 (Filep *et al.*, 2016).

Attenuated total reflectance (ATR) spectroscopy

ATR spectroscopy is based on the same principles of infrared energy absorption by particles and molecular bonds. However, rather than exposing samples to infrared spectrums and capturing diffuse reflectance, this technique

Table 1: Soil properties predicted using Vis-NIR and chemometrics.

Technique	Predicted soil properties	References
Vis-NIR	P, K, Ca, Mg, Al, soil organic carbon (SOC), CEC	Marmette *et al.*, (2018)
	Clay, sand, silt, $CaCO_3$, CEC, SOC, pH	Gomez *et al.*, (2012)
	Bulk density	Katuwal *et al.*, (2019)
	Soil organic nitrogen	Sithole *et al.*, (2018)
	Hot water extractable carbon	Vohland and Emmerling (2011)
	TC	McDowell *et al.*, (2012)
	pH, CEC, Sand, Clay, Silt, TC, TN, K, P, S, Fe, Cu, Mn, Zn	Clingensmith *et al.*, (2019)
	SOC, pH	Vohland *et al.*, (2014)
	CEC, SOC, pH, exch. Ca	Araujo *et al.*, (2015)
	TC, SOC	Knox *et al.*, (2015)
NIR	Total carbon (TC) and total nitrogen (TN)	Barthes *et al.*, (2019)
	Mineralizable N	Russel *et al.*, (2002)
	Clay, silt, sand, pH, TC, P, Ca, Mg, K, Al, CEC	Vendrame *et al.*, (2012)
	Exchangeable Na	Cozzolino (2010)
	SOC, Nitrate	Freschet *et al.*, (2011)
	K, S	Cozzolino *et al.*, (2013)
	SOC, pH, As, Cu, Zn, Pb, Cr	Dong *et al.*, (2011)
	SOC, TN	Xie *et al.*, (2011)
	Avail. N, P, K	Shao and He (2011)
	TC	Arachchi *et al.*, (2016)
	SOC, TC	Sharififar *et al.*, (2019)
MIR	pH, CEC, SOC, Ca, Mg, TN, TP, Fe, Cu, K, Na	Ji *et al.*, (2016)
	SOC, CEC, pH, TN, Sand, Silt, Clay	Sanderman *et al.*, (2020)
	SOC, TN, Sand, Silt, Clay	Ludwig *et al.*, (2019)
	pH, SOC, TC, TN, clay, silt, sand, P, K, CEC, S	Wijewardane *et al.*, (2018)
	TC, SOC, TN, pH, sand, clay, silt	Sepahvand *et al.*, (2019)
LIBS	TC, SOC	Bricklemyer *et al.*, (2018)
	SOC	Xu *et al.*, (2019)
EDXRF	K, Fe	Buchele *et al.*, (2019)
FTMIR-ATR	pH (H_2O, KCl), $CaCO_3$, SOC, CEC, Sand, Silt, Clay	Dunne *et al.*, (2020)

involves directing incident IR energy directly onto a crystal in direct contact with the sample. Typically composed of materials like diamond, zinc selenide, germanium, or thallium iodide, the selection of an appropriate ATR crystal requires consideration of factors such as refractive index, spectral range, and the physical and chemical properties of the sample. When incident radiation strikes the crystal, an evanescent field forms between the crystal and the sample due to the reflected incident energy. Subsequently, the energy exits the ATR crystal and travels towards a spectrometer, generating a spectrum for the sample. ATR in the 2,500–50,000 nm spectrum was used to distinguish nitrate in various soil pastes. Although ATR minimizes the need for extensive sample preparation in soil nutrient detection, challenges persist in mitigating interferences from soil moisture and existing minerals (Jahn *et al.*, 2006). Additionally, ATR instrumentation is costly and fragile, making it impractical for field use (Alkhuder, 2022; Burton *et al.*, 2020).

Raman spectroscopy

Raman spectroscopy is a technique that involves observing changes in the wavelength and intensity of scattered light as it interacts with a sample. The incident light energy is absorbed and then re-emitted from the sample at distinct frequencies, known as the Raman scatter or Raman effect. The resulting Raman spectrum serves as a unique identifier for a sample, providing valuable information about its chemical structure, identity, polymorphism, intrinsic strain, and potential contamination (Burton *et al.*, 2020).

Laser Induced Breakdown Spectroscopy (LIBS)

A high-energy laser pulse serves as the energy source, inducing the ablation of atoms from the sample surface and leading to the formation of a transient, high-temperature plasma. Plasma temperatures generally exceed 10,000 K, providing sufficient energy to excite electrons in outer orbitals. During the cooling phase of

the plasma, the excited electrons transition to lower-energy orbitals, releasing photons with wavelengths inversely proportional to the energy difference between the excited and base orbitals. Due to the numerous possible excited states, each element exhibits multiple emitted wavelengths (Senesi and Senesi, 2016). Various types of lasers can be employed for Laser-Induced Breakdown Spectroscopy (LIBS), with the crucial factor for successful readings being the laser's spot size on the material, typically on the order of micrometers.

Immediately following the laser pulse, the light received by the spectrometer appears white, given the small and extremely hot nature of the plasma bubble. The plasma conditions break down the majority of chemical bonds, leaving the constituent elements as free atoms, often in an ionized state. As the plasma expands and cools, ions and atoms relax from their excited states, emitting the useful spectral lines (Musazzi and Pereni, 2014).

The accuracy of single pulse and double pulse LIBS with PLSR modelling in predicting various soil properties was compared and it showed better accuracy and reliability with double pulse than those of single-pulse signals when measured in terms of R^2 and RMSE. More than 93% prediction accuracy was observed with models trained using double pulse LIBS (He *et al.*, 2018).

Energy Dispersed X-Ray Fluorescence (EDXRF)

X-ray fluorescence (XRF) spectrometry is a contemporary approach to elemental analysis, extensively utilized in various disciplines such as forensic science, archaeology, geochemistry, and other related fields. Direct excitation involves exciting atoms in a specimen using primary photons from external sources like an X-ray tube, radioactive source, or synchrotron beam to generate primary fluorescence. On the other hand, indirect excitation results in observed fluorescence as a secondary process caused by photons or particles (electrons) originating from direct excitation or other secondary processes within the specimen. X-rays, electromagnetic radiation generated by high-energy particles bombarding atoms, exhibit wave-particle duality. X-ray fluorescence (XRF) spectrometry utilizes primary X-ray photons or microscopic particles to excite atoms in the tested material, generating secondary XRF for material composition analysis and chemical state research (Marguí *et al.*, 2022).

In EDXRF spectrometers, all elements in the sample are simultaneously excited, and an energy dispersive detector, combined with a multi-channel analyzer, is employed to collect fluorescence radiation emitted from the sample. This process separates different energies of characteristic radiation from each element in the sample. The resolution of EDXRF systems depends on the detector and typically ranges from 150 eV to 600 eV. Key advantages of EDXRF systems include simplicity, rapid operation, the absence of moving parts, and high source efficiency (Chen *et al.*, 2008).

A multivariate PLSR model trained and validated for total K and Fe in soil showed a higher R^2 for both K and Fe (0.98 and 0.90 respectively) and lower error value (RMSE = 0.17), which indicated that EDXRF can be effectively used along with PLSR modelling for accurate prediction of soil properties (Buchele *et al.*, 2019).

Data pre-processing

Many chemometric methods primarily focus on the diversity and extent of variation within the data. This preference is based on a simple principle: variables with the same values across all samples don't provide substantial learning value. However, large datasets also include variation unrelated to the specific problem under investigation. It is ideal to remove this irrelevant variation from the data to avoid interference with data analysis methods. This process is commonly known as data preprocessing (Gerretzon, 2023). The spectral data acquired may exhibit various artifacts such as outliers, missing data, noise, baseline effects, light scattering, temporal and spatial misalignment, multiplicative effects, and peak shifts. To eliminate or diminish these undesired artifacts from the spectra, various pre-processing techniques are employed. Data pre-processing involves transforming raw data into a format that can be comprehended and analyzed by computers (Mishra *et al.*, 2021).

Noise

Noisy data refers to data containing a substantial amount of additional meaningless information known as noise. This encompasses data corruption, often used synonymously with corrupt data, and any data that a user system cannot understand and interpret accurately. Unstructured text, for instance, may be challenging for many systems to utilize. Improper handling of noisy data can adversely impact the outcomes of data analysis and distort conclusions. Statistical analysis is sometimes employed to filter noise from noisy data (Gupta and Gupta, 2019).

Outliers

An outlier is an observation that significantly deviates from the other values within a random sample taken from a population. The definition of what is considered abnormal is typically determined by the analyst or through a

Table 2: Various pre-processing techniques.

Method	Objective	Reference
Smoothing	Removes high frequency noise from the sample	Mishra *et al.*, (2020)
Wavelet transform		
Derivatives	Reduce the drift of baseline	
Normalization	Minimize the errors presented due to sample preparation	
Standard normal variate (SNV)	Eliminate the effect of uncontrolled variations	
Multiplicative scatter corrections (MSC)	Mitigate problems arising from scattered light	
Savitzky-Golay derivative	Reduce high frequency noise in a signal	

consensus process. It is essential to first characterize normal observations before identifying abnormal ones (Zhang, 2013).

Missing data

Missing data, also known as missing values, pertains to data that is not recorded for a variable in the observed data set. The challenge of dealing with missing data is pervasive across various research studies and can have a substantial impact on the conclusions drawn from the data analysis (Rawal *et al.*, 2017).

Baseline effect

The baseline effect denotes the number of counts generated by the detector when there is no light. As various sources of noise are invariably present, achieving zero counts is impossible. The primary components of baseline offset include electronic offset, dark current, and readout noise. The primary pre-processing technique of utmost importance is the Savitzky-Golay derivation, which serves to diminish high-frequency noise in a signal. In this method, each variable within a sample undergoes subtraction from its immediate neighbouring variable to yield a first derivative. This process is iteratively repeated to obtain a second derivative, resulting in a more precise spectrum (Mishra *et al.*, 2020).

Model calibration

Subsequently, the pre-processed data is subjected to analysis by machine learning tools, facilitating the training of models to predict soil properties from spectra that have not been previously encountered.

Machine learning

Machine learning, a subset of artificial intelligence, entails endowing machines with the capability to emulate intelligent human behaviour. In the context of soil diagnosis, machine learning serves as a replacement for a chemical analyst, predicting soil properties (Carleo *et al.*, 2019).

Types of machine learning

Machine learning can be categorized into supervised, unsupervised, and reinforced learning. Supervised learning necessitates labeled data for training, involving instructional-based processes that require supervision. Unsupervised learning identifies concealed data patterns from unlabeled datasets, while reinforcement learning does not rely on predefined data; instead, it learns through interaction with the environment. The primary goal of reinforcement learning is for the agent to acquire an optimal or nearly-optimal policy that maximizes the "reward function" or another user-provided reinforcement signal accumulated from immediate rewards (Ayodele, 2010).

Statistical tools used in chemometric modeling

Partial Least Square Regression (PLSR)

Partial Least Square (PLS) analysis stands out as a preferred tool in chemometrics for constructing calibration models. Introduced by Herman Wold, the PLS technique not only captures the maximum variation associated with predictor variables (i.e., spectra) and predicted variables (i.e., concentration) but also maximizes the correlation between them. This technique aids in regressing predictor variables (x, the spectra) against predicted variables (y, soil properties/concentration). An advantageous aspect of the PLSR algorithm is its equal emphasis on both predictors and predicted variables (Kumar, 2021).

The soil reflectance decreases with increasing organic matter content. Knox *et al.*, (2015) observed that the wavelengths centered at approximately 400, 450, 510, 550, 700, 870, and 1000 nm indicate the presence of ferrous and ferric iron oxides, attributed to electronic transitions of iron cations in addition to soil components.

Physical soil properties, such as particle size distribution and aggregate size and density, also influence the reflectance intensity and shape of soil spectra through the phenomena of light scattering and reflection (Bellon-Maurel and McBratney, 2011; Conforti *et al.*, 2015).

The soil reflectance exhibited relatively high values for loamy sand soils due to the abundance of quartz in the sand fraction. However, reflectance decreased with an increase in clay content dominated by phyllosilicates, resulting in an increase in soil organic carbon (SOC) concentration (Conforti *et al.*, 2015).

The prediction models for micronutrients (Zn, Mn, Cu and Fe) and heavy metals (Cd, Pb, Ni and Cr) were developed using PLSR and showed the highest prediction accuracy was observed for Cu (94%), while the lowest was for Cd (26%). The model exhibited excellent predictions for Cu, Pb, Mn, Zn, and Ni, acceptable predictions for Fe and Cr, but poor prediction for Cd. The accuracy of the model's prediction was quantified using the R^2 value and followed the order Cu > Pb > Mn = Zn > Ni > Fe > Cr > Cd (Krzebietke *et al.*, 2023).

The PLSR models developed for the prediction of heavy metals in soils of orthic-anthrosols showed a prediction accuracy in the order Hg > Cr > Ni > Pb > Cu > As > Cd (Liu *et al.*, 2020).

The usefulness of PLSR with either PLSR combined with a genetic algorithm (GA-PLSR) or support vector machine regression (SVMR) was compared for an estimation of soil organic carbon (SOC), total nitrogen (N), pH, cation exchange capacity (CEC) and soil texture for surface soils. The order of estimation accuracies for the random validation sample was SVMR > GA-PLSR > PLSR for SOC, N, pH, and CEC, whereas for clay the order changed to SVMR > PLSR > GA-PLSR (Ludwig *et al.*, 2018).

The model constructed based on the PLSR machine learning method and vis-NIR (350-2500 nm) spectral data has a good predictive power (R^2 > 0.9, ratio of performance to deviation (RPD) > 3.0). For physical and chemical properties, the bulk density (BD, R^2 = 0.97, RPD = 5.90), soil organic matter (SOM, R^2 = 0.98, RPD = 8.56), pH (R^2 = 0.95, RPD = 4.40), and TN (R^2 = 0.98, RPD = 6.67) were predicted (Liu *et al.*, 2020).

Cubist Regression Model (CRM)

Cubist is a rule-based model extending Quinlan's M5 model tree. The tree grows, with terminal leaves containing linear regression models based on predictors used in prior splits. Intermediate linear models exist at each step, and predictions are made using the linear regression model at the terminal node. This prediction is "smoothed" by considering the linear model's prediction in the previous node, occurring recursively up the tree. The tree is eventually transformed into a set of rules through pruning and/or combination for simplification (Kuhn *et al.*, 2012).

Random Forest (RF)

Random Forest Regression, a supervised learning algorithm, employs an ensemble learning method for regression. Ensemble learning combines predictions from multiple machine learning algorithms to enhance accuracy compared to a single model (Liu *et al.*, 2012).

Support Vector Machine Regression (SVMR)

Support Vector Machines, supervised learning models with associated learning algorithms, are utilized for data analysis in classification and regression. In Support Vector Regression, the hyperplane is employed to fit the data and predict discrete values (Brereton and Lloyd, 2010).

The prediction accuracy of three chemometric regression techniques (PLSR, RF and SVMR) was compared to identify the most suitable model for predicting various soil properties. Among the different models, PLSR-based predictive models outperformed the other two regression techniques for all soil properties, except for EC (Hati *et al.*, 2022).

Multiple Linear Regression (MLR)

Multiple Linear Regression is a statistical technique utilizing several explanatory variables to predict the outcome of a response variable. The primary objective of multiple linear regression is to model the linear relationship between explanatory (independent) variables and response (dependent) variables (Pentos *et al.*, 2022).

Validation of the model

The validation of the trained model is conducted to assess prediction accuracy, employing four parameters (Coblinski *et al.*, 2020).

Coefficient of determination

The coefficient of determination, denoted as R^2, is a statistical measure evaluating a model's capability to predict or explain outcomes in a linear regression context. Specifically, R^2 indicates the proportion of variance in the dependent variable (Y) predicted or explained by linear regression and the predictor variable (X, also known as the independent variable). A higher R^2 value suggests a better fit for the model to the data. Interpretation of fit varies contextually, with an R^2 of 0.35 signifying 35 percent of the variation in the outcome being explained by the model. Excellent prediction is indicated if R^2 is greater than 0.80, acceptable if in the range of 0.60-0.80, and poor if less than 0.60 (Cozzolino and Moron, 2006).

Root mean square error (RMSE)

The root-mean-square error (RMSE) serves as a commonly used measure for assessing differences between values predicted by a model and observed values. Representing the square root of the second sample moment of the differences, RMSE aggregates the magnitudes of errors in predictions across various data points into a single measure of predictive power. Lower RMSE values are generally preferred, as they indicate higher accuracy. RMSE is scale-dependent and compares forecasting errors within a specific dataset rather than

between datasets (Chai and Draxler, 2014).

Ratio of performance to inter-quartile (RPIQ)

It is defined as the interquartile range of observed values divided by the Root Mean Square Error (RMSE), RPIQ considers both prediction error and the variation of observed values. This metric provides an objective measure of model validity, facilitating comparison across different model validation studies. A higher RPIQ signifies better predictive capacity, with values above 2.0 indicating excellent prediction accuracy, 1.4-2.0 considered reasonable, and values below 1.4 indicating poor prediction accuracy (Nawar and Mouazen, 2017).

Ratio of Performance to Deviation (RPD)

RPD is the ratio between the standard deviation of a variable and the standard error of prediction of that variable by a given model.

Advantages of chemometrics in soil analysis

- Rapid
- Environment friendly
- Cost-effective once established
- Multi analytical technique
- Non destructive
- Requires minimum or no sample preparation

Disadvantages of chemometrics in soil analysis

- Dynamic nature of soil
- Lack of technology and instrumentation
- High initial establishment and maintenance cost
- Chances of field level variation high (due to changes in temperature, sunlight and moisture)
- Difficulty in model transfer

Conclusion

Chemometrics is an effective alternative to routine analytical techniques which are time consuming, costly and which uses harmful chemical reagents. However, further research is required to standardize the working methods and to make its efficient use in predicting soil biological properties also, so that it will be an effective tool for soil health assessment.

Acknowledgement

The authors express their gratitude to the Department of Soil Science and Agricultural Chemistry, College of Agriculture, Vellanikkara, Kerala Agricultural University, India, for providing research facilities and technical support.

Conflict of interest: The authors declare that they have no conflicts of interest.

References

Adams, M.J. (1990). *Chemometrics in Analytical Spectroscopy*. The Royal Society of Chemistry, Cambridge.

Alkhuder, K. (2022). Attenuated total reflection-Fourier transform infrared spectroscopy: A universal analytical technique with promising applications in forensic analyses. *Int. J. Legal Med.*, **136(6)**, 1717-1736.

Arachchi, M.P.N.K.H., Field D.J. and McBratney A.B. (2016). Quantification of soil carbon from bulk soil samples to predict the aggregate-carbon fractions within using near- and mid-infrared spectroscopic techniques. *Geoderma*, **267**, 207-214.

Araujo, S.R., Soderstrom M., Eriksson J., Isendahl C., Stenborg P. and Dematte J.A.M. (2015). Determining soil properties in Amazonian Dark Earths by reflectance spectroscopy. *Geoderma*, **237**, 308-317.

Ayodele, T.O. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*, **3**, 19-48.

Barra, I., Haefele S.M., Sakrabani R. and Kebede F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances – A review. *Trends Anal. Chem.*, **135**, 116166.

Barthes, B.G., Kouakoua E., Clairotte M., Lallemand J., Chapuis-Lard L., Rabenarivo M. and Roussel S. (2019). Performance comparison between a miniaturized and a conventional near infrared reflectance (NIR) spectrometer for characterising soil carbon and nitrogen. *Geoderma*, **338**, 422-429.

Bellon-Maurel, V. and McBratney A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils-Critical review and research perspectives. *Soil Biol. Biochem.*, **43**, 1398-1410.

Bellon-Maurel, V., Fernandez-Ahumada E., Palagos B., Roger J.M. and McBratney A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Anal. Chem.*, **29(9)**, 1073-1081.

Brereton, R.G. and Lloyd G.R. (2010). Support vector machines for classification and regression. *Analyst*, **135**, 230-267.

Brereton, R.G. (2003). Experimental design. In: *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. University of Bristol, UK, 497.

Bricklemyer, R.S., Brown D.J., Turk P.J. and Clegg S. (2018). Comparing vis-NIRS, LIBS, and combined vis-NIRS-LIBS for intact soil core soil carbon measurement. *Soil Sci. Soc. Am. J.*, **82**, 1482-1496.

Buchele, D., Chao M., Ostermann M., Leenen M. and Bald I. (2019). Multivariate chemometrics as a tool for prediction of K and Fe in a diverse German agricultural soil-set using EDXRF. *Sci. Rep.*, **9**, 17588.

Burton, L., Jayachandran K. and Bhansali K. (2020). Review – The "real time" revolution for in situ soil nutrient sensing.

J. Electrochem. Soc., **167**, 037569.

Bystrzanowska, M. and Tobiszewski M. (2020). Chemometrics for selection, prediction, and classification of sustainable solutions for green chemistry—A review. *Symmetry*, **12(12)**, 2055.

Carleo, G., Cirac I., Cranmer K., Daudet L., Schuld M., Tishby N., Vogt-Maranto L. and Zdeborova L. (2019). Machine learning and the physical sciences. *Rev. Mod. Phys.*, **91**, 1-39.

Cecillon, L., Barthes B.G., Gomez C., Ertlen D., Génot V., Hedde M., Stevens A. and Brun J.J. (2009). Assessment and monitoring of soil quality using near infrared reflectance spectroscopy (NIRS). *Eur. J. Soil Sci.*, **60(5)**, 770-784.

Chai, T. and Draxler R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geosci. Model Dev. Discuss.*, **7**, 1525-1534.

Chen, Z.W., Gibson W.M. and Huang H. (2008). High-definition X-ray fluorescence: principles and techniques. *X-Ray Opt. Instrum.*, **2008**, 1-10.

Clingensmith, C.M., Grunwald S. and Wani S.P. (2019). Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data limited environment. *Eur. J. Soil Sci.*, **70(1)**, 107-126.

Coblinski, J.A., Giasson E., Dematte J.A.M., Dotto A.C., Costa J.J.F. and Vasat R. (2020). Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. *Catena*, **189**, 104485.

Conforti, M., Castrignano A., Robustelli G., Scarciglia F., Stelluti M. and Buttafuoco G. (2015). Laboratory-based Vis-NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *Catena*, **124**, 60-67.

Cozzolino, D. and Moron A. (2006). Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil Tillage Res.*, **85**, 78-85.

Cozzolino, D., Cynkar W.U., Dambergs R.G. and Shah N. (2013). In situ measurement of soil chemical composition by near-infrared spectroscopy: a tool toward sustainable vineyard management. *Commun. Soil Sci. Plant Anal.*, **44(10)**, 37-41.

Cozzolino, D. (2010). Influence of soil particle size on the measurement of sodium by near-infrared reflectance spectroscopy. *Commun. Soil Sci. Plant Anal.*, **41(19)**, 2330-2339.

Dong, Y.W., Yang S.Q., Xu C.Y., Li Y.Z., Bai W., Fan Z.N., Wang Y.N. and Li Q.Z. (2011). Determination of soil parameters in apple-growing regions by near- and mid-infrared spectroscopy. *Pedosphere*, **21**, 591-602.

Dunne, S., Holden N.M., O'Rourke S.M., Fenelon A. and Daly K. (2020). Prediction of phosphorus sorption indices and isotherm parameters in agricultural soils using mid-infrared spectroscopy. *Geoderma*, **358**, 113981.

Ferrari, M., Mottola L. and Quaresima V. (2004). Principles, techniques, and limitations of near infrared spectroscopy. *Can. J. Appl. Physiol.*, **29(4)**, 463-487.

Filep, T., Zachary D. and Balog K. (2016). Assessment of soil quality of arable soils in Hungary using DRIFT spectroscopy and chemometrics. *Vibrational Spectrosc.*, **84**, 16-23.

Freschet, G.T., Barthes B.G., Brunet D., Hien E. and Masse D. (2011). Use of near infrared reflectance spectroscopy (NIRS) for predicting soil fertility and historical management. *Commun. Soil Sci. Plant Anal.*, **42(14)**, 1692-1705.

Gerretzon, J. (2023). Data artifacts and preprocessing. In: *Chemical Data Preprocessing*. Radboud University, Nijmegen.

Gomez, C., Lagacherie P. and Coulouma G. (2012). Regional predictions of eight common soil properties and their spatial structures from hyperspectral Vis-NIR data. *Geoderma*, **189**, 176-185.

Gupta, S. and Gupta A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Comput. Sci.*, **161**, 466-474.

Hati, K.M., Sinha N.K., Mohanty M., Jha P., Londhe S., Sila A., Towett E., Chaudhary R.S., Jayaraman S., Coumar M.V. *et al.* (2022). Mid-infrared reflectance spectroscopy for estimation of soil properties of Alfisols from Eastern India. *Sustainability*, **14**, 4883.

He, Y., Liu X., Lv Y., Liu X., Peng J., Shen T., Zhao Y., Tang Y. and Luo S. (2018). Quantitative analysis of nutrient elements in soil using single- and double-pulse Laser Induced Breakdown Spectroscopy. *Sensors*, **18**, 1526.

Ismail, A.A., van de Voort F.R. and Sedman J. (1997). Fourier transform infrared spectroscopy: principles and applications. In: *Techniques and Instrumentation in Analytical Chemistry*, **18**, 93-139.

Jahn, B.R., Linker R., Upadhyaya S.K., Shaviv A., Slaughter D.C. and Shmulevich I. (2006). Mid-infrared spectroscopic determination of soil nitrate content. *Biosyst. Eng.*, **94(4)**, 505-515.

Ji, W., Adamchuk V.I., Biswas A., Dhawale N.M., Sudarsan B., Zhang Y., Rossel R.A.V. and Shi Z. (2016). Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosyst. Eng.*, **152**, 14-27.

Katuwal, S., Knadel M., Norgaard T., Moldrup P., Greve M.H. and de Jonge L.W. (2019). Predicting the dry bulk density of soils across Denmark: comparison of single-parameter, multi-parameter, and Vis-NIR based models. *Geoderma*, **361**, 114080.

Knox, N.M., Grunwald S., McDowell M.L., Bruland G.L., Myers D.B. and Harris W.G. (2015). Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*, **239**, 229-239.

Krzebietke, S., Daszykowski M., Czarnik-Matusewicz H., Stanimirova I., Pieszczek L., Sienkiewicz S. and Wierzbowska J. (2023). Monitoring the concentrations

of Cd, Cu, Pb, Ni, Cr, Zn, Mn and Fe in cultivated Haplic Luvisol soils using near-infrared reflectance spectroscopy and chemometrics. *Talanta*, **251**, 123749.

Kuhn, M., Weston S., Keefer C. and Coulter N. (2012). Cubist models for regression. *R package Vignette R package version 0.0*, **18**, 480.

Kumar, K. (2021). Partial last square (PLS) analysis. *Resonance*, **26(3)**, 429-442.

Lehmann, J., Bossio D.A., Kogel-Knabner I. and Rillig M.C. (2020). The concept and future prospects of soil health. *Nat. Rev. Earth Environ.*, **1(10)**, 544-553.

Liu, J., Han J., Xie J., Wang H., Tong W. and Ba Y. (2020). Assessing heavy metal concentrations in earth-cumulic-orthic-anthrosols soils using Vis-NIR spectroscopy transform coupled with chemometrics. *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, **226**, 117639.

Liu, Y., Wang Y. and Zhang J. (2012). New machine learning algorithm: random forest. In: *Information Computing and Applications: Third International Conference, ICICA*. Springer Berlin Heidelberg, 246-252.

Ludwig, B., Murugan R., Parama V.R. and Vohland M. (2018). Use of different chemometric approaches for an estimation of soil properties at field scale with near infrared spectroscopy. *J. Plant Nutr. Soil Sci.*, **181(5)**, 704-713.

Ludwig, B., Murugan R., Ramakrishna P.V.R. and Vohland M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: implications of different chemometric approaches and software packages related to calibration sample size. *Soil Sci. Soc. Am. J.*, **83**, 1542-1552.

Marguí, E., Queralt I. and de Almeida E. (2022). X-ray fluorescence spectrometry for environmental analysis: Basic principles, instrumentation, applications and recent trends. *Chemosphere*, **303**, 135006.

Marmette, M., Adamchuk V., Nault J., Tabatabai S. and Cocciardi R. (2018). Comparison of the performance of two Vis-NIR spectrometers in the prediction of various soil properties. *Int. Soc. Precision Agric.*, **2018**, 1-12.

McDowell, M.L., Bruland G.L., Deenik J.L., Grunwald S. and Knox N.M. (2012). Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma*, **189**, 312-320.

Mishra, P., Biancolillo A., Roger J.M., Marini F. and Rutledge D.N. (2020). New data pre-processing trends based on ensemble of multiple pre-processing techniques. *Trends Anal. Chem.*, **132**, 116045.

Mishra, P., Rutledge D.N., Roger J.M., Wali K. and Khan H.A. (2021). Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta*, **229**, 122303.

Musazzi, S. and Pereni U. (2014). LIBS instrumental techniques. In: Musazzi, S. and Pereni, U. (Eds.), *Laser Induced Breakdown Spectroscopy*. Springer Heidelberg New York Dordrecht London, 575.

Nawar, S. and Mouazen A.M. (2017). Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena*, **151**, 118-129.

Olatunde, K.A. (2021). Estimation of soil organic carbon using chemometrics: a comparison between mid-infrared and visible near infrared diffuse reflectance spectroscopy. *West Afr. J. Appl. Ecol.*, **29(2)**, 1-11.

Patra, A.K., Lenka N.K. and Biswas A.K. (2015). Soil health assessment and management. *Indian J. Fertil.*, **11**, 16-27.

Pentos, K., Mbah J.T., Pieczarka K., Niedbala G. and Wojciechowski T. (2022). Evaluation of multiple linear regression and machine learning approaches to predict soil compaction and shear stress based on electrical parameters. *Appl. Sci.*, **12**, 8791.

Rawal, S., Gupta S.C. and Singh S. (2017). Predicting missing values in a dataset: challenges and approaches. *Int. J. Recent Res. Asp.*, **4(3)**, 34-38.

Rohman, A., Windarsih A., Hossain M.A.M., Johan M.R., Ali M.E. and Fadzilah N.A. (2019). Application of near-and mid-infrared spectroscopy combined with chemometrics for discrimination and authentication of herbal products: A review. *J. Appl. Pharma. Sci.*, **9(3)**, 137-147.

Rossel, R.V., Walvoort D.J.J., McBratney A.B., Janik L.J. and Skjemstad J.O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, **131(1-2)**, 59-75.

Russel, C.A., Angus J.F., Batten G.D., Dunn B.W. and Williams R.L. (2002). The potential of NIR spectroscopy to predict nitrogen mineralization in rice soils. *Plant Soil*, **247**, 243-252.

Sanderman, J., Savage K. and Dangal S.R.S. (2020). Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Sci. Soc. Am. J.*, **84**, 251-261.

Santos, M.C., Nascimento P.A.M., Guedes W.N., Pereira-Filho, E.R., Filletti E.R. and Pereira F.M.V. (2019). Chemometrics in analytical chemistry – An overview of applications from 2014 to 2018. *Eclectic Chem. J.*, **44(2)**, 11-25.

Sawant, S.D., Baravkar A.A. and Kale R.N. (2011). FTIR spectroscopy: principle, technique and mathematics. *Int. J. Pharma Bio Sci.*, **2(1)**, 513-519.

Senesi, G.S. and Senesi N. (2016). Laser Induced Breakdown Spectroscopy (LIBS) to measure quantitatively soil carbon with emphasis on soil organic carbon. A review. *Anal. Chem. Acta*, **938**, 7-17.

Sepahvand, H., Mirzaeitalarposhti R. and Beiranvand K. (2019). Prediction of soil carbon levels in calcareous soils of Iran by mid-infrared reflectance spectroscopy. *Environ. Pollut. Bioavailab.*, **31**, 9-17.

Shao, Y. and He Y. (2011). Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Res.*, **49**, 166-172.

Sharififar, A., Singh K., Jones E., Ginting F.I. and Minasny B. (2019). Evaluating a low cost portable NIR spectrometer for the prediction of soil organic and total carbon using different calibration models. *Soil Use Manag.*, **35**, 607-616.

Sithole, N.J., Ncama K. and Magwaza L.S. (2018). Robust Vis-NIR models for rapid assessment of soil organic carbon and nitrogen in Feralsols Haplic soils from different tillage management practices. *Comput. Electron. Agric.*, **153**, 295-301.

Vendrame, P.R.S., Marchao R.L., Brunet D. and Becquer T. (2012). The potential of NIR spectroscopy to predict soil texture and mineralogy in Cerrado Latosols. *Eur. J. Soil Sci.*, **63(5)**, 743-753.

Vohland, M. and Emmerling C. (2011). Determination of total soil organic C and hot water extractable C from VIS NIR soil reflectance with partial least squares regression and spectral feature selection techniques. *Eur. J. Soil Sci.*, **62(4)**, 598-606.

Vohland, M., Ludwig M., Thiele-Bruhn S. and Ludwig B. (2014). Determination of soil properties with visible to near- and mid-infrared spectroscopy: effects of spectral variable selection. *Geoderma*, **223**, 88-96.

Voncina, D.B. (2009). Chemometrics in analytical chemistry. *Appl. Nat. Sci.*, **211**, 211-216.

Wang, Y., Huang H.Y., Zuo Z.T. and Wang Y.Z. (2018). Comprehensive quality assessment of Dendrobium officinale using ATR-FTIR spectroscopy combined with random forest and support vector machine regression. *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, **205**, 637-648.

Wijewardane, N.K., Ge Y., Wills S. and Libohova Z. (2018). Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. *Soil Sci. Soc. Am. J.*, **82**, 722-731.

Wilkins, C.L. (1990). *Computer-Enhanced Analytical Spectroscopy*. The Royal Society of Chemistry, Cambridge.

Xie, H.T., Yang X.M., Drury C.F., Yang J.Y. and Zhang X.D. (2011). Predicting soil organic carbon and total nitrogen using mid- and near-infrared spectra for Brookston clay loam soil in Southwestern Ontario, Canada. *Can. J. Soil Sci.*, **91**, 53-63.

Xu, X., Du C., Ma F., Shen Y., Wu K., Liang D. and Zhou J. (2019). Detection of soil organic matter from laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (FTIR-ATR) coupled with multivariate techniques. *Geoderma*, **355**, 1-13.

Zhang, J. (2013). Advancements of outlier detection: A survey. *ICST Trans. Scalable Inf. Syst.*, **13(1)**, 1-26.