MAPPING OF SOIL PROPERTIES USING MACHINE LEARNING TECHNIQUES

Abstract

We aimed to estimate Soil Nutrients and relate the spectral signatures to that of the Laboratory reference Measurements utilizing CART analysis. Sustainable agriculture aims at controlled and/or precise soil fertility interventions based on spatial soil information. The profound advancements in remote sensing and geospatial techniques provide means for determining the spatial coverage and variability of the soil properties through the survey and image data incorporated in the mapping procedures (i.e.) Digital Soil Mapping. The soil moisture content at varying levels influences crop growth and decides the yield, as the crop requires water at critical crop growth stages. Machine learning techniques provide the means of optimized model calibration when compared to conventional geostatistical or statistical approaches.

Keywords: CART analysis, geostatistical technique, Machine learning techniques, soil properties mapping

INTRODUCTION

The demand for Quantitative and spatial information on soil properties among others, has increased the research on the soil properties. The major worldwide concern in the twenty-first century is the ever-increasing population, and food security must be considered in this context. By 2025, total food demand in developing countries will be increased by 150%. To meet the food demand of the increasing population, agricultural productivity must be scaled to the present average through sustainable agriculture.

Sustainable agriculture aims at controlled or precise soil fertility interventions based on spatial soil information. The spatial information on various soil properties can aid crop growers in critical decision-making and helps in implementing policies for increasing agricultural productivity and the livelihood of small-scale farmers. The conventional means of estimating the soil's physical and chemical properties through sampling and laboratory analyses are time-consuming and expensive when the mapping is done at the regional or national level. The profound advancements in remote sensing and geospatial techniques provide the means of determining the spatial remotely sensed data through its extractable soil spectral information, large spatial coverage, and temporal consistency helps in mapping remote and inaccessible locations (Forkuor et al., 2017).

This eliminates the drawbacks of the traditional method of assessment and provides a venture for non-destructive sampling procedures. Though the spatial variability (i.e.) non-uniformity associated with soil properties can be included through Digital soil mapping, the process is always constrained by the within-site variability. The variability is accounted several natural processes influenced by the factors such as climate, soil type, and land use. The within-site variability can be excluded by the use of variable rate technology which facilitates precision agriculture. Precision agriculture tends to the site-specific needs of the soil and the crop through remote sensing and geospatial techniques, DEM, and other climatic variables (John et al., 2020).

Similarly, the hazards of soil erosion and its associated land degradation can be predicted by assessing soil properties. Soil erosion is the removal of the top portion of the soil which leads to the associated nutrient leaching and land degradation. Soil erosion is coverage and variability of the soil properties through the survey and image data incorporated in the mapping procedures (i.e.) Digital Soil Mapping.

Some soil physical properties such as porosity, soil aggregate stability, permeability, texture, structure, and chemical properties such as soil organic matter and calcium carbonate equivalent were considered responsible (Alexakis et al., 2019).

The Heavy metal assessment in the soil provides an insight into the contaminants that retard the quality and the biological properties of the soil. Most of the soil spatial information is facilitated through the model calibration employing several of the remotes sensed image variables, spectral information, and climatic and environmental variables. The major limitation in the model calibration is the selection of the spectral variable or bands (Gomez et al., 2008).

The use of geostatistical framework was prevalent in the spatial prediction of the soil information, which is a linear combination of the environmental covariates and spatial autocorrelated residuals and the prediction at unobserved location estimated through interpolation technique. The geostatistical models are considered for its assumptions on spatial variations and the uncertainty associated with the prediction measures. Conversely, the geostatistical models have several limitations which affect the fitness and the prediction accuracy. The limitations were the stationarity of the residuals, increase in the parameter estimated, and the increased computational load due to increased sample size. As an alternative, Machine learning approaches are employed for their increased efficiency when compared to the geostatistical models. *"Machine learning techniques refer to a large class of non-linear data-driven algorithms employed primarily for data mining and pattern recognition purposes, and now frequently used for regression and classification tasks in all fields of science"*(Wadoux et al., 2020).

associated with climatic parameters such as wind speed and temperature,

Unlike geostatistical models, machine learning techniques are void of assumptions and can process a large number of parameters. Geostatistical and statistical are model-oriented and the predictive accuracy depends on the assumptions that make up the model whereas machine learning techniques are data-driven and the predictions are made from the predictive model calibrated using an error-minimization process making the calibration more accurate than conventional (McBratney et al., 2003).

Several of the machine learning techniques have been utilized in much of the literatures as a comparative analysis and each technique have been scrutinized for its efficiency over other. The physicochemical soil properties are reviewed for its efficient mapping techniques, intercorrelation among the properties and the machine learning methods adopted for each of the mapping methodologies are detailed in the following.

SOIL PROPERTIES

The physicochemical properties that characterize soil fertility and their intercorrelations over other soil properties are discussed. Important soil physicochemical properties and their processes are depicted in the Table 1.

Soil Physical Properties

The fine movement of the air, water, and nutrient uptake by the plants is determined by the soil physical properties which affect the germination and soil erosion processes. The germination capability of the seed is determined by the water holding capacity i.e., soil moisture considering other parameters at their optimums. Several of the physical properties contribute to the soil erosion process as specified. (Abd-Elmabod et al., 2017).

Table 1. Some of the important soil properties and its associated soil processes

Soil Property	Soil Processes
Soil structure	Aggregation, organic matter turnover, retention, and transportation of water and chemicals
Porosity	Plant available water capacity, soil crusting, aeration, water entry
Infiltration	Soil water availability and movement, leaching of nutrients, erosion
Bulk density	Soil structural conditions, compaction
Available water	Field capacity, permanent wilting point, water flow
pH	Soil acidification, salinization, soil structural stability, biological and chemical activity thresholds
Electrical conductivity	Plant and microbial activity thresholds, leaching of salts, soil structure decline, salinization
Plant available N, P, and K	Availability of nutrients for plant uptake, losses from the soil-plant system
Soil organic matter	Organic matter storage and quality, plant residue decomposition, the metabolic activity of soil organisms, mineralization–immobilization turnover, microbial activity, nutrient supply
Total soil C and N	C and N mass and balance, soil structure, nutrient supply.

Adapted from Jat, Mangi L., Clare M. Stirling, Hanuman S. Jat, Jagdish P. Tetarwal, Raj K. Jat, Rajbir Singh, Santiago Lopez-Ridaura, and Paresh B. Shirsath. "Soil processes and wheat cropping under emerging climate change scenarios in South Asia." Advances in

Soils are differentiated based on the soil particle size and classified into textural classes of sand, silt, clay, and loamy soils. Based on the relative portion of the classes, the soils are classified (Jat et al., 2018). Soil Texture drives crop management and production. The particle size is associated with a particular textural class as depicted in Table 2.

Through pedo-transfer functions, the property is closely associated with any of the textural classes and can be quantified (Schaap & Leij, 1998). The absorption peak of the high clay content particularly of those of smectitic mineralogy can shrink and result in the formation of large cracks and fissures. Thus, soil with high shrink-swell potential is difficult to manage when dry. Soil structure is another physical property that is determined by management practices, environmental factors, and other properties. It is an indicator of the soil and determines the porosity, infiltration erodibility, C accumulation, and other processes (Jat et al., 2018). The soil structure is a measure that refers to the ability of the particles to resist disruption when outside forces are applied. Since soil structure is strongly affected by the amount of organic matter, any changes in the soil organic content will affect the structure resulting in lower infiltration rates and increased run-off. Soil structure can be used as an influencing variable (Boruvka et al., 2002).

The soil moisture influences crop growth and decides the yield, as the crop requires water at critical growth stages. A saturated soil will have a soil moisture tension of about -0.001 bars or less which requires less energy for plant uptake. At field capacity, the moisture will be available between -0.05 and -0.33 bars, and on the other hand when the plant requires much energy

(15 bar) to extract the water the situation is called wilting point (Adab et al., 2020).

The air is necessary for crop growth and water storage is determined by the porosity and the pore size distribution. Soil porosity influences the aeration capacity and water-holding capacity. It determines root development and soil enzyme activities. The soil infiltration capacity is related to that of the soil structure and texture (Jat et al., 2018). The soil infiltration capacity determines erodibility and surface run-off but can significantly change over time, use, and management. The bulk density refers to the state of soil compaction, aeration, and infiltration. Bulk density is inversely proportional to that of the soil organic matter (Pittman & Hu, 2020).

Table 2. The particle size of the respective textural classes

Textural Class	**Particle Size (Diameter)**
Sand	2 to 0.2 mm
Slit	0.2 to 0.002 mm
Clay	< 0.002 mm

Source: Adapted from Abd-Elmabod, Sameh K., Antonio Jordán, Luuk Fleskens, Jonathan D. Phillips, Miriam Muñoz-Rojas, Martine van der Ploeg, María Anaya-Romero, Soad El-Ashry, and Diego de la Rosa. "Modeling agricultural suitability along soil transects under current conditions and improved scenario of soil factors." In Soil mapping and process modeling for sustainable land use management, pp. 193-219. Elsevier, 2017.

Soil Chemical Properties

Most of the chemical properties of the soil are influenced by the climatic parameters or drivers that affect the soil's organic matter, carbon, nutrient cycling, and crop productivity. The soil pH and EC are also likely to be affected by the climatic drivers. Soil pH affects a wide range of soil properties such as salinization and soil nutrient availability. Intensive rainfalls lead to the leaching of the soil nutrient properties resulting in acidification as dictated by the buffering pools.

The extent of leaching and evaporation will depend on the degree of rainfall and temperature. The electrical conductivity is indirectly associated with the soil's structural properties. Sorption capacity and the cation exchange capacity in response to the high rainfall content and low organic content leads to low cation exchange capacity and an increase in the leaching of the base ions. The Cation exchange capacity of the soil also influences some of the major cations such as Mg, Ca, and K, and the immobilization of the Al and Mn content (Jat et al., 2018).

Soil organic carbon is an important indicator and helps in improved soil structure formation. The effect of SOC on other properties and ecosystem functioning requires less precision. Climatic factors influence the quantity of SOC in the soil (i.e.) Humid and cool temperatures increase the SOD content and vice versa. Although major influential factors that affect the SOC are climatic drivers and environmental variables, other properties were used as a measure such as soil structure, porosity, aggregate stability, pore connectivity, and clay mineralogy that affect soil affect SOC storage. (Jat et al., 2018).

The SOM has a strong influence at the top soil level to 10 or 20 cm at the most. Though the effect and prevalence of SOM are limited, SOM is considered critical as most of the agricultural activities take place at the topsoil level. The increased SOM will have a negative impact on the soil reflectance spectra. B. W.

formation of the absorption peak at a particular

Murphy *et al.*, 2015). Total Nitrogen and Phosphorous in the soil are fixed through natural biochemical processes. But the recalcitrant of OM and phosphorous adsorption by Sesquioxide has a larger effect on the availability of the N and P in the soil (Wang et al., 2012). TN and TP, when compared to other soil nutrients have a considerable effect on soil productivity. The SOC cycle, clay, and silt content of the soil have a strong influence over the SOC cycle (Liu et al., 2013).

The Total Potassium in the soil favors several physiological functions such as stomatal opening and closure, translocation of sugars and starch, enzyme activation, respiration, and ATP production. The TP availability is influenced by the soil moisture content (Goldberg, 1989). The micronutrients such as Ca, Mg, Fe, and Al have a significant role in determining soil fertility and have a strong influence over the physical properties (Gao et al., 2019). Flocculated clay facilitate by the Calcium ions favors the soil's stabilization by promoting aggregation. Ca ions have more affinity for the exchange site resulting in aggregation (Norton, 2013). Decreased Mg ions affect plants by limiting the formation of chlorophyll, activation of the enzyme, and decreasing the quality and yield. The levels of Mg and Ca ions are affected by the increased soil acidity facilitating increased aluminum exchange sites. (Yan & Hou, 2018).

The effects of Al and Fe ions include stabilizing clay minerals by decreasing coagulation and having a significant effect on the soil's physical properties(Behnood, 2018). Increased Ca, Al, and Fe content in the soil leads to the absorption peak in the soil reflectance spectra (Goldberg, 1989). The effect of the chemical properties has a profound change in the spectra either through the overall decrease in the reflectance or through the wavelength. The estimation of the soil chemical properties for digital soil mapping is made through the model calibration of the significant bands. Some of the soil textural classes and soil mineral composition that greatly influences the spectral properties of the soils and forms a prominent absorption peak at wavelengths are depicted in Table 3.

Table 3. Absorption peaks associated with some of the soil properties

Soil Properties	Absorption Peak (Bandwidth)
Clay (kaolinite, montmorillonite and illite)	2200nm
Calcium Carbonate	2340nm
Hematite (Iron)	550, 630, and 860 nm
Goethite (Iron)	480, 650, and 920 nm
Liquid water and O-H bonds	1400 and 1900 nm

Source: Adapted fromGomez, Cécile, and Phillipe Lagacherie. "Mapping of primary soil properties using optical visible and near-infrared (Vis-NIR) remote sensing." In Land surface remote sensing in agriculture and forest, pp. 1-35. Elsevier, 2016.

MACHINE LEARNING TECHNIQUES

The requirement and the application of machine learning techniques have increased exponentially in the past decade. The increased availability of remotely sensed data and many open-source algorithms lead to the increased adoption of ML techniques. "*Machine learning techniques refer to a large class of non-linear data-driven algorithms employed primarily for data mining and pattern recognition purposes, and now frequently used for regression and classification tasks in all fields of science.*" The use of Machine learning techniques in calibrating the predictive models has been employed in the Digital Soil Mapping or Pedotransfer functions and the analysis of the

The most often used Machine learning

soil Vis-NIR Spectra. Machine learning techniques can be divided into shallow and Deep. (Wadoux et al., 2020)

Shallow learning may be referred to as learning methods that have been adopted before 2005 and “Deep learning is neural networks with multiple hidden layers. Compared with shallow learning-based applications, deep learning models require large amounts of training data. Furthermore, the structures of the network have a great impact on the performance of the deep learning models.” (Xu et al., 2021). The percentage of the literature that has adopted machine learning approaches recently is depicted in Figure 1.

approaches are depicted in Table 5 in the majority of the literature evaluated. Binary Trees (BT), Support Vector Machines (SVM), Nave Bayes (NB), Artificial Neural Networks (ANN), Cubist Regression (CB), Principal Component Regression (PCR), Partial Least Square Regression (PLSR), Least-Square SVM (LS- SVM), Extreme Learning Machines (ELM), Ordinary Least Square Estimation (OLSE), Ant Colony Optimization-interval Partial Least Squares (ACO-iPLS (CNN). (Trontelj ml & Chambers, 2021).The advantages and disadvantages of the major Shallow and deep learning techniques are depicted in Table 4.

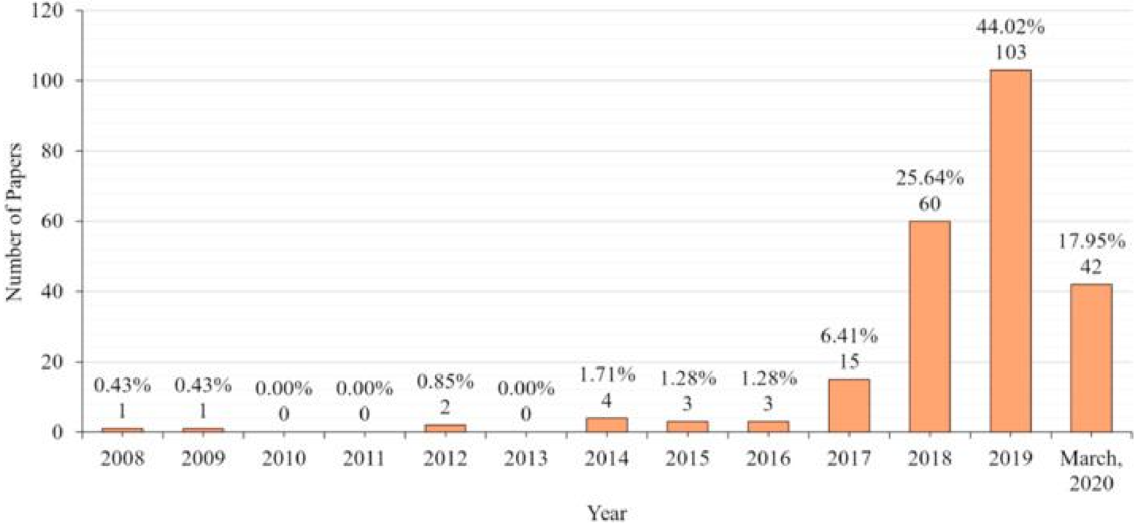

Figure 1. Percentage of literature adopted machine learning approaches over the years Source: Xu, Yayin, Ying Zhou, Przemyslaw Sekula, and Lieyun Ding. "Machine learning in construction: From shallow to deep learning." Developments in the Built Environment (2021): 100045.

Table 4. Advantages and disadvantages of major machine learning techniques implicated.

Method	**Advantages**	**Disadvantages**
Regression	Low computation time Performs well with large datasets Reduce data dimensionality Provide a feature selection Easy to implement	Do not deal with nonlinear problems over-fitting may occur

DT	Can be effectively applied to the nonlinear problem Performs well with large datasets In the built feature selection procedure Easy to implement	Over-fitting may occur Non-robust to small dataset changes Input parameters, such as nods numbers, need to be defined manually
SVM	Can be effectively applied to the nonlinear problem Performs well when data dimensionality is greater than the number of samples Low risk of the over-fitting	Non-robust to small dataset changes Is not suitable for large datasets, where data dimensionality is smaller than the number of samples An effective kernel function is not easy to define The large computational time for large datasets Different impact of the weights parameters that are not easy to visualize the impact Needs adaptation for multi-class problems Not easy to implement
NB	Can be effectively applied to the nonlinear problem Low computation time Suitable for multi-class problems Effective for small training datasets Easy to implement Robust to small dataset changes Probabilistic predictions can be obtained	Assigning zero probability to a categorical variable is not available and loss of accuracy
RF	Can be effectively applied to the nonlinear problem This may apply to soils under a great variety of environments Act to reduce bias Performs well with large datasets Overfitting is less common Accommodate random inputs and random features Can be used for classification as well as for regression	Had difficulty predicting high and low laboratory measured values, underestimating and overestimating them, respectively The number of trees needs to be defined manually Long computation time Large computational power is required due to the large number of trees created by the algorithm
NN	Can be effectively applied to the nonlinear problem Effective in many applications Defined fault tolerance that makes classification more robust Robust to small dataset changes Can perform in parallel without affecting the system	Effective architecture parameters need to be defined manually and classes are translated to respective numeric values Require large computation power and training dataset Weights are assigned randomly, to acquire high accuracy, the process of training the data must be iterative The duration of the network is unknown

Source: Chambers, Olga. "Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method." Sensors 21, no. 12 (2021): 4208. (NB – Naïve Bayes; RF- Random Forest; SVM – Support Vector Machine; NN- Neural Networks; DT – Decision Trees.

Shallow Learning

Shallow learning techniques require prior insight into the data that are to be learned. Different types of shallow learning techniques that have been developed are 1. Supervised, 2. Unsupervised, and 3. Reinforcement Learning. Supervised learning recognizes the pattern and optimizes the model based on the user-defined training datasets while Unsupervised learning recognizes the pattern purely based on the prediction i.e., patterns are determined based on the clustering techniques. Some of the supervised learning that was considered as a foundation were Logistic Regression, Perceptron, and kNN. While Perceptron undoubtedly laid the foundation for machine learning algorithms, they were fragmented and unstructured before the publication of the Decision Tree algorithm. A standardized algorithm concerning unsupervised learning includes Principal Component Analysis (PCA), kernel PCA and t-SNE. K-mean, EM, mean shift and spectral clustering are typical clustering algorithms (Xu et al., 2021). Reinforcement learning is typically avoided because of its trial-and-error method of model optimization resulting in computational load and fitting errors.

Deep Learning

Though most of the Deep learning methods adopted in the soil property mapping and estimations are less, model calibration through deep learning has evident significance when compared to the shallow methods.

Deep learning methods incorporate the neural networks and backpropagation is considered as the earlier artificial neural networks and insignificant due to their inability to train multilayer neural networks. Currently, two major network structures used in deep learning are CNNs and RNNs. CNNs and considered significant for digital image processing from than fully connected neural networks where each input feature is connected to each neutron in a fixed way. Two popular deep Recurrent Neural Networks (RNNs) types are LSTM and GRU (gated Recurrent Units). GRU is used when the datasets are smaller and LSTM is used for larger datasets (Xu et al., 2021).

MAJOR MACHINE LEARNING METHODS IMPLICATED

Support Vector Machine

Vapnik (1999) introduced the SVM principle. SVM is the ability to minimize the algorithm's structural error (i.e.) and ideally separate hyperplane for distinguishing classes that overlap but are not linearly separable. It was created for classification purposes, but it can also be used to solve regression problems. There are two types of SVM models: classification and regression (Elisseeff & Weston, 2001). To solve data categorization challenges, classification models are utilized. Forecasting difficulties are solved using regression models. The SVM has the benefit of being extremely effective in high-dimensional spaces (Zhang et al., 2020).

Random Forest

Breiman (2001) combined the bagging approach (Breiman, 1996) with the random 20-variable selection to create a random forest (RF). The aim was to combine a group of "weak learners" to build a "strong learner." For each RF tree, bootstrap sampling is employed, and the binary split data criteria for regression and classification tasks are distinct. (Zhang et al., 2020). The general premise of group training is that it improves the accuracy of other trained models, which is related to the principle that ensemble models are more accurate than solo models (Yousefi et al., 2021).

spectral feature detection and classification because of their less parameter estimation and entire image processing rather

The use of RFs has the advantage of allowing ensembles of trees to be employed

5. Machine Learning Techniques Used in the Soil Properties Mapping: Pieces of Literature Reviewed

Table 5. Some of the Machine Learning Techniques used in Soil Properties Mapping (Literature reviewed)

S.NO	Machine Learning Technique	Study area	Property estimated	Best ML feature selected	Validation measure used	Reference
1.	ELM, PLS, and BPNN	Morocco	SOC and TN	ELM	R^2, RMSE, and RPD	Reda et al. (2019)
2.	RF	Northeast China	STN	RF	R^2 and RMSE	Zhang et al. (2019)
3.	SVM, BRT, and RF	Switzerland	SOC and C: N ratio	BRT	R^2, RMSE, and MAE	Zhou et al. (2021)
4.	RF and SVM	Eastern Tunisian Atlas	soil texture	RF	OA	Bousbih et al. (2019)
5.	CNN - LucasCNN, LucasResNet, LucasCoordConv and RF	Europe	soil texture	LucasCoord Conv	OA, AA, Kappa	Riese and Keller (2019)
6.	ANN-Backpropagation	The northwestern province of Qazvin	STP		Fitted vs original plot	Keshavarzi et al. (2015)
7.	RF	Canada	Bulk Density and Soil Carbon		MAE and R2	Pittman and Hu (2020)
8.	PLSR, Cubist Regression, LS-SVM, and ELM	Middle-lower Yangtze Plain	SOM and pH	ELM	R^2, RMSE, and RPIQ	Yang et al. (2019)
9.	MLR, RFR, SVM, and SBG	South-western Burkina Faso	Sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC), and nitrogen	RFR	R^2, sMAPE, and RMSE	Forkuor et al. (2017)
10.	KNN, MLP, RF, SVM, XGB, ALR,	Northwest of China	soil texture and Soil Particle Size	RF and XGB	R^2, MAE, RMSE, AD and	Zhang et al. (2020)

	CLR, and ILR		Fractions		STRESS	
11.	MLR, RF, SVM, cubist regression, and ANN	Calabar, Cross River State	SP and SN	RF	RMSE, MAE, and R^2	John et al. (2020)
12.	MLR and RF	Morocco	Soil Aggregate	Both methods provided the same estimates	R^2 and RMSE	Bouslihim et al. (2021)
13.	boosted BRT, RF, and SVM	Northwestern China	STN	BRT and RF	R^2, MAE, and RMSE	Zhou et al. (2019)
14.	RF, EN, SVM, and ANN	Iran	Soil Moisture	RF	Nash–Sutcliffe efficiency value	Adab et al. (2020)
15.	PCR, PLSR, LS-SVM, and CB	Germany	Soil Moisture, Soil Total Nitrogen, and Soil Organic content	LS-SVM for OC, CB for N	RMSEP and RPD	Morellos et al. (2016)
16.	ANN, RF PLSR, and CB	Brazil	Environmental vulnerability	CB for OC, PLSR for N	-	Costa et al. (2020)
17.	PLSR, BPNN and GA-BPNN	Guangdong, China	Total Nitrogen, Total phosphorus, and total potassium	GA-BPNN for N, P, K	RRMSE	Liu et al. (2013)
18.	PLS and SVR	Anhui, China	Soil Available Potassium content	SVR for available K	R^2 AND RMSE	Jin et al. (2020)
19.	LS-SVM and PLSR	Zhejiang, China.	Total Nitrogen, Total phosphorus, and total potassium	LS-VM for N, P, and K	RMSEP	Shao and He (2011)
20.	OLSE, RF, and ELM	LUCAS Soil (23 countries)	Soil total Nitrogen	ELM for N	R^2, RMSEP, RPD	Wang et al. (2020)
21.	AOC-iPLS, RF, and RF-SVM	Xinjiang Uyghur Autonomous Region, China	Soil organic carbon (SOC)	RF-SVM for OC	R^2 and RMSE	Ding et al. (2018)
22.	PLSR, SVM, RF, ANN, and DL	Czech Republic	selected PTEs (Cr, Cu, Pb, Zn, and Al) in	ANN for Cr and Al	R^2 and RMSE	Gholizadeh et al. (2020)

			forest organic horizons				
23.	PCR, PLSR, LS-SVM, BP-NN	Qingdao Fushan Mountain foothills, Qingdao Zaoshan Mountain farmland, and Qingdao Licun River	total carbon (TC), total nitrogen (TN), total phosphorus (TP), total potassium (TK), available nitrogen (AN), available phosphorus (AP), available potassium (FK), and slowly available potassium (SK)	BPNN and LS-SVM for different nutrients	R^2, RMSEP, RPD	Li et al. (2019)	
24.	SVM and 13 ANN models	-	Soil Nutrient	GRNN for nutrients	Prediction accuracy through fitting	Li et al. (2014)	

without the need for pruning. Since RF is indifferent to the range of values, it is generally resistant to overfitting and does not require standardization or normalization. The number of trees (ntree) and the number of features randomly sampled at each split need both be changed for the RF model (mtry). (Zhou et al., 2020). This modeling technique is commonly used in soil mapping investigations because it can assess the importance of variables, is resistant to overfitting, and produces consistent and reliable results (Wiesmeier et al. (2019); Yang et al. (2019)).

Artificial Neural Networks

Artificial neural networks have shown good performance in predictive modeling and forecasting, as well as nonlinear and impermanent time series of processes where there is no definite answer and clear relationship plots the input to the output using training data and then uses the model to predict the output when the outcome is unknown. This model is sometimes used in place of a feed-forward network. (John et al., 2020). The training of the weight matrix characterizes an ANN model with a feed-forward network. The weights are randomly assigned to appropriate ranges and then adjusted using various training processes (Pachepsky et al. (1996); Schaap et al. (1998)). Different approaches, such as gradient descent (GD), Levenberg–Marquardt (LM), and Conjugate Gradient, are used to reduce error in feed-forward networks (CG). Backpropagation (BP) is based on the gradient descent (GD) algorithm, which is relatively stable when using a modest learning rate but has sluggish convergence properties (Farjam et al., 2014)

K – Nearest Neighbor

to detect and explain them. The multilayer perceptron model is the most commonly used ANN model (MLP). To learn and train the network, the MLP requires a well-understood output; this sort of neural network is known as a supervised network. MLP creates a model that

The K-nearest neighbor (KNN) classifier is a simple non-parametric that labels unknown instances based on the known instance (Cover & Hart, 1967). K-nearest training set vectors (k) were identified for the test set, and maximum summed kernel densities of five were computed for classification. Continuous variables can also be predicted using the average values of K nearest neighbors for regression. The maximum value of k (kmax), the distances between the nearest neighbors (distance), and the types of kernel function are all parameters of KNN (kernel). (Zhang et al., 2020).

Cubist Regression

The cubist model is a rule-based model which is an extension of the M5 tree model. (Quinlan, 1992). The structure of the cubic regression model contains an MLR model coupled with a conditional component acting as a decision tree. The simplification of the model is done by eliminating or pruning the rules. The key advantage of the cubist system is that it allows you to add additional training committees and enhance the weights to make them more balanced (Kuhn and Johnson (2013); Quinlan (1992); Wang and Witten (1997)). The cubist model includes boosting with training committees (typically more than one), which is comparable to the approach of "boosting" by constructing a sequence of trees with changed weights successively (John et al., 2020).

MACHINE LEARNING IN MAPPING METHODOLOGIES:

The soil properties estimation and mapping were employed majorly based on two methodologies or applications

1. Spectral-based modeling and Mapping,
2. Digital Soil Mapping

Spectral Based Modelling and Mapping

out over the years. So far affordable alternatives using different statistical studies have been universally employed in estimating soil properties.

Various statistical approaches have been used to translate the spectral information to those quantified soil properties and to develop spectral models for soil properties characterization. Hyperspectral Data is rich in information but the processing associated with the data is usually a bit complicated and poses several challenges considering the data complexity, information redundancy, modeling accuracy, and high correlation between the spectral bands (i.e.) High collinearity. Considering the above-mentioned vulnerabilities and disadvantages, an attempt is made to develop a methodology to identify the Bands of a particular wavelength that corresponds with the soil nitrogen properties (Gomez et al., 2008).

Once spectrum reflectance is known and a relationship between the spectral feature and soil characteristic is known a priori, the specificity allows for the assessment of various soil nutrients. As a result, spectral fingerprints are frequently regarded as inherent soil features that differ amongst soils. Das et al. (2015) also defined soil reflectance as a confluence of the responses of the electromagnetic radiation from different soil factors referred to as chromophores Essentially, physical (particle size and sample geometry) and chemical (moisture content, organic matter, clay minerals, and iron oxides) chromophores contributes to the spectral characteristics of the soil under study. The spectral reflectance of soil increases

The extractable soil information from that remote sensed image data or Vis-NIR Spectroscopy has been utilized for mapping the soil properties and to evaluate and estimate the soil properties employing hyperspectral remote sensing (Vis-NIR or MIR Spectroscopy and Hyperspectral remote sensing).

studies have been implicated regarding the estimation of several soil properties such as N, P, K, EC, CEC, Fe, soil moisture, carbonates, and hydraulic properties. Usually, two major approaches are considered for the hyperspectral band selection 1. Using a specific absorption band (Table 2). The area of the selected absorption band is considered the explanatory variable to estimate the response value. 2. full spectra. Each spectral band value is an explanatory variable to estimate the response value (Gomez & Lagacherie, 2016).

The various mathematical and statistical methods conveyed the research into more on quantitative prediction with reflectance spectroscopy of various soil constituents, most of which have absorptions within the Vis-NIR spectral region such as water, organic matter, and carbonates. Because of the intercorrelations among these spectral constituents, soil properties such as cation exchange capacity, pH, and phosphorus can be estimated indirectly.

(Bajcsy & Groves, 2004) discussed methodology incorporating mathematical and statistical procedures for optimal band selection from the hyperspectral data. As the high collinearity and information redundancy are concerned, the best band selection is complicated. A methodology is implicated by separating the techniques of band Selection into supervised and unsupervised band selection Procedures. In unsupervised band selection, the approaches (i.e.) Information Entropy, Contrast Measures, Correlation, Derivative Analysis, and PCA have been utilized for best band selection. Supervised Band Selection is performed by training data sets to filter exponentially as particle size decreases, with the increment below 0.4 mm diameter. As the roughness increases with an increase in the particle size, more energy is trapped in the inter-aggregate gaps, resulting in poorer reflectance (Sadeghi et al., 2018). The soil spectra can be utilized for the estimation of the primary properties and some

best bands obtained, by utilizing approaches such as Regression Trees, Regression, and Instance-Based Methods.

Several statistical analysis has been used for the spectral band selection and validated through MLR which is associated with the error in the form of outliers, over or underfitting of the variables, and the multicollinearity associated with the if the variables selected. Hence, most of the pieces of literature recently have adopted ML-based variable selection and modeling. The pieces of literature have shown a better performance of the ML model over statistical models. Several of the literature determined the efficiency of the RF method and Generic Algorithm for the model selection.

(Gmur et al., 2012) studied soil samples by a field-based analysis of spectroradiometer providing a spectral range from 400-1000 nm. The ranking is done to find out the similarities between the replicated soil series and other soil series. The statistical analysis and classification are based on Regression Trees.

Regression Trees are formed keeping the spectral responses as the independent variable and the respective soil nutrient concentrations (i.e.) nitrogen, carbon, carbonate, and organic matter as dependent variables. The aim was to estimate soil nutrients and relate the spectral signatures to that of the laboratory reference measurements utilizing CART analysis. The spectral analysis of the SOC and TN is done utilizing three ML approaches ELM, PLS, and BPNN. The spatial estimation utilizing the ML is conducted with or without variable selection measures. The External Learning Machine approach provided the best estimate for the SOC and TN. (Reda et al., 2019).

Digital Soil Mapping

The term "digital soil mapping" refers to the use of geospatial technology for mapping soils (DSM). Digital soil mapping is the construction of geographically referenced soil databases based on quantitative correlations between spatially explicit environmental data and measurements taken in the field and laboratory (McBratney et al., 2003). The progress of digital soil mapping followed four different transformations. 1. Small to larger areas. 2. Simpler to complex landscapes 3. 2D to 3D, and 4. Agricultural and ecosystem management (ZHANG et al., 2017)

In a nutshell, digital soil mapping refers to the geographic prediction of soil parameters based on model calibration. The science of soil surveying has greatly advanced thanks to the availability and accessibility of geographic information systems (GIS), global positioning systems (GPS), remotely sensed spectral data, topographic data derived from digital elevation models (DEMs), predictive or inference models, and data analysis software (Boettinger et al., 2010).

Conventional soil mapping includes statistical and geostatistical modeling. The use of inference models to predict the soil parameters in Geodatabase makes the DSM more advantageous than the conventional. Besides the use of data mining, statistical analysis, and machine learning approaches have enhanced the accuracy of soil mapping (Wadoux et al., 2020).

The Generic framework has been implicated by the Generic framework McBratney et al. (2003). For regions where soil resource information is lacking, the scorpan SSPFe (soil spatial prediction function with spatially autocorrelated other or previously measured attributes of the soil at a point; (2) c: climate, climatic properties at a point; (3) o: organisms, including land cover and natural vegetation; (4) r: topography, including terrain attributes and classes; (5) p: parent material (6) a: age, or the passage of time; (7) n: location, either spatially or geographically.

The derived SCORPAN factors of Jenny 1941, are essentially used for the mapping of soil attributes. Several methodologies have been implemented for deriving data layers respective to a particular factor (Avello).

Trials have been made to implicate Random Forest based machine learning methods and also compared the efficiency of each of the major Machine learning methods. The RF has been considered uniquely efficient when compared to other machine learning and MLR technique (Table 5). (Zhang et al. (2019);Bousbih et al. (2019); Zhang et al. (2020); John et al. (2020); Bouslihim et al. (2021); Zhou et al. (2021); Adab et al. (2020)). Some of the pieces of literature have implemented ANN-Back Propagation means of classification and CNN means of soil properties estimation (Keshavarzi et al., 2015).

The default tuning methodology provided is sufficient considering the application used for the Digital soil Mapping. Parameter optimization is usually performed to increase the model calibration accuracy (Jat et al., 2018). The model validation is usually performed by comparing the fitted variable to that of the original data. The most popularly used measures of validation are R2 and RMSE values. Based on the validation parameters the efficiency of the model calibrated can be determined.

errors technique is applied. “The seven predicted scorpan factors, which are a generalization of Jenny's five factors, are as follows: (1) s: soil,

CONCLUSION

Machine learning techniques provide the means of optimized model calibration when compared to conventional geostatistical or statistical approaches. Though most of the research has used shallow learning approaches for soil properties estimation, deep learning neural networks were also implicated. The important highlights of Machine learning in Soil properties mapping are:

1. Machine learning approaches are considered efficient compared to the conventional means,

2. Many of the approaches in soil properties estimation (i.e.) Digital Soil Mapping, Pedo-transfer functions, and spectral-based mapping are shifting from geo-statistical or statistical modeling to Machine learning approaches.

3. In most of the soil properties estimation procedures, A random Forest method has been considered efficient.

4. From parameter optimization or tuning in the DSM to variable selection in spectral-based hyperspectral mapping, Machine learning tools have been utilized for its qualitative detection.

Through increased research into this concern, the means of rapid mapping of properties and optimization of the band selection methodologies specific to a soil parameter can be standardized.

References

Abd-Elmabod, S. K., Jordán, A., Fleskens, L., Phillips, J. D., Muñoz-Rojas, M., van der Ploeg, M., . . . de la Rosa, D. (2017). Modeling agricultural suitability along soil transects under current conditions and improved scenario of soil factors. In *Soil mapping and process modeling for sustainable land use management* (pp. 193-219). Elsevier.

Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M., & Ghalhari, G. A. F. (2020). Machine learning to estimate surface soil moisture from remote sensing data. *Water*, *12*(11), 3223.

Alexakis, D. D., Tapoglou, E., Vozinaki, A.-E. K., & Tsanis, I. K. (2019). Integrated use of satellite remote sensing, artificial neural networks, field spectroscopy, and GIS in estimating crucial soil parameters in terms of soil erosion. *Remote Sensing*, *11*(9), 1106.

Avello, T. D. https://www.nrcs.usda.gov/wps/PA_NRCSConsumption/download?cid=nrcseprd1079006&ext=pdf

Bajcsy, P., & Groves, P. (2004). Methodology for Hyperspectral Band Selection. *Photogrammetric Engineering and Remote Sensing journal*, *70*, 793-802. https://doi.org/10.14358/PERS.70.7.793

Behnood, A. (2018). Soil and clay stabilization with calcium-and non-calcium-based additives: A state-of-the-art review of challenges, approaches and techniques. *Transportation Geotechnics*, *17*, 14-32.

Boettinger, J. L., Howell, D. W., Moore, A. C., Hartemink, A. E., & Kienast-Brown, S. (2010). *Digital soil mapping: Bridging research, environmental application, and operation*. Springer Science & Business Media.

Boruvka, L., Valla, M., Donátová, H., & Nemecek, K. (2002). Vulnerability of soil aggregates in relation to soil properties. *Rostlinná Vyroba*, *48*(8), 329-334.

Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N., . . . Mougenot, B. (2019). Soil Texture Estimation Using Radar and Optical Data from Sentinel-1 and Sentinel-2. *Remote Sensing*, *11*(13), 1520. https://www.mdpi.com/2072-4292/11/13/1520

Bouslihim, Y., Rochdi, A., & Paaza, N. E. A. (2021). Machine learning approaches for the prediction of soil aggregate stability. *Heliyon*, *7*(3), e06480.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Costa, E. M., dos Anjos, L. H. C., Pinheiro, H. S. K., Gelsleichter, Y. A., & Marcondes, R. A. T. (2020). Spatial Bayesian belief networks: a participatory approach for mapping environmental vulnerability at the Itatiaia National Park, Brazil. *Environmental Earth Sciences*, *79*(14), 1-13.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Das, B., Sarathjith, M., Santra, P., Sahoo, R., Srivastava, R., Routray, A., & Ray, S. (2015). Hyperspectral remote sensing: opportunities, status and challenges for rapid soil assessment in India. *Current science*, 860-868.

Ding, J., Yang, A., Wang, J., Sagan, V., & Yu, D. (2018). Machine-learning-based quantitative estimation of soil organic carbon content by VIS/NIR spectroscopy. *PeerJ*, *6*, e5714.

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. *Advances in neural information processing systems*, *14*.

Farjam, A., Omid, M., Akram, A., & Fazel Niari, Z. (2014). A neural network based modeling and sensitivity analysis of energy inputs for predicting seed and grain corn yields. *Journal of Agricultural Science and Technology*, *16*(4), 767-778.

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PloS one*, *12*(1), e0170478.

Gao, X.-s., Yi, X., Deng, L.-j., Li, Q.-q., Wang, C.-q., Bing, L., . . . Min, Z. (2019). Spatial variability of soil total nitrogen, phosphorus and potassium in Renshou County of Sichuan Basin, China. *Journal of integrative agriculture*, *18*(2), 279-289.

Gholizadeh, A., Saberioon, M., Ben-Dor, E., Rossel, R. A. V., & Borůvka, L. (2020). Modelling potentially toxic elements in forest soils with vis–NIR spectra and learning algorithms. *Environmental Pollution*, *267*, 115574.

Gmur, S., Vogt, D., Zabowski, D., & Moskal, L. (2012). Hyperspectral Analysis of Soil Nitrogen, Carbon, Carbonate, and Organic Matter Using Regression Trees. *Sensors (Basel, Switzerland)*, *12*, 10639-10658. https://doi.org/10.3390/s120810639

Goldberg, S. (1989). Interaction of aluminum and iron oxides and clay minerals and their effect on soil physical properties: a review. *Communications in Soil Science and Plant Analysis*, *20*(11-12), 1181-1207.

Gomez, C., & Lagacherie, P. (2016). Mapping of primary soil properties using optical visible and near infrared (Vis-NIR) remote sensing. In *Land surface remote sensing in agriculture and forest* (pp. 1-35). Elsevier.

Gomez, C., Viscarra Rossel, R. A., & McBratney, A. B. (2008). Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma*, *146*(3), 403-411. https://doi.org/https://doi.org/10.1016/j.geoderma.2008.06.011

Jat, M. L., Stirling, C. M., Jat, H. S., Tetarwal, J. P., Jat, R. K., Singh, R., . . . Shirsath, P. B. (2018). Soil processes and wheat cropping under emerging climate change scenarios in South Asia. *Advances in agronomy*, *148*, 111-171.

Jin, X., Li, S., Zhang, W., Zhu, J., & Sun, J. (2020). Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. *Applied Sciences*, *10*(4), 1520.

John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., & Marcus Afu, S. (2020). Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land*, *9*(12), 487.

Keshavarzi, A., Sarmadian, F., Omran, E.-S. E., & Iqbal, M. (2015). A neural network model for estimating soil phosphorus using terrain analysis. *The Egyptian Journal of Remote Sensing and Space Science*, *18*(2), 127-135. https://doi.org/https://doi.org/10.1016/j.ejrs.2015.06.004

Kuhn, M., & Johnson, K. (2013). Regression trees and rule-based models. In *Applied predictive modeling* (pp. 173-220). Springer.

Li, H., Leng, W., Zhou, Y., Chen, F., Xiu, Z., & Yang, D. (2014). Evaluation models for soil nutrient based on support vector machine and artificial neural networks. *The Scientific World Journal*, *2014*.

Li, X.-Y., Fan, P.-P., Liu, Y., Hou, G.-L., Wang, Q., & Lv, M.-R. (2019). Prediction results of different modeling methods in soil nutrient concentrations based on spectral technology. *Journal of Applied Spectroscopy*, *86*(4), 765-770.

Liu, Z.-P., Shao, M.-A., & Wang, Y.-Q. (2013). Spatial patterns of soil total nitrogen and soil total phosphorus across the entire Loess Plateau region of China. *Geoderma*, *197*, 67-78.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1-2), 3-52.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., . . . Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104-116.

Norton, R. (2013). Focus on Calcium: Its role in crop production. *GRDC Updates Pap. Available online: https://grdc. com. au/resources-and-publications/grdc-update-papers/tab-content/grdc-updatepapers/2013/02/focus-on-calcium-its-role-in-crop-production (accessed on 25 September 2020)*.

Pachepsky, Y. A., Timlin, D., & Varallyay, G. (1996). Artificial neural networks to estimate soil water retention from easily measurable data. *Soil Science Society of America Journal*, *60*(3), 727-733.

Pittman, R., & Hu, B. (2020). Estimation of Soil Bulk Density and Carbon Using Multi-Source Remotely Sensed Data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *3*, 541-548.

Quinlan, J. R. (1992). Learning with continuous classes. 5th Australian joint conference on artificial intelligence,

Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L., & El Hadrami, E. M. (2019). A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometrics* *and Intelligent Laboratory Systems*, *195*, 103873. https://doi.org/https://doi.org/10.1016/j.chemolab.2019.103873

Riese, F. M., & Keller, S. (2019). Soil texture classification with 1D convolutional neural networks based on hyperspectral data. *arXiv preprint arXiv:1901.04846*.

Sadeghi, M., Babaeian, E., Tuller, M., & Jones, S. (2018). Particle size effects on soil reflectance explained by an analytical radiative transfer model. *Remote Sensing of Environment*, *210*. https://doi.org/10.1016/j.rse.2018.03.028

Schaap, M. G., & Leij, F. J. (1998). Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science*, *163*(10), 765-779.

Schaap, M. G., Leij, F. J., & Van Genuchten, M. T. (1998). Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Science Society of America Journal*, *62*(4), 847-855.

Shao, Y., & He, Y. (2011). Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Research*, *49*(2), 166-172.

Trontelj ml, J., & Chambers, O. (2021). Machine Learning Strategy for Soil Nutrients Prediction Using Spectroscopic Method. *Sensors*, *21*(12), 4208.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Wadoux, A. M.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359.

Wang, S., Wang, X., & Ouyang, Z. (2012). Effects of land use, climate, topography and soil properties on regional soil organic carbon and total nitrogen in the Upstream Watershed of Miyun Reservoir, North China. *Journal of Environmental Sciences*, *24*(3), 387-395.

Wang, Y., Li, M., Ji, R., Wang, M., & Zheng, L. (2020). Comparison of soil total nitrogen content prediction models based on Vis-NIR spectroscopy. *Sensors*, *20*(24), 7078.

Wang, Y., & Witten, I. (1997). Induction of model trees for predicting continuous classes. *Induction of Model Trees for Predicting Continuous Classes*.

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., . . . Garcia-Franco, N. (2019). Soil organic carbon storage as a key function of soils-A review of drivers and indicators at various scales. *Geoderma*, *333*, 149-162.

Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, *6*, 100045.

Yan, B., & Hou, Y. (2018). Effect of soil magnesium on plants: a review. IOP Conference Series: Earth and Environmental Science,

Yang, M., Xu, D., Chen, S., Li, H., & Shi, Z. (2019). Evaluation of machine learning approaches to predict soil organic matter and pH using Vis-NIR spectra. *Sensors*, *19*(2), 263.

Yousefi, S., Pourghasemi, H. R., Avand, M., Janizadeh, S., Tavangar, S., & Santosh, M. (2021). Assessment of land degradation using machine-learning techniques: A case of declining rangelands. *Land Degradation & Development*, *32*(3), 1452-1466.

ZHANG, G.-l., Feng, L., & SONG, X.-d. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of integrative agriculture*, *16*(12), 2871-2885.

Zhang, M., Shi, W., & Xu, Z. (2020). Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences*, *24*(5), 2505-2526.

Zhang, Y., Sui, B., Shen, H., & Ouyang, L. (2019). Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. *Computers and Electronics in Agriculture*, *160*, 23-30. https://doi.org/10.1016/j.compag.2019.03.015

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Science of The Total Environment*, *729*, 138244.

Zhou, T., Geng, Y., Chen, J., Sun, C., Haase, D., & Lausch, A. (2019). Mapping of Soil Total Nitrogen Content in the Middle Reaches of the Heihe River Basin in China Using Multi-Source Remote Sensing-Derived Variables. *Remote Sensing*, *11*(24), 2934.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., . . . Lausch, A. (2021). Prediction of soil organic carbon and the C: N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Science of The Total Environment*, *755*, 142661.