Approximating Soil Organic Carbon (SOC) using Remote Sensing and Machine Learning

Manuka Stratta
Stanford University
mstratta@stanford.edu

Andrew Hojel
Stanford University
ahojel@stanford.edu

Sasankh Munukutla
Stanford University
sasankh@stanford.edu

Abstract

Globally, soils have the potential to sequester 1.85 gigatons of carbon per year. With rising CO2 emissions and climate change, it is critical to leverage soils as carbon sinks. To implement and monitor practices that increase the carbon sequestration ability of soils, we need to measure soil organic carbon (SOC) which currently involves significant manual labor and high lab costs. This paper investigates the application of machine learning to satellite imagery (Remote Sensing) to approximate the amount of soil organic carbon, which would significantly decrease the cost of measurement. We examine the use and limitation of three different leading soil characterization datasets, combined with data from Landsat 7, Sentinel-1, and Sentinel-2. In addition, we investigate the use of various spectral indices and features to improve model performance. Our best model achieved an R^2 of 0.534 on an unseen test set.

1 Introduction

Climate change is a pressing problem that will require innovative solutions. The world's emissions amount up to 50 gigatons of CO_2 per year, with the average American emitting 15 tons of CO_2 per year. To reach net zero by 2050, climate scientists and technologists will need to turn to gigaton-scale solutions, such as soil organic carbon sequestration.

Soil organic carbon (or SOC) sequestration is defined as the "process of transferring CO_2 from the atmosphere into the soil through plants, plant residues and other organic solids, which are stored or retained as part of the soil organic matter" (Olson 2010). Given the past and current rates of emissions, it will not be enough to reduce emissions: we also need to remove excess CO_2 that has been building up in the atmosphere. Nature-based carbon removal solutions such as SOC sequestration have immense potential to reach scale, by leveraging the existing carbon cycle and helping soils and plants do what they do best: sequester carbon through photosynthesis. This nature-based carbon solution has the promise to reach scale, as soil organic carbon accounts for the largest proportion of terrestrial carbon, between 50 and 80% (Odebiri, Mutanga, et al. 2021). Soil organic carbon sequestration is also an appealing solution from a food security perspective. Soil degradation is a major problem that is affecting yields worldwide—problems that are exacerbated by changing weather conditions and drought. Soil degradation threatens food security, and the United Nations estimates that there could be 200 million soil refugees by 2050. The decline in soil fertility threatens the ability to feed a growing population, so it is increasingly important to regenerate our soils. Increasing the amount of soil organic carbon improves soil health while restoring biodiversity and resilience.

Globally, soils have the potential to sequester up to 1.85 gigatons of carbon per year (Zomer 2017). There is a massive opportunity to implement sustainable agriculture practices that sequester carbon accompanied with large environmental and economic gains. However, to implement such practices, it is critical to be able to accurately measure and monitor soil organic carbon through time to better adjust soil management practices, allocate resources, and measure impact. There are inherent challenges to measuring SOC, as soil is a complex mixture of organic and inorganic constituents with different physical and chemical properties, with large variability from site to site or even within the same field as found by Angelopoulou et al. 2019. However, it is currently very costly and difficult to measure SOC. Scientists need to visit the field, take many soil samples due to SOC stock variation, send these samples to a lab, and do dry combustion to measure the organic matter content of the soil samples, and then generalize the organic carbon content to the entire field using statistical methods. This process is simply too costly and not scalable enough to reach gigaton-scale sequestration potential.

Remote sensing (RS) could be a key tool to overcome these barriers and truly scale soil organic carbon sequestration (Angelopoulou et al. 2019). SOC values are highly correlated with variables that can be derived from remote sensing, such as vegetation condition indices, moisture, brightness. There are four key advantages of remote sensing applications to measure SOC: 1) RS is a non-destructive way to get information about soil properties, 2) RS data covers large geographical areas, 3) RS provides information about inaccessible areas, 4) RS provides the means to reduce traditional and costly soil sampling techniques (Angelopoulou et al. 2019). Remote sensing's promise to make SOC sequestration scalable is what motivated our research project of approximating soil organic carbon using remote sensing and machine learning.

2 Literature Review

2.1 Current Non-Remote Sensing Methods to Measure Soil Organic Carbon

Current methods to measure soil organic carbon include direct sampling, soil spectroscopy, in situ sensing, and process-based models. Direct sampling is the most common method to measure SOC, but it is expensive, laborious, and unscalable, as discussed above. Soil spectroscopy in the lab is an effective measurement method: Janik et al. conducted an in situ SOC spectra measurement using an ATR-FTIR (Fourier transform infrared spectroscopy) spectrometer in the mid-infrared region (2500–2500 nm) and obtained excellent results ($R^2 = 0.93$) (Janik, Forrester, and Rawson 2009). Although soil spectroscopy eliminates the cost of doing dry combustion, it still requires samples to be taken and sent to the lab. In situ sensing, which combines on-the-ground sensors and spectroscopy techniques, is promising as it cuts out the laboratory, but these sensors can be expensive and this technique wouldn't be applicable to inaccessible areas or large geographical areas.

Process-based models are also commonly used, such as the DNDC model (Denitrification–Decomposition) which, based on inputs like soil characteristics, temperature, and crop management, predicts greenhouse gas emissions (CO_2, CH4, N2O) and other aspects like crop growth (Giltrap, Li, and Saggar 2010). While process-based models perform well, they do not directly measure soil organic carbon and are thus not a reliable source of information that can be used to monitor the change in soil organic carbon over time with the implementation of a new agriculture practice.

Remote sensing techniques—using hyperspectral, multispectral, and radar data—could help estimate important indicators for soil monitoring purposes in a cost-effective and scalable way (Angelopoulou et al. 2019).

2.2 Remote Sensing + Deep Learning Approaches to Measure SOC

Soils are incredibly complex and dynamic systems, and remote sensing has the potential to drastically improve scientists' ability to understand and estimate the dynamics of soil. Soil organic carbon is highly correlated with variables that can be derived from remote sensing data, including brightness, wetness, and vegetation condition indices (Angelopoulou et al. 2019). RS techniques for SOC measurement can be split in three categories: hyperspectral RS, multispectral RS, and radar RS. Hyperspectral RS data can contain as many as 200 contiguous spectral bands, while multispectral data typically has 3-10 different band measurements for each pixel. For each of these types, we should distinguish spaceborne data (obtained with satellites) from airborne data (obtained with planes/drones flying over areas of interest).

Models using hyperspectral airborne data achieved the best results, such as Žížala et al. who predicted SOC using an airborne spectrographic imager and multilayer perceptron neural network, achieving $R^2 = 0.88$. However, hyperspectral airborne data is difficult and expensive to obtain, particularly over large areas, as it requires flying planes over the areas of interest (Odebiri, Mutanga, et al. 2021). Hyperspectral spaceborne imagery is promising and is becoming increasingly available, with authors such as Jaber et al predicting SOC using hyperspectral data from the Hyperion/Earth Observing-1 satellite. In the near future, more hyperspectral sensors (including PRISMA, HyspIRI, and EnMAP) with better spatial-spectral and temporal attributes will become available.

Multispectral RS data is more readily available and is the primary type of RS data used for estimating SOC. Were and Bui 2015 and Mirzaee et al. 2016 used multispectral data from Landsat 7 and derived vegetation indices as predictors for SOC, achieving a reasonable performance of $R^2 = 0.5 - 0.7$ (Odebiri and Odindi 2021). Radar RS data can be valuable in generating different topographic variables (slope, elevation) that affect SOC stocks. Radar is promising because of its long wavelength that can penetrate canopy and cloud cover and can be used in all weather conditions. Radar is also very sensitive to soil conditions such as moisture, which can improve SOC estimation (Odebiri, Mutanga, et al. 2021).

The application of deep learning for environmental remote sensing has rapidly increased in the past decade, and so have the use of neural networks and remote sensing-based techniques for SOC estimation. Nonetheless, the use of DL for retrieving environmental parameters such as SOC has been largely ignored compared to other applications of DL in agriculture (Odebiri, Mutanga, et al. 2021). DL is helpful as it can capture non-linear correlations between environmental properties. DL approaches that have been adopted for SOC modeling include extreme learning machine, multilayer perceptron, backpropagation neural networks, radial basis function, convolutional neural networks, and recurrent neural networks. Studies have demonstrated that DL produces better accuracy than geostatical and traditional ML approaches for SOC estimation (Odebiri, Mutanga, et al. 2021). However, due to computational requirements, neural networks have not been as used in practice as much as other computationally efficient geostatical and ML models such as partial least square regression, principal component analysis, support vector machine, and random forest.

Significant challenges with remote sensing for SOC measurement remain, such as low resolution, atmospheric and geometric corrections, vegetation cover, and soil moisture (Angelopoulou et al. 2019). Yet various spectral unmixing techniques have been promising for segregating bare soils from vegetation cover (Angelopoulou et al. 2019). There is still a large variation between models using RS data to estimate SOC, and the results are not consistently reproducible across datasets. Indeed, there is no universally accepted calibration method for SOC retrieval (Odebiri, Mutanga, et al. 2021). Moreover, DL approaches require large sample sizes which can be hard to obtain in the soil science space.

Despite the greater number of studies recently, the effectiveness of DL techniques for SOC modeling is still in its infancy with new freely available multispectral sensors, such as Sentinel with moderate resolution (Odebiri, Mutanga, et al. 2021). Reading about this in the literature motivated our project to leverage freely available multispectral sensors (Landsat 7 and Sentinel 1, 2) and apply deep learning techniques on different global soil datasets.

3 Task

The main objective of our ML models was to, given remote sensing data for a specific point and other soil properties, predict the amount of carbon in soil. When measuring the amount of carbon in soil, we can consider multiple metrics including SOC concentration and SOC stock. However, not all of our datasets provided both, so the task and final predicted output varied by dataset.

3.1 Predicting SOC Concentration or SOC Stock

Task 1 is predicting SOC Concentration (g C / kg soil). SOC Concentration is the amount of grams of carbon per kilogram of soil. This typically varies by the range of depths of soil considered. Most soil datasets have this value.

Task 2 is predicting SOC Stock (g C / m^2). SOC Stock is the amount of grams carbon per m^2 (unit of area). The calculation of SOC stock of a site requires determination of SOC concentration, bulk densities (BD), stone contents, and soil depth, which all vary in space and have different measurement errors associated as detailed by Schrumpf et al. 2011. Not all soil datasets have this value.

3.2 Metrics

Given both tasks are regression problems, we used the coefficient of determination (R^2) as the metric to evaluate performance. R^2 is the statistical measure of how well predictions approximate the real data points, with values ranging from 0 (worst) to 1 (best). The coefficient of determination is the most popular metric in the literature for this type of SOC modeling.

4 Soil Characterization Data

After consultation with soil scientists, we identified and compiled three primary global soil datasets which had the potential to be used for machine learning. None of these datasets were ready for machine learning out of the box, as we had to pull the corresponding satellite data that aligned temporally. The task and predicted output varied by dataset depending on whether SOC Concentration or SOC Stock was available.

4.1 Dataset Preparation

After mapping the soil data to the corresponding satellite data for each dataset, we split the data into train, val, and test sets using a 70%, 20% and 10% split respectively. Note that the same location can have multiple measurements across different time periods and soil depths, so to avoid location leakage between train and val or test data, we made sure to split the data by unique location—so that if multiple points are associated with the same location, they are all in the same split. We then normalized each of the dataset splits by their respective mean and standard deviation.

4.2 Datasets

In Table 1, we have provided the number of unique locations in each dataset used in the Baseline model (filtered to have SOC concentration or stock values and filtered by relevant imaging dates). If there was no satellite imagery available for a particular date, the data point had to be excluded.

Table 1: Dataset Unique Location Overview

Dataset	Original Locations	Filtered Locations	Train Locations	Val Locations	Test Locations
WoSIS	>196,000	8569	5998	1714	857
RaCA	6237	5790	4050	1161	579
SoDaH	1470	591	413	118	60

4.2.1 World Soil Information Service (WoSIS)

WoSIS (Batjes, Ribeiro, and Van Oostrum 2020) has the best global coverage of soil carbon data, which has been developed by harmonizing many country-level datasets by the International Soil Reference and Information Centre (ISRIC). WoSIS serves as a great candidate to develop a model that generalizes across regions. WoSIS has varying levels of geographic accuracy for its points and lacks spatial density (concentrations of points). Moreover, in harmonizing many datasets, WoSIS can be sparse in data across various features and also is terse in its features, for example, it does not have any land use data for its points. Despite WoSIS being the largest among the three datasets, when analyzing WoSIS data, we observed that the majority of the data is comprised of samples prior to 1999. This left only a small subset data that aligned with the temporal range of publicly available satellite data. WoSIS has only SOC Concentration meaning only Task 1 as defined in Section 3 is possible with this dataset.

4.2.2 Rapid Carbon Assessment (RaCA)

The RaCa dataset (Wills et al. 2014) was initiated by the USDA-NRCS in 2010, with the explicit purpose of mapping carbon stocks for U.S. soils under various land covers with differing agricultural management. More than 24 universities collaborated in assembling this dataset. Temporally, the dataset is limited to data from 2010 to 2011, restricting the type of satellite data with which it aligns. While geographically limited, the RaCA dataset does benefit from a greater spatial density of points and the availability of more features like land use. RaCA has both SOC Concentration and SOC Stock, making both tasks as defined in Section 3 possible with this dataset.

Acquiring the RaCa dataset proved particularly challenging in that we had to work with the USDA-NRCS to obtain the data and go through multiple layers of clearances.

4.2.3 Soil Data Harmonization (SoDaH)

The SoDaH dataset (Wieder et al. 2021) was assembled to bring together soil carbon data from diverse research networks (five networks) into a unified global dataset. Temporally, it has very recent readings which align with the more recent multispectral satellites like Sentinel-1 and Sentinel-2. The SoDaH dataset also has a variety of additional environmental and soil features like mean annual temperature and precipitation making it well suited to machine learning. However, it is a relatively small dataset and also suffers from a lack of spatial density. SoDaH has both SOC Concentration and SOC Stock making both tasks as defined in Section 3 possible with this dataset.

5 Satellite Data

Given the objective of this project, acquisition of satellite data for each individual measurement of soil organic carbon is essential. A major constraint on the options for data sources was imaging time periods. We tended to favor quantity of data available from a given dataset when choosing which satellite to use for a specific dataset. Although this prevented access to newer satellite equipment with higher resolution and improved technology, the amount of measurements available from WoSIS and RaCA within the time frame of newer imaging options was negligible. Therefore, we chose to use USGS Landsat 7 Surface Reflectance Tier 2 for WoSIS and RaCA given the ability to access imaging starting from 05-28-1999. For SoDaH, we decided to use Sentinel-1 SAR GRD and Sentinel-2 MSI Level-2A because there were enough point measurements past Sentinel-2's first imaging date of 03-28-2017. Now we will provide further analysis of each of these imaging options:

5.1 USGS Landsat 7 Surface Reflectance Tier 2

Landsat 7 provides atmospherically corrected surface reflectance and land surface temperature data. The imaging contains 4 visible and near-infrared (VNIR) bands, 2 short-wave infrared (SWIR) bands, and one thermal infrared (TIR) band processed to focus on surface reflectance. In addition, it includes intermediate bands used in calculations of the surface temperature products. Finally, quality assessment (QA) bands are reported to understand the usefullness of a given pixel. See Table 20 for in depth information about the bands returned by Landsat 7.

Landsat 7 began imaging on 05-28-1999 and images the same location on earth every 16 days at a 30m pixel resolution.

5.2 Sentinel-1 SAR GRD: C-band Synthetic Aperture Radar Ground Range Detected

Sentinel-1 provides data from a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument at 5.405GHz (C band). Nguyen et al. 2022 found that incorporating SAR with Multispectral bands could improve ability to predict soil organic carbon. See Table 21 for in depth information about the bands returned by Sentinel-1.

Sentinel-1 began imaging on 10-03-2014 and images the same location on earth every 10 days at a 10m pixel resolution.

5.3 Sentinel-2 MSI: MultiSpectral Instrument, Level-2A

Sentinel-2 is a high-resolution, multi-spectral imaging instrument designed to support the monitoring of vegetation, soil, and water cover. It has 9 VNIR bands, 2 SWIR bands, 2 bands to detect water vapor, 1 band for aerosol thickness, and 3 true color image (TCI) bands. Sentinel-2 provides significantly more spectral information than Landsat 7 at higher resolutions (bands vary between 10-60m resolution). See Table 22 for in depth information about the bands returned by Sentinel-2.

5.4 Geographic Coordinate System

When pulling band values for a certain location, we needed to take into account the geographic coordinate system (GCS), the way that latitude and longitude values are projected onto the Earth, for each dataset. The most popular GCS and the default used by Google Earth Engine is WGS_1984 (EPSG 4326). We identified that SoDaH used EPSG 4326 and reached out to the developers of RaCA and WoSIS to confirm that they used EPSG 4326, but we did not receive a response. We therefore assumed that all datasets used EPSG 4326.

5.5 Collecting Satellite Date

Given that the satellite imaging frequency is not daily for any of the satellites we used, is is not guaranteed that there is an image on the date that soil samples were taken. As a result, we pulled band values using date ranges. For some experiments, we took a single image from this range and for others we took the median over the images within the range. For each experiment, we will specify the date range and technique used to acquire the band values.

6 Models

Our approach to modeling was to build up from basic means of assessing data correlation to basic neural models and larger gradient boosting models. Our primary concern with the more basic models was gauging and understanding the quality of the data that we are generating. Given the fact that we worked with filtering the datasets across various features and independently pulled satellite data corresponding to point measurements, we wanted to understand the quality of the data we were generating.

6.1 Correlation Matrix

Correlation matrices provide the linear correlation between various input features. We focused on analyzing the correlation of our input features with SOC stock / concentration. Gholizadeh et al. 2018 found that the B4/B5 and B11/B12 bands from Sentinel-2 have the highest negative correlation with SOC concentration. Given that B4/B5 are Red/Red Edge and B11/B12 are SWIR, we can assume that SR_B4 (near infrared) and SR_B5/SR_B6 (SWIR) from Landsat 7 will likely have high negative correlation with SOC concentration as well. Based on this information, we compared our correlations of individual features to SOC concentration / stock to Gholizadeh et al. 2018 findings to tentatively gauge quality of data.

6.2 Linear Regression

The first model we applied to data was a linear regression to gauge correlation of all the features with SOC concentration / stock. This provided a good baseline before testing any non-linear neural models.

6.3 Neural Model V1

Neural Model V1 is the first neural model that we applied to our data. This model was meant to investigate the delta between linear and non-linear modeling methods. The model was very small to prevent over-fitting. The model architecture details can be found in Table 2.

Table 2: Architecture of Neural Model V1

Network Architecture	Optimizer	Learning Rate	Epochs
(0) Linear(in,64) (1) ReLU (2) Linear(64,1)	SGD	1e-3	100

6.4 Neural Model V2

Neural Model V2 includes more and larger linear layers to improve on the modeling capacity of Neural Model V1. Given the small size of many of our datasets, we added features such as Batch Normalization and Dropout to prevent overfitting. The model architecture details can be found in Table 3.

Table 3: Architecture of Neural Model V2

Network Architecture	Optimizer	Learning Rate	Epochs	Dropout Percentage	Batch Size
(0) Linear(in,128) (1-3) Dropout/BatchNorm/ReLU (4) Linear(128,128) (5-7) Dropout/BatchNorm/ReLu (8) Linear(128,1)	ADAM	1e-3	100	0.1	256

6.5 XGBoost

XGBoost by Chen and Guestrin 2016 is a scalable and distributed library for gradient boosted tree (GBDT) machine learning. It is often the default choice for a variety of machine learning tasks including regression and classification, given its strong performance and quick training time. Moreover, prior work by Nguyen et al. 2022 and Wang et al. 2021 has shown XGBoost based models achieve the best performance for regression tasks on both Task 1 and Task 2.

7 Features

Each of our chosen datasets came with a number of features related to soil properties. With the satellite imagery that we pulled (from either Landsat 7 or Sentinel-1/2, depending on the dataset), we augmented our data with band values and other important features. Feature engineering was important to improve model performance while making sure to maintain realistic assumptions about what kind of data would be available to us without taking extensive on-the-ground soil samples.

7.1 Spectral Indices

Spectral indices were key features for model performance. Spectral indices are combinations of spectral reflectance from one or more spectral bands that indicate the relative abundance of features of interest. Gholizadeh et al. 2018 calculated 18 spectral indices as covariates as they are expected to improve the prediction capability. This is because mineral composition, soil moisture, organic matter content, and soil texture influence soil optical properties. To retrieve variables through inter-correlation between target variables, it is helpful to calculate spectral indices such as water and vegetation indices. Here are the following indices we experimented with including as features:

- Normalized Differences Vegetation Index (NDVI): $\frac{NIR-Red}{NIR+Red}$.
- Bare Soil Index (BSI): $\frac{(R+SWIR)-(NIR+BLUE)}{(R+SWIR)+(NIR+BLUE)}$
- Modified Soil Adjusted Vegetation Index (MSAVI): $\frac{2NIR+1-\sqrt{2(2NIR+1)-8(NIR-R)}}{2}$

Using the band values we pull from the satellite, we were able to calculate these indices for all our datasets.

7.2 Regional Features

The SODAH dataset included other data specific to location, such as mean average precipitation (MAP) and mean average temperature (MAT). We included MAP and MAT as features to our model, since temperature and moisture affect soil organic carbon (Fissore). Additionally, these features can help calibrate our model to a particular region.

7.3 Non-Carbon Soil Characterization Features

Finally, our datasets contained data about soil properties besides SOC, such as Soil pH, Nitrogen (N), and Carbon to Nitrogen ratio (C:N). Although SOC and features like nitrogen are highly correlated, we decided to exclude most of these soil properties from our features because we wanted our model to primarily rely on purely spectral, remote data. In practice, it is just as difficult to collect information about nitrogen as it is to measure SOC. Our original motivation was to leverage remote sensing to globally scale SOC measurement and eliminate the need for ground samples, which is why we removed those features. The soil property features we kept include pH and elevation, which can be cheaply measured or are already known by the landowner/farmer.

8 Experiments & Results

8.1 Data Ablation Experiments

A particular difficulty of working with band values over a large range of locations is the lack of ability to confirm that we are imaging bare soil. Imaging bare soil is important because the bands measure surface reflectance, which is drastically different when the land is covered by trees or crops in comparison to imaging bare soil. Therefore, for each of our datasets we performed an ablation study comparing different tactics to isolate bare soil pictures with various trade offs. The rationale behind this decision was to ensure that we were using the best data to train our models before experimenting with various features.

To test different methods for isolating bare soil images and improving the data, we performed the following experiments:

- **Baseline**: Baseline where all measurements from a dataset were included. To acquire satellite imagery, we took the median of all images taken in a range of +/- 1 month of the sampling date.

- **Land Use (LU)**: The baseline data is filtered by land use (if available). To filter for land use we removed forest and wetland from the data, but kept cropland, rangeland, and pastureland. To acquire satellite imagery, we took the median of all images taken in a range of +/- 1 month of the sampling date.
- **Land Use + Max Bare Soil Index (LU + Max BSI)**: The baseline and land use datasets do not ensure that the soil is bare when imagery is pulled. Since none of the datasets provide information whether the soil was bare or not when sampled, we incorporated the Bare Soil Index into our satellite data acquisition for this dataset. BSI takes a value in [-1,1], where more positive values correlate with soil being more bare. Given that some of the soil samples were likely taken when there was vegetation at the site, when acquiring satellite imagery we increased the range to +/- 6 months of the sampling date and took the image with the highest BSI value (ie. the most bare soil).
- **Land Use + Max Bare Soil Index + NDVI Thresholding (LU + Max BSI + NDVI Thresh)**: When analyzing the Normalized Difference Vegetation Index (NDVI) of the values pulled using the BSI maximization strategy, we saw a large range of NDVI values. Government and academic websites find that NDVI within the range of 0.1-0.3 tends to indicate bare or near bare soil. Therefore, we decided to filter all of the satellite data pulled using BSI maximization to only include the values where NDVI falls within the range of [0.1,0.3] (). The imaging used was the same as **LU + Max BSI**.

8.2 WoSIS

WoSIS does not have land use data available with its soil measurements. Thus, the data ablation study for WoSIS performed Baseline, Max BSI, Max BSI + NDVI Thresh using Landsat 7.

The results for the data ablation are summarized by Tables 4, 5, 6. Looking at the results, we see that Max BSI + NDVI Thresh has the best validation performance with its Neural Model V1 having the best overall R^2 of 0.213, so we used as the baseline for the WoSIS Feature Ablation Study. From Table 5, we see that SR_B4, SR_B5, and SR_B6 are not very negatively correlated, but do become more negatively correlated using Max BSI + NDVI Thresh, which further justified the use of this data strategy. Note that literature does finds higher negative correlation for these bands (see Section 6.1)

Table 4: WoSIS Data Ablation Study - Dataset Splits

Data Ablation Type	Total Unique Locations	Total Points	Train Points	Val Points	Test Points
Baseline	8569	35749	25054	7122	3573
Max BSI	8433	35310	24691	7072	3547
Max BSI + NDVI Thresh	3600	15740	11079	3112	1549

Table 5: WoSIS Data Ablation Study - Feature Correlation with SOC Concentration

Variable	Baseline	Max BSI	Max BSI + NDVI
orgc_value_avg	1.00	1.00	1.00
lower_depth	-0.271	-0.272	-0.299
SR_B1	0.027	0.052	-0.001
SR_B2	0.028	0.048	-0.023
SR_B3	0.011	0.042	-0.057
SR_B4	0.014	0.053	-0.006
SR_B5	-0.051	0.025	-0.078
SR_B6	-0.051	-0.061	-0.011
SR_B7	-0.083	-0.003	-0.090
SR_ATMOS_OPACITY	0.002	0.047	0.051
SR_ATRAN	0.032	-0.008	-0.004

Table 6: WoSIS Data Ablation Study - Model Evaluation

Data Ablation Type	Linear Regression	Neural Model V1	Neural Model V2	XGBoost
Baseline	0.084 / 0.123	0.178 / 0.210	0.158 / 0.199	0.837 / 0.191
Max BSI	0.082 / 0.090	0.170 / 0.171	0.166 / 0.202	0.831 / 0.138
Max BSI + NDVI Thresh	0.101 / 0.112	0.196 / **0.213**	0.201 / 0.171	0.826 / 0.123

For each model, train/val R^2 performance is reported.

8.2.1 RaCA

The data ablation study for RaCA performed the Baseline, LU, LU + Max BSI, LU + Max BSI + NDVI Thresh tests using Landsat 7. In addition the following test was performed on the dataset:

- **Land Use + Custom BSI/NDVI Threshold (LU + BSI/NDVI Thresh)**: Sentinel-2 returns a scene classification value for each pixel pulled in an image. One of the classes is Bare Soils. Yet, given that all of the RaCA measurements were taken in 2011 and 2012, Sentinel-2 imagery was not available. For this strategy, we shift the year of all of the sampling dates to 2018 and pull Sentinel-2 data for +/- 6 months where only images classified as Bare Soil are included in the median returned. We then found the mean and standard deviation of BSI and NDVI for all the images pulled using Sentinel-2. These values are reported in Table 7. We then pulled Landsat 7 images where we took the median over images from +/- 6 months and where the NDVI was in the range [0.197, 0.393] (within one standard deviation of mean) and BSI was greater than 0.084 (one standard deviation below the mean values). The rationale for a lower threshold for BSI was that as BSI increases the likelihood of bare soil increases. The idea behind this strategy was to improve our bare soil filtering by using knowledge from Sentinel-2's classification algorithm.

It is worth noting that for RaCA we were attempting to estimate SOC stock instead of SOC concentration to see how this affected model performance given that the calculated SOC stock values were provided in the dataset. We approximated SOC stock at a depth of 5cm. There is a 1-1 mapping of SOC Stock calculations to unique locations (unlike other datasets).

The results for the data ablation are summarized by Tables 8, 9. 10. Looking at the results, we see that LU + Max BSI + NDVI Thresh has the best validation performance with Neural Model V2, achieving the best overall R^2 of 0.195, so we used this as a baseline for the RaCA Feature Ablation Study. The Baseline performance model is relatively high in comparison to other models tested, but we can see in Table 9 that SR_B4, SR_B5, and SR_B6 are not very negatively correlated and become significantly more negatively correlated for LU + Max BSI + NDVI Thresh. Note that literature finds higher negative correlation for these bands (see Section 6.1)

Table 7: RaCA: NDVI and BSI Mean and STD from Sentinel-2

Index	Mean	STD
NDVI	0.295	0.098
BSI	0.173	0.089

8.2.2 SoDaH

The data ablation study for SoDaH performed the Baseline, LU, LU + Max BSI, LU + Max BSI + NDVI Thresh tests using Sentinel-2. Since Sentinel-2 returns a land use classification, the reason we included LU + Max BSI and LU + Max BSI + NDVI Thresh was to see how effective these strategies were at isolating bare soil in comparison to Sentinel-2's classification algorithm. In addition the following test was performed on the dataset:

Table 8: RaCA Data Ablation Study - Dataset Splits

Data Ablation Type	Total Unique Locations	Train Points	Val Points	Test Points
Baseline	5790	4050	1161	579
LU	3631	2541	726	364
LU + Max BSI	3631	2541	726	364
LU + Max BSI + NDVI Thresh	1487	1041	296	150
LU + BSI/NDVI Thresh	911	639	182	91

Since we used SOC Stock values from RaCA each unique location maps to one SOC Stock value.

Table 9: RaCA Data Ablation Study - Feature Correlation with SOC Stock

Variable	Baseline	LU	LU+Max BSI	LU+Max BSI+NDVI Thresh	LU+BSI/NDVI Thresh
SOCstock5	1.000	1.000	1.000	1.000	1.000
SR_B1	0.048	0.033	0.067	-0.095	-0.099
SR_B2	0.037	0.008	0.060	-0.134	-0.147
SR_B3	0.021	-0.004	0.051	-0.142	-0.118
SR_B4	0.053	0.054	0.073	-0.063	-0.115
SR_B5	-0.054	-0.044	0.032	-0.134	-0.096
SR_B6	-0.095	-0.094	-0.183	-0.188	-0.231
SR_B7	-0.156	-0.144	-0.042	-0.205	-0.210
SR_ATMOS_OPACITY	0.184	0.115	0.108	-0.080	-0.101
SR_ATRAN	-0.143	-0.119	-0.032	0.107	0.198

Table 10: RaCA Data Ablation Study - Model Evaluation

Data Ablation Type	Linear Regression	Neural Model V1	Neural Model V2
Baseline	0.095 / 0.072	0.214 / 0.182	0.228 / 0.175
LU	0.080 / 0.066	0.111 / 0.080	0.156 / 0.035
LU + Max BSI	0.087 / 0.080	0.119 / 0.118	0.205 / 0.078
LU + Max BSI + NDVI Thresh	0.097 / 0.052	0.109 / 0.091	0.197 / **0.195**
LU+ BSI/NDVI Thresh	0.090 / 0.092	0.092 / 0.001	0.155 / 0.123

For each model, train/val R^2 performance is reported.

- **Baseline + Sentinel-1 (Baseline + S1)**: This is the same as the Baseline experiment but we include data from Sentinel-1 (which is also pulled for +/- 1 month and the median of these images is taken). The bands from Sentinel-1 are used in addition to the bands from Sentinel-2.
- **Land Use + Sentinel-1 (LU + S1)**: This is the same as the Land Use experiment but we include data from Sentinel-1 (which is also pulled for +/- 1 month and the median of these images is taken). The bands from Sentinel-1 are used in addition to the bands from Sentinel-2.
- **Land Use + Sentinel Classification (LU + SC)**: For this experiment we take the dataset filtered by land use and pull Sentinel-2 imagery +/- 6 months where we take the median image of all images classified as bare soil (by Sentinel-2's land use classification algorithm).

Given that RaCA had poor performance compared to WoSIS, we decided to approximate SOC Concentration using SoDaH instead of attempting to approximate SOC Stock.

The results for the data ablation are summarized by Tables 11, 12. 13. Looking at the results, we see that LU + SC has the best validation performance across all models and achieved the best overall R^2 of 0.391 using Neural Network V2, so we used this as a baseline for the SoDaH Feature Ablation Study. It was inconclusive whether Sentinel-1 improved predictive capacity of the models given it increased validation R^2 values in some cases and decreased in others. In addition, it is worth noting that LU + Max BSI + NDVI Thresh may have achieved such poor validation performance because of the small size of the dataset (79 unique locations overall and only 60 samples in the validation set) because train performance was similar to other models. Finally, in terms of comparing our mechanisms for finding bare soil it appears that Max BSI + NDVI Thresh is the most similar to Sentinel-2's classification algorithm in terms of getting high negative correlation for B4/B5 and B11/B12. This makes sense given the performance of Max BSI + NDVI Thresh on WoSIS and RaCA.

Table 11: SoDaH Data Ablation Study - Dataset Splits

Data Ablation Type	Total Unique Locations	Total Points	Train Points	Val Points	Test Points
Baseline & Baseline + S1	591	2916	1881	580	455
LU & LU + S1	239	1147	714	194	239
LU + Max BSI	115	759	463	121	175
LU + Max BSI + NDVI Thresh	79	339	240	60	39
LU + SC	218	996	646	177	173

Since we used SOC concentration, each unique location has measurements at different depths and dates (multiple measurements per locations).

Table 12: SoDaH Data Ablation Study - Feature Correlation with SOC Concentration

Variable	Baseline	LU	LU+Max BSI	LU+Max BSI+NDVI Thresh	LU+SC
lyr_soc	1.000	1.000	1.000	1.000	1.000
B1	0.152	0.226	0.102	-0.006	0.054
B2	0.126	0.108	0.085	-0.170	0.013
B3	0.108	0.047	0.078	-0.138	-0.035
B4	0.095	0.032	0.082	-0.132	-0.106
B5	0.105	0.096	0.096	-0.130	-0.050
B6	0.111	0.115	0.117	-0.108	0.052
B7	0.098	0.102	0.123	-0.101	0.072
B8	0.096	0.105	0.133	-0.134	0.118
B8A	0.092	0.125	0.132	-0.094	0.113
B9	0.151	0.242	0.136	-0.091	0.134
B11	-0.087	-0.143	-0.037	-0.229	-0.208
B12	-0.078	-0.158	-0.070	-0.207	-0.231
VV	0.093	0.077	-	-	-
VH	0.100	0.089	-	-	-
angle	-0.044	0.027	-	-	-
map	0.021	-0.116	-0.189	-0.145	-0.078
mat	-0.283	-0.437	-0.459	-0.408	-0.447
layer_top	-0.283	-0.233	-0.273	-0.225	-0.228
layer_bottom	-0.301	-0.244	-0.295	-0.238	-0.240
elevation	0.025	0.010	0.039	0.116	-0.000

Note that Sentinel-1 bands are reported for both Baseline and LU but they were only used on their respective experiments. In addition, mat stands for mean annual temperature, and map stands for mean annual precipitation.

Table 13: SoDaH Data Ablation Study - Model Evaluation

Data Ablation Type	Linear Regression	Neural Model V1	Neural Model V2
Baseline	0.289 / 0.166	0.465 / 0.306	0.584 / 0.387
Baseline + S1	0.295 / 0.180	0.457 / 0.301	0.532 / 0.380
LU	0.430 / 0.262	0.422 / 0.212	0.666 / 0.311
LU + S1	0.450 / 0.290	0.441 / 0.208	0.701 / 0.358
LU + Max BSI	0.418 / 0.265	0.308 / 0.269	0.542 / 0.372
LU + Max BSI + NDVI Thresh	0.393 / 0.073	0.172 / 0.053	0.779 / 0.084
LU+SC	0.384 / 0.332	0.461 / 0.384	0.679 / **0.391**

For each model, train/val R^2 performance is reported.

8.3 Feature Ablation Experiments

Once we determined the best strategy to isolate for bare soil pixels, we performed an ablation study using different features available from each dataset. The purpose of experimenting with features was to better understand how they affected predictive capacity. Given the limited size of our datasets, we didn't want to include features that were not valuable to help models learn more effectively.

We experimented with the following features individually for all models: NDVI, BSI, MSAVI, before using an ensemble of all the best features. Note for the case of NDVI, although we previously filtered for NDVI, we never included it explicitly as a feature. These features are used in conjunction with the best data strategy as determined by the Data Ablation Study for each specific dataset. Note that same dataset splits (from the best data strategy) are used for the feature ablation study.

8.3.1 WoSIS

The results for the feature ablation are summarized by Tables 14, 15. In addition to NDVI, BSI, and MSAVI, pH (soil) is available for WoSIS, so this was also used as a feature to study. These features are used on top of the Max BSI + NDVI Thresh data strategy as determined by the Data Ablation Study.

Looking at the features added and their correlation, we see that relative to the existing features (band values), they have a relatively high correlation in terms of magnitude, justifying their inclusion in the feature ablation study. After adding each of the features individually, NDVI, BSI and pH improved performance relative to the baseline, while MSAVI seemed to significantly worsen performance. As a result, NDVI, BSI, and pH were selected for the ensemble. Looking at the results, we see that just adding BSI has the best validation performance with its Neural Model V1 having the best overall R^2 of 0.261, which is an improvement from the best model from Data Ablation Study (R^2 of 0.213). It is interesting to note that just adding BSI performed better than the ensemble.

Table 14: WoSIS Feature Ablation Study - Added Feature Correlation with SOC Concentration

Feature Added	Correlation
NDVI	0.098
BSI	-0.127
MSAVI	0.099
pH	-0.068

Table 15: WoSIS Feature Ablation Study - Model Evaluation

Feature Added	Linear Regression	Neural Model V1	Neural Model V2	XGBoost
Best Data Strategy	0.101 / 0.112	0.196 / 0.213	0.201 / 0.171	0.826 / 0.123
NDVI	0.110 / 0.091	0.203 / 0.186	0.203 / 0.142	0.812 / 0.126
BSI	0.093 / 0.151	0.180 / **0.261**	0.185 / 0.173	0.811 / 0.174
MSAVI	0.146 / 0.054	0.282 / 0.098	0.295 / 0.081	0.567 / 0.072
pH	0.101 / 0.140	0.208 / 0.238	0.259 / 0.156	0.874 / 0.182
Ensemble (NDVI + BSI + pH)	0.095 / 0.149	0.208 / 0.239	0.248 / 0.216	0.863 / 0.184

For each model, train/val R^2 performance is reported. Best Data Strategy refers to Max BSI + NDVI Thresh from the WoSIS Data Ablation study.

8.3.2 RaCA

RaCA was tested with adding each of the three indices individually on top of the Max BSI + NDVI Threh data strategy as determined by the Data Ablation Study. Since none of the indices improved performance, we did not test an ensemble model. The results for the feature ablation are summarized by Tables 16, 17.

One potential interpretation for the fact that none of the indices improved the results is that the indices provide less value when attempting to predict SOC stock.

Table 16: RaCA Feature Ablation Study - Added Feature Correlation with SOC Stock

Feature Added	Correlation
NDVI	0.139
BSI	-0.183
MSAVI	-0.016

Table 17: RaCA Feature Ablation Study - Model Evaluation

Data Ablation Type	Linear Regression	Neural Model V1	Neural Model V2
Best Data Strategy	0.097 / 0.052	0.109 / 0.091	0.197 / **0.195**
NDVI	0.100 / 0.054	0.107 / 0.079	0.201 / 0.128
BSI	0.099 / 0.043	0.102 / 0.096	0.22 / 0.178
MSAVI	0.110 / 0.053	0.121 / 0.067	0.201 / 0.193

For each model, train/val R^2 performance is reported. Best Data Strategy refers to Max BSI + NDVI Thresh from the RaCA Data Ablation study.

8.3.3 SoDaH

The results for the feature ablation are summarized by Tables 18, 19. In addition to NDVI, BSI, and MSAVI, ph_h2o (soil pH in water), lyr_n_tot (nitrogen concentration), and two additional bands from Sentinel-2 (AOT-Aerosol Optical Thickness and WVP-Water Vapor Pressure) are available for SoDaH and are also used as a feature to study. These features are used on top of the Land Use + Sentinel-2 Classification data strategy as determined by the Data Ablation Study.

From the individual feature ablations, we saw that AOT, WVP, ph_h2o, lyr_n_tot, and MSAVI all improved model performance. We decided not to include nitrogen in the emsemble model given its extremely high intial correlation with SOC concentration. The resulting emsembled outperformed all other models tested achieving and R^2 of 0.568 on validation and achieved an R^2 of 0.534 on the test set, which was held out throughout all other testing of models using SoDaH. This shows that the model can generalize well on unseen locations.

Table 18: SoDaH Feature Ablation Study - Added Feature Correlation with SOC Stock

Feature Added	Correlation
NDVI	0.335
BSI	-0.318
MSAVI	0.076
AOT	0.105
WVP	-0.051
lyr_n_top	0.919
ph_h2o	-0.373

9 Discussion

The results from our study were promising given that our final test R^2 of 0.534 using our model began to get closer to state-of-the-art values found in existing literature. We will discuss potential reasons why others have been able to achieve higher values and investigate areas of uncertainty that contributed to the poor performance of RaCA and WoSIS particularly.

Table 19: SoDaH Feature Ablation Study - Model Evaluation

Data Ablation Type	Linear Regression	Neural Model V1	Neural Model V2
Best Data Strategy	0.384 / 0.332	0.461 / 0.384	0.679 / 0.391
NDVI	0.387 / 0.337	0.448 / 0.374	0.688 / 0.306
BSI	0.410 / 0.203	0.445 / 0.440	0.754 / 0.330
MSAVI	0.409 / 0.357	0.451 / 0.447	0.740 / 0.337
ph_h2o	0.428 / 0.228	0.482 / 0.430	0.795 / 0.438
lyr_n_tot	0.848 / 0.849	0.875 / 0.857	0.941 / 0.824
AOT + WVP	0.420 / 0.206	0.486 / 0.420	0.739 / 0.530
Ensemble (AOT + WVP + ph_h2o + BSI + MSAVI)	0.425 / 0.254	0.492 / 0.392	0.791 / **0.568 / 0.534**

For each model, train/val R^2 performance is reported. Best Data Strategy refers to Land Use + Sentinel-2 Classification as determined by the SoDaH Data Ablation study.

The large delta in performance between RaCA/WoSIS and SoDaH could be caused by many different factors. One potential explanation is that Sentinel-2 provides more relevant data to approximating SOC. Despite WoSIS having many more points than SoDaH, WoSIS performs significantly worse, further underscoring the importance of data quality and accounting for various satellite types. The poor performance of RaCA, despite its relatively high spatial density (RaCA has points more spatially concentrated) was rather unexpected and is also worth noting. In addition, it appears that predicting SOC stock may have affected RaCA's performance as well, as this is arguably a more complex task. Another explanation for this general performance discrepancy is that our attempts to identify bare soil for RaCA and WoSIS were suboptimal compared to the Sentinel-2 Classification algorithm.

The delta between our results and state-of-the-art results may be due to the fact that SoDaH is such a small dataset in comparison to other larger private datasets. Another likely explanation is the high uncertainty surrounding our acquisition of satellite imagery. For all of the datasets, we were pulling satellite data from multiple-month ranges, and for all of the top performing models we were looking at the median band values over a year. Ideally, we would like to just look at values from a smaller period when there is bare soil.

10 Future Work

Moving forward, the best way to improve our models would be to systematically address these points of uncertainty. For example, we could limit the region of our measurements to a small region where we understand the agricultural system. This would allow us to significantly tighten the range of dates used to pull bare soil satellite images. It could also be interesting to take multiple pixels from a small region directly surrounding the point measurement and run CNN-based models, which would also allow us to take advantage of pretrained CNN models. There is also an exciting opportunity to work with time series data and attempt to predict the delta in SOC concentration between two measurements instead of predicting SOC concentration just using imaging.

Given current data, the need for model innovation is clear. Data augmentation such as synthetic data generation using GANs to augment the limited data available would be interesting to explore. Conversion between satellite data types such as from Landsat to Sentinel as done by Isa et al. 2021 could be a promising approach to overcome the temporal restrictions posed by existing soil datasets. Fusion of various RS data types like hyperspectral, multispectral and radar data has also shown to improve performance as shown by Schulte to Bühne and Pettorelli 2018.

11 Conclusion

This study has proven the importance of the need for high quality data, and most importantly, organizing this data for SOC estimation. We see this as our best results were obtained using SoDaH, which allowed us to use Sentinel-2 imagery due to its recency. With appropriate permissions, we aim to release our data as it is ML-ready, to encourage more research in this space. Our study has also laid the ground work for more complex modeling. As we focused on evaluating multiple datasets and a variety of features, we used relatively simple DL models.

On the whole, we are excited to help advance innovation in the remote-sensing and SOC estimation space, with this project serving as just the beginning.

12 Contributions

All members contributed to this research, the writing of the report, and creation of the final poster. Here we will provide a breakdown of each person's specific contributions:

Manuka

- Managed satellite data acquisition with Google Earth Engine API, focusing on Landsat 7/8 and Sentinel 2. Implemented the acquisition of satellite data while maximizing certain criteria, such as the BSI and NDVI spectral indices.
- Derived spectral indices from band values and contributed to feature engineering. Experimented with variations of our neural model V1 architecture and hyperparameter tuning.
- Worked with Andrew on data processing, modeling, and satellite data acquisition for SoDaH. Together we performed ablation study for data and features on SoDaH.

Andrew

- Managed acquisition of Sentinel-1 satellite data for RaCA and incorporated Sentinel-2's land use classification algorithm to pull bare soil images for SoDaH.
- Built data pipelines for opening and processing WoSIS (gpkg format). Constructed WoSIS dataset by merging various databases and filtering for relevant soil samples. Performed initial modeling for WoSIS.
- Managed data acquisition, data processing, modeling, and satellite data acquisition for RaCA. Performed ablation study for data and features on RaCA.
- Worked with Manuka on data processing, modeling, and satellite data acquisition for SoDaH. Together we performed ablation study for data and features on SoDaH.

Sasankh

- Managed modeling and satellite data acquisition for WoSIS. Performed ablation study for data and features on WoSIS.
- Worked on refining and finalizing the neural network and XGBoost models used by all datasets.
- Worked with soil scientists and experts to identify SoDaH dataset and understand advantages and limitations of all existing datasets. Investigated other datasets that showed potential, but were not suitable for study/ML-based approach.

13 Appendix

Table 20: Detailed Band Information for Landsat 7

Name	Wavelength (μm)	Description
SR_B1	0.45-0.52	Band 1 (blue)
SR_B2	0.52-0.60	Band 2 (green)
SR_B3	0.63-0.69	Band 3 (red)
SR_B4	0.77-0.90	Band 4 (near infrared)
SR_B5	1.55-1.75	Band 5 (shortwave infrared 1)
SR_ATMOS_OPACITY	-	gauge of atmospheric opacity
ST_B6	10.40-12.50	Band 6 surface temperature
ST_ATRAN	-	Atmospheric Transmittance
ST_CDIST	-	Pixel distance to cloud
ST_DRAD	-	Downwelled Radiance
ST_EMIS	-	Emissivity estimated from ASTER GED
ST_EMSD	-	Emissivity standard deviation
ST_QA	-	Uncertainty of the ST band
ST_TRAD	-	Thermal band converted to radiance
ST_URAD	-	Upwelled Radiance
QA_PIXEL	-	Pixel quality from CFMASK algorithm
QA_RADSAT	-	Radiometric saturation QA

Table 21: Detailed Band Information for Sentinel-1

Name	Units	Description
HH	dB	Single co-polarization, horizontal transmit/horizontal receive
HV	dB	Dual-band cross-polarization, horizontal transmit/vertical receive
VV	dB	Single co-polarization, vertical transmit/vertical receive
VH	dB	Dual-band cross-polarization, vertical transmit/horizontal receive
andle	Degrees	Approximate incidence angle from ellipsoid

Table 22: Detailed Band Information for Sentinel-2

Name	Pixel Size (m)	Wavelength (nm)	Description
B1	60	443.9 (S2A) / 442.3 (S2B)	Aerosols
B2	10	496.6 (S2A) / 492.1 (S2B)	Blue
B3	10	560 (S2A) / 559 (S2B)	Green
B4	10	664.5 (S2A) / 665 (S2B)	Red
B5	20	703.9 (S2A) / 703.8 (S2B)	Red Edge 1
B6	20	740.2 (S2A) / 739.1 (S2B)	Red Edge 2
B7	20	782.5 (S2A) / 779.7 (S2B)	Red Edge 3
B8	10	835.1 (S2A) / 833 (S2B)	NIR
B8A	20	864.8 (S2A) / 864 (S2B)	Red Edge 4
B9	60	945 (S2A) / 943.2 (S2B)	Water vapor
B11	20	1613.7 (S2A) / 1610.4 (S2B)	SWIR 1
B12	20	2202.4 (S2A) / 2185.7 (S2B)	SWIR 2
AOT	10	-	Aerosol Optical Thickness
WVP	10	-	Water Vapor Pressure
SCL	20	-	Scene Classification Map
TCI_R	10	-	Truce Color Image (red)
TCI_G	10	-	Truce Color Image (green)
TCI_B	10	-	Truce Color Image (bed)
MSK_CLDPRB	20	-	Cloud Probability Map
MSK_SNWPRB	10	-	Snow Probability Map
QA60	60	-	Cloud mask

References

[1] Theodora Angelopoulou et al. “Remote sensing techniques for soil organic carbon estimation: A review”. In: *Remote Sensing* 11.6 (2019), p. 676.

[2] Niels H Batjes, Eloi Ribeiro, and Ad Van Oostrum. “Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019)”. In: *Earth System Science Data* 12.1 (2020), pp. 299–320.

[3] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[4] Asa Gholizadeh et al. “Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging”. In: *Remote Sensing of Environment* 218 (2018), pp. 89–103.

[5] Donna L. Giltrap, Changsheng Li, and Surinder Saggar. “DNDC: A process-based model of greenhouse gas fluxes from agricultural soils”. In: *Agriculture, ecosystems environment* 136.3-4 (2010), pp. 292–300.

[6] Sani M Isa et al. “Supervised conversion from Landsat-8 images to Sentinel-2 images with deep learning”. In: *European Journal of Remote Sensing* 54.1 (2021), pp. 182–208.

[7] L. Janik, S. Forrester, and A. Rawson. “The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis”. In: *Chemometrics and Intelligent Laboratory Systems* 97.2 (2009), pp. 179–188.

[8] S. Mirzaee et al. “Spatial variability of soil organic matter using remote sensing data”. In: *CATENA* 145 (2016), pp. 118–127.

[9] Thu Thuy Nguyen et al. “A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion”. English. In: *Science of the Total Environment* 804 (Jan. 2022), pp. 1–12. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2021.150187.

[10] Omosalewa Odebiri, Onisimo Mutanga, et al. “Deep learning approaches in remote sensing of soil organic carbon: a review of utility, challenges, and prospects”. In: *Environmental monitoring and assessment* 193.12 (2021), pp. 1–18.

[11] Omosalewa Odebiri and John Odindi. “Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review”. In: *International Journal of Applied Earth Observation and Geoinformation* 102 (2021).

[12] Kenneth R Olson. “Impacts of tillage, slope, and erosion on soil organic carbon retention”. In: *Soil science* 175.11 (2010), pp. 562–567.

[13] Marion Schrumpf et al. “How accurately can soil organic carbon stocks and stock changes be quantified by soil inventories?” In: *Biogeosciences* 8.5 (2011), pp. 1193–1212.

[14] Henrike Schulte to Bühne and Nathalie Pettorelli. “Better together: Integrating and fusing multispectral and radar satellite imagery to inform biodiversity monitoring, ecological research and conservation science”. In: *Methods in Ecology and Evolution* 9.4 (2018), pp. 849–865.

[15] Huan Wang et al. “Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed”. In: *Remote Sensing* 13.7 (2021), p. 1229.

[16] K. Were and D. T. Bui. “A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape.” In: *Ecological Indicators* 52 (2015), pp. 394–403.

[17] William R Wieder et al. “SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0”. In: *Earth System Science Data* 13.5 (2021), pp. 1843–1854.

[18] Skye Wills et al. “Overview of the US rapid carbon assessment project: sampling design, initial summary and uncertainty estimates”. In: *Soil carbon*. Springer, 2014, pp. 95–104.

[19] R. J. Zomer. “Global Sequestration Potential of Increased Organic Carbon in Cropland Soils”. In: *Sci Rep* 7 (2017), p. 15554. URL: `https://doi.org/10.1038/s41598-017-15794-8`.