Methodology Proposal for Soil Organic Carbon Estimation

REMOTE SENSING AND MACHINE LEARNING

Authors:

Hywel Evans, MBE PhD BA- *Chief Technical Officer*

David Wright, BSc – *Chief Executive Officer*

Version: 1.0

Document ID: ecometric-0038

Last updated: 01/08/2022

https://www.ecometric.co.uk

Table of Contents

List of Tables

List of Figures

List of Equations

Units

Metric Tonnes	t
Hectare	Ha
Total Soil Volume	cm^3
Soil Bulk Density	g/cm^3
Dry Soil Weight	g
SOC Percentage	%
SOC	t/ha
SOC Stock	t
CO_2eq	t

Abbreviations

Soil Organic Carbon	SOC
CO2 equivalents	CO2e
Cation Exchange Capacity	CEC
Environmental Systems Research institute	ESRI
Keyhole Markup Language	KML
Keyhole Markup Language Zipped	KMZ
Geographic Java Script Object Notation	GeoJSON
Artificial Neural Network	ANN
Machine Learning	ML
Earth Observing System	EOS
Compact Geographic Stratification	CGS
Geographic Information System	GIS
Global Navigation Satellite System	GNSS

1. METHODOLOGY OVERVIEW

1.1. SCOPE

This methodology protocol uses soil sampling and remotely sensed multispectral imagery, coupled with machine learning, to estimate SOC stocks in a project area and areas beyond those which have been sampled.

The main soil health indicators assessed in this methodology are:

1. Carbon Sequestration
    - Soil Organic Carbon (SOC) stocks
    - CO2 equivalents (CO2e)
2. Additional Co-benefits are:
    - pH
    - Nitrogen, Phosphorus, Potassium
    - Cation Exchange Capacity - CEC
    - Calcium, Magnesium, Potassium, Sodium, Aluminum

1.2. MOTIVATION

Soil contains approximately 2344 Gt of organic carbon globally and is the largest terrestrial pool of organic carbon (Stockmann *et al.*, 2013). Small changes in the soil organic carbon stock could result in significant impacts on the atmospheric carbon concentration. By monitoring the carbon levels in the soil, farmers and landowners will be able to measure the impact of their stewardship.

1.3. OUTLINE

The following steps will be followed to estimate change in soil organic carbon stocks within a project area:

1. Develop a soil sampling plan for the project area.
2. Sample collection and preparation.
3. Laboratory analysis of soil samples.
4. Estimation of SOC stocks using Machine Learning (ML) and remotely sensed multispectral imagery.
5. Convert SOC stocks to CO2 equivalent stocks.
6. Calculate the change in CO2e stocks between monitoring periods.

SOC stocks measured in the first sampling round (Baseline), will be compared to those calculated in subsequent sampling rounds to quantify changes in carbon stocks after project commencement. This methodology outlines an innovative approach based on using remote sensing data to train a neural network to estimate SOC stocks, within and beyond the project area. This approach also allows for a significant reduction in the number of soil samples and in turn reduces cost.

2. PROJECT BOUNDARIES

2.1. SPATIAL BOUNDARIES

Spatial boundaries of the project area will be defined, including any parcels or stratification schemes, using one or more of the following data formats:

1. ESRI polygon shapefiles
2. KML/KMZ
3. GeoJSON.

2.1.1. MASKING

To ensure correct estimation of SOC stocks, any man-made objects not included within the bounds of the project area will be excluded using masking.

2.2. TEMPORAL BOUNDARIES

The project timeframe will be defined as the period during which SOC stocks will be monitored. The monitoring period and frequency defining the temporal boundaries will adhere to the following:

1. The minimum number of soil sampling rounds for a 10-year crediting period will be five.
2. Soil sampling rounds will be conducted on the first and last years of the project.
3. If possible two soil sample rounds will occur consecutively during the first two years.
4. If possible two soil sample rounds will occur consecutively during the last two years.
5. The minimum duration between monitoring periods will be one year.
6. The maximum time between soil sampling rounds will be three years.

This proposal will initially be based on annual sampling rounds. The schema described above will be modified if an extreme climatic event or disaster is declared in or near the project area.

3. CALCULATING CARBON SEQUESTRATION USING REMOTELY SENSED MULTISPECTRAL IMAGERY AND NEURAL NETWORKS

3.1. REMOTE SENSING IMAGERY AS PROXY FOR SOC

Satellite imagery and other remote sensing data has been shown to provide a proxy for SOC, approaches were mainly based on spectral indices and some used machine learning. An example of the spectral index approach, Thaler, Larsen and Yu, (2019), developed a SOC index (SOCI) using three bands of WorldView-2 imagery, with central wavelength (ρ):

$$SOCI = \frac{\rho_{478}}{\rho_{659} - \rho_{546}} \tag{1}$$

Bartholomeus *et al.*, (2008) tested, in laboratory conditions, the performance of several spectral indices which had been developed to detect biochemical constituents (e.g., cellulose, lignin) for their ability to retrieve SOC. They found correlation for indices based on the visible part of the spectrum (R2 = 0.80) and for the absorption features related to cellulose (around 2100 nm) (R2 = 0.81). Rasel *et al.*, (2017) used remotely sensed variables such as elevation and forest type rather than image pixel values to estimate SOC. Gardin *et al.*, (2021) used meteorological data, a land use map and MODIS Normalised Difference Vegetation Index (NDVI) imagery. This information was processed by advanced statistical methods to map SOC spatial distribution. Guo *et al.*, (2021) estimated SOC and soil bulk density (SBD) through partial least square regression (PLSR) and extreme learning machine (ELM) neural networks. They found that the combination of Sentinel 2 images and ELM obtained the best prediction results. ELMs are not as accurate as traditional backpropagation networks; they are generally used with problems that require real-time retraining of the network. Their method resulted in a correlation between image reflectance and SOC% with R^2=0.67. In this methodology Sentinel-2 multispectral data will be used as the proxy (inputs) and soil sample samples will be used to provide ground truth SOC (targets) data.

3.2. NEURAL NETWORKS

This methodology will use a Machine Learning (ML) in the form of an Artificial Neural Network (ANN). ML has been used by other carbon sequestration methodologies, as reviewed by Odebiri, Odindi and Mutanga, (2021). A further review investigated remote sensing techniques for SOC estimation, highlighting the significant wavelengths and the use of ML (Angelopoulou *et al.*, 2019). Both shallow and deep learning networks have been reviewed with reference to using Sentinel-2 imagery by Odebiri, Odindi and Mutanga, (2021). The advantage of neural networks is that they can train directly from high dimensional data, such as the bands in multispectral imagery. However, neural network training is stochastic, which produces slightly different predictions from each training session. Therefore, the network in this methodology will be run multiple times and the results and reported accuracy will the mean of the multiple iterations. This approach reduces the potential for extremes to occur and smooths the data to provide more reliable and conservative estimates. The neural network code will be compiled and packaged as a standalone App.

3.3. SOIL SAMPLING METHODS

3.3.1. INITIAL

Compact Geographic Stratification (CGS) will be the initial soil sampling scheme used to collect the Year 0 baseline soil samples. The main reason for choosing the CGS method was its suitability to areas where there is no previous knowledge of variability of SOC or a proven suitable proxy to provide accurate stratification (De Gruijter, Minasny and Mcbratney, 2015). Also, in laboratory experiments targeted at SOC and remote sensing data CGS was used by Bartholomeus *et al.*, (2008). CGS makes the assumption that the variability is spatially correlated and the sub-sample areas are smaller than the global variability (de Gruijter *et al.*, 2016).

3.3.2. SUBSEQUENT

1. To scale the initial sampling method and reduce sampling costs the following schemes may be implemented in follow on sampling rounds:
2. Knowledge of the variability in the Yr0 baseline SOC may enable stratified sampling. To enable the use of stratified sampling the variability will need to be shown to be consistent or predictable year on year.
3. Reducing the number of soil samples will be dependent on there being no significant change of ANN prediction accuracy.

4. The CGS grid size may be increased in areas where the variability has been shown to be larger than the chosen grid size.
5. In parallel with ANN training on annual soil samples, the baseline and subsequent years' networks will be tested on following years data. If there is no significant reduction in prediction accuracy this will pave the way forward for a method without or very reduced soil sampling.

Schemes, where the number of samples are reduced, will select core positions randomly from the recorded Yr0 core positions to ensure multitemporal comparability. In order to implement any of the above subsequent sampling schemes several annual sampling rounds will be required to provide evidence of viability.

3.4. PROCESSING WORKFLOW

The following workflow outlines the method used to estimate SOC stocks using Sentinel-2 imagery, ancillary data and ML. Other high-resolution imagery may be used. All images and ancillary data included in the analysis will be specified in the project report. The workflow sequence will be:

1. Soil sampling.
2. Ancillary data, if used.
3. Extracting samples
4. Sample analysis.
5. Image and ancillary data pre-processing.
6. Neural network training data.
7. Neural network training.
8. Estimating SOC with the previously trained neural network.
9. Project reporting.

3.4.1. SOIL SAMPLING SCHEME

The proposed CGS method (3.3.1) will use a hectare grid system with 10 or more composite cores per grid square. The location of the cores will be selected at random, (Gisel Booman *et al.*, 2021). A GIS random points within polygon tool (QGIS[2] or similar) will be used to select the sampling locations in each section of the grid and the geo-locations will be recorded in the report. The sample locations will remain consistent between monitoring rounds.

3.4.2. ANCILLARY DATA

Ancillary data may be used to augment the ML training dataset. The ancillary data may include:

1. Date.
2. Soil type.
3. Temperature.
4. Rainfall.
5. Soil moisture.

[2] https://qgiscloud.com/

6. Nitrogen levels.
7. Slope.
8. Altitude.

The ancillary data will meet the following requirements:

1. The sample dates for the project area and the sample dates for the area providing the ancillary data will fall within one month of each other.
2. The project area and the area providing the ancillary data will be within the same climatic region according to the Köppen Climate Classification System,(Beck et al., 2018).
3. The project area and the area providing the ancillary data will have been under the same management practices for at least 3 years.
4. The project area and the area providing the ancillary data will have similar soils and vegetation cover.
5. The sample extraction methods and sample analysis methods at the ancillary area will match the protocols used for the primary area.

3.4.3. EXTRACTING SAMPLES

To maintain the integrity of the results the Regen Network soil sampling guide (G Booman *et al.*, 2021) will be the main reference. The following method will be used to collect soil samples:

1. Prior to core extraction, the sample location will be cleared of living plants, plant litter and surface rocks.
2. Recommended sampling depth in the Regen Network guide is 15cm, however, for more accurate measurement SOC, this proposal will use a core depth of 30 cm.
3. The sampling depth will be the same for all samples. The only exception to this will be where the nominated sampling depth cannot be reached due to bedrock or impenetrable layers. The sampling depth will be recorded.
4. The sampling depth will be consistent between all sampling rounds.
5. Each core will be georeferenced using a GNSS device with have an accuracy 4 metres or better.
6. Samples will be taken at least 10 meters away from any tree, structure, or body of water.
7. The date/time of each core will be recorded for each sampling round.
8. All sampling rounds will occur at least 6 months after the application of non-synthetic fertilizer.

3.4.4. SAMPLE ANALYSIS

To quantify SOC (t/ha), percent soil organic carbon (%), total soil volume (cm^3) and bulk density (g/cm^3) will be calculated for each sample:

$$Total\ Soil\ Volume = No\ of\ Cores * \pi * Core\ Radius^2 * Sample\ Depth \tag{2}$$

$$Soil\ Bulk\ Density = \frac{Dry\ Soil\ Weight}{Total\ Soil\ Volume} \tag{3}$$

$$SOC = Core\ depth * Bulk\ Density * SOC \tag{4}$$

Table 1 in Regen Network soil sampling guide (G Booman *et al.*, 2021) will be used to comply with Regen Network laboratory specific instructions, laboratory accreditation requirements and approved laboratories.

3.4.5. IMAGERY SOURCE

Twelve bands of the Sentinel-2[3] multispectral image data (Table 1) will be used as the input for ML training; band 10 will not be used as it is only used for cirrus detection. The data will be downloaded from EOS Landviewer[4]. The download will include the respective metadata. Higher resolution data is available from EOS, including 0.4 meter per pixel and may be used in future projects.

Table 1 Sentinel-2 Multispectral Bands

Band Number	Central Wavelength (nm)	Bandwidth (nm)	Spatial Resolution (m)	Remarks
1	443	20	60	Aerosols
2	490	65	10	Blue
3	560	35	10	Green
4	665	30	10	Red
5	705	15	20	Red Edge 1
6	740	15	20	Red Edge 2
7	783	20	20	Red Edge 3
8	842	115	10	Near IR
8a	865	20	20	Red Edge 8
9	945	20	60	Water Vapour
~~10~~	1375	~~30~~	~~60~~	~~Cirrus Detection~~
11	1610	90	20	SWIR 1
12	2190	180	20	SWIR 2

[3] https://eos.com/find-satellite/sentinel-2/

[4] https://eos.com/products/landviewer/

3.4.6. IMAGE AND ANCILLIARY DATA PROCESSING

Imagery with a sensing date +/- 4 months around the sampling date will be used and will be Cloud and defect free in the project area. Ancillary data dates will not have to fall within the +/- 4-month sensing period if no significant change in the measured variable has occurred. Sentinel-2 image processing will include:

1. Processed to Level 2-A to provide BOA reflectance values.
2. Coordinate conversion as required (British National Grid for UK projects); Sentinel-2 tiles are ortho-images natively in UTM/WGS84 projection.
3. Resampling imagery and ancillary data to the same resolution (normally 10m).
4. Band stacking to create a 12-band multispectral image.
5. If multiple images are available within the +/- 4-month period, images will be averaged to smooth the data and reduce the effect outlying spectral values could have on analysis. Image preprocessing will be done before the average is performed.

3.4.7. NEURAL NETWORK TRAINING DATA

The QGIS[5] Point Sampling tool (or analogous tool) will be used to extract multispectral reflectance data at each soil sampling location. This data will then be paired with its respective SOC values to create a training dataset for the neural network.

3.4.8. NEURAL NETWORK TRAINING

The neural network will be trained using a set of design parameters that will remain fixed for each sampling project. The training may be run multiple times to determine the optimum parameters. Multiple training sessions will also be used to smooth the effects of the stochastic nature of neural networks.

3.4.9. ESTIMATE SOC USING A TRAINED NEURAL NETWORK

The trained network and Sentinel-2 imagery of the required project area will be used to estimate SOC at sampled and unsampled locations. The estimated SOC [t/ha] will be exported as a raster map(s).

3.4.10. ACCURACY

Cross validation will be used to quantify accuracy and uncertainty for the network results. The dataset will be split into two parts, a training set which will be used to train the network and a test set used to test the network accuracy. Validation will use a subset of the ground truth soil sample data. The accuracy will be quantified using the mean absolute percentage error (MAPE), where number of values (n), actual value (A_t), estimated value (F_t):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_{t-}F_t}{A_t}\right| \quad (5)$$

3.4.11. CALCULATING SOC STOCKS

To ensure only the required areas are included in the final soil organic carbon stock estimate, the project area and/or sub-areas will use masking. The QGIS[5] zonal statistics tool (or equivalent

5 https://qgis.org/en/site/

tool) will be used to sum all pixels contained within the project area or sub-areas (fields) to calculate the SOC stocks[t]. The spatial resolution of the pixels will be adjusted, if required, to match units of tonnes per hectare before calculating the sum. The results will be compared with calculating the SOC stocks using the mean sampled SOC[t/ha], in those areas where sampling was carried out.

3.4.12. CONVERTING SOC STOCKS TO CO_2 EQUIVALENTS

Converting soil organic carbon stocks to equivalent stocks (CO_{2e}) will be done by multiplying the SOC stocks [t] by a conversion factor of 3.67 (the ratio of the molecular mass [Da] of carbon dioxide (44) to that of carbon (12):

$$CO_2e[t] = SOC[t] \times \left(\frac{44}{12}\right) \tag{6}$$

3.4.13. WORKFLOW AUTOMATION

Automating the workflows within the methodology reduces the amount of time and work needed to complete a monitoring round. The automation also provides method integrity between monitoring dates and projects. The workflows that have been automated are image processing, SOC estimation using ML and the calculation of CO_2 equivalent. The automated workflows use QGIS modeler and a custom ANN App.

4. CALCULATING THE CREDITABLE CARBON CHANGE

4.1. BASELINE DEFINITION

The baseline SOC stocks or CO2e are defined here as the total carbon stocks calculated for the project's Initial Monitoring Date (IMD), or date of the first sampling round. The methodology will use a static baseline for each project, calculated as the total SOC stocks[t], from the IMD. All sampling rounds after the IMD will be compared to the baseline to calculate creditable carbon change.

4.2. CHANGES IN CO_2E BETWEEN REPORTING PERIODS

The change in SOC stocks between reporting periods is estimated as the difference between the total SOC stocks ($tSOC_{(t+1)}$) from the second monitoring period, minus total SOC stocks ($tSOC_{(t)}$) from the previous period:

$$SOC\ Stock = tSOC_{(t+1)} - tSOC_{(t)} \tag{7}$$

The same applies for estimating the change in the total SOC converted into CO_2 equivalents (CO_2e) between two sampling periods:

$$CO_2change = CO_2e_{(t+1)} - CO_2e_t \tag{8}$$

4.3. NET CO_2E REDUCTION

The net CO_2e reduction in the project area for a given reporting period is calculated as the difference between the changes in SOC, expressed as metric tons of CO2e, minus the total Greenhouse Gas Emissions (GHG), also in CO2e units:

$$NET\ CO_2eReduction = CO_2e\ Change - E_{Livestock} - E_{Fertilizer} \tag{9}$$

4.4. UNCERTAINTY AND DEDUCTIONS

Under this methodology framework, the total uncertainty for the project will be the sum of the uncertainties calculated throughout the methodology during a given monitoring period. Sources of uncertainties for creditable carbon stock calculations may include:

1. Percent soil organic carbon estimates.
2. Bulk density estimates.

3. Deviations from the original methodology which might have introduced additional errors.

Acceptable methods for calculating uncertainty for percent soil organic carbon and bulk density will be in line with Regen Network Supplement S.2 (Gisel Booman *et al.*, 2021). If the uncertainty (U) for the reporting period is less than or equal to 20%, the net CO2 reduction value will be used without making any deductions to account for uncertainty (Uncertainty Deduction (UD) = 0). If uncertainty is greater than 20%, the UD values in Table 2 will be used to calculate the amount of uncertainty to deduct from the creditable carbon stocks. UD values are based on the Gold Standard LUF activity requirements Version 1.2.126 (Gold Standard, 2020).

Table 2 Ranges of uncertainties and the corresponding discounts

UNCERTAINTY (U)	Uncertainty Deduction (UD) (% of U)
U≤20%	-No Deduction-
20%<U≤30%	50% of U
30%<U≤40%	75% of U
40%<U≤50%	100% of U

The maximum uncertainty allowed for any measurement in a project is 50%. The Creditable Carbon Change after UD will then be estimated as:

$$Creditable\ Carbon\ Change = (Net\ CO_2e\ Reduction) \times (1 - UD) \qquad (10)$$

5. DATA REPORTING

5.1. REPORT

After each monitoring round, a report will be submitted to the Regen Registry including a description of the methods used for soil sampling, analysis of samples, as well as the equations and references that were used. The reported results for each section of this Methodology will be accompanied by the supporting data. In the case of GIS or remote sensing data the SOC maps will be included as images within the report for illustrative purposes. The original vector and raster files will be kept by ecometric ltd. Any documentation containing calculations and statistical analysis will also be saved.

5.1.1. LABORATORY REPORT

The laboratory report will include:

1. Tools and methods used to estimate number of samples.
2. Sample stratification method and stratification map.
3. Tool used to extract soil cores.
4. If core sampler used, include tool diameter in mm.
5. Georeferenced location of each sample location and sub-samples (if applicable).
6. Coordinate reference system of sample and estimate locations.
7. GNSS device used to record sample locations.

8. The raw lab reports.

6. DATA STORAGE

All data used during the analysis will stored for monitoring verification. This data includes:

1. All raster and vector data used in geospatial analysis to generate results for any section of the methodology.
2. A copy of all laboratory reports.
3. All the relevant field data from the soil sample collection process (dates, tools, procedures, sample locations).
4. Documentation outlining calculations and results of statistical analysis.

7. DATA VERIFICATION

The following data will be verified:

1. Data reported in the soil lab reports.
   - Percent soil organic carbon
   - Bulk density
2. Spectral values extracted from the satellite imagery and any ancillary data.
   - Download the original imagery and ancillary data and follow the image pre-processing steps.
   - Spectral values should be extracted and compared to the data used as ML training data.
3. SOC Results
   - ML results used to estimate SOC should be re-created and compared to sampled values, where available.
   - Final SOC stock estimates should be recreated and compared to sampled values, where available.

ecometric.

References

Angelopoulou, T. *et al.* (2019) 'Remote sensing techniques for soil organic carbon estimation: A review', *Remote Sensing*, 11(6), pp. 1–18. doi: 10.3390/rs11060676.

Bartholomeus, H. M. *et al.* (2008) 'Spectral reflectance based indices for soil organic carbon quantification', *Geoderma*, 145(1–2), pp. 28–36. doi: 10.1016/J.GEODERMA.2008.01.010.

Beck, H. E. *et al.* (2018) 'Present and future köppen-geiger climate classification maps at 1-km resolution', *Scientific Data*, 5, pp. 1–12. doi: 10.1038/sdata.2018.214.

Booman, Gisel *et al.* (2021) 'Regen Methodology for GHG and Co-Benefits in Grazing Systems', *Regen*, (Version 0.91), pp. 1–42. Available at: https://regen-registry.s3.amazonaws.com/Methodology+for+GHG+and+Co-Benefits+in+Grazing+Systems.pdf.

Booman, G *et al.* (2021) *Regen Network Whitepaper*.

Gardin, L. *et al.* (2021) 'Mapping soil organic carbon in Tuscany through the statistical combination of ground observations with ancillary and remote sensing data', *Geoderma*, 404, p. 115386. doi: 10.1016/J.GEODERMA.2021.115386.

Gold Standard (2020) 'Land Use & Forests Activity Requirements', *Gold Standard*, 1.2.1(October 2019), pp. 1–44. Available at: https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fglobalgoals.goldstandard.org%2F200-gs4gg-land-use-forests-activity-requirements%2F&data=02%7C01%7Csusan.lecomte%40student.hogent.be%7Ccc6e834ee20440e03fe208d8355547e0%7C5cf7310e091a4bc5acd7.

de Gruijter, J. J. *et al.* (2016) 'Farm-scale soil carbon auditing', *Geoderma*, 265, pp. 120–130. doi: 10.1016/j.geoderma.2015.11.010.

De Gruijter, J. J., Minasny, B. and Mcbratney, A. B. (2015) 'Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error', *Journal of Survey Statistics and Methodology*, 3(1), pp. 19–42. doi: 10.1093/jssam/smu024.

Guo, L. *et al.* (2021) 'Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas', *Geoderma*, 398, p. 115118. doi: 10.1016/J.GEODERMA.2021.115118.

Odebiri, O., Odindi, J. and Mutanga, O. (2021) 'Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review', *International Journal of Applied Earth Observation and Geoinformation*, 102(June), p. 102389. doi: 10.1016/j.jag.2021.102389.

Rasel, S. M. M. *et al.* (2017) 'Proxies for soil organic carbon derived from remote sensing', *International Journal of Applied Earth Observation and Geoinformation*, 59, pp. 157–166. doi: 10.1016/J.JAG.2017.03.004.

Stockmann, U. *et al.* (2013) 'The knowns, known unknowns and unknowns of sequestration of soil organic carbon', *Agriculture, Ecosystems & Environment*, 164, pp. 80–99. doi: 10.1016/j.agee.2012.10.001.

Thaler, E. A., Larsen, I. J. and Yu, Q. (2019) 'A New Index for Remote Sensing of Soil Organic Carbon Based Solely on Visible Wavelengths', *Soil Science Society of America Journal*, 83(5), pp. 1443–1450. doi: 10.2136/sssaj2018.09.0318.