Table of Contents

Funding: Research in this publication was supported by the Foundation for Food and Agriculture Research under award number – Grant ID: DFs-18-0000000012

Attribution: This report was commissioned by ESMC/ESMRC and authored by Dr. Patrick Lawrence[1], Dr. Gisel Booman[2], Alexandra Thornton[1] & Dan Kane under contract to ESMC, January 2022.

[1]*Repliculture LLC*
[2]*Regen Network Development, INC*

Submission Contact: Patrick Lawrence; patrick@repliculture.com

Executive Summary

The profusion of geospatial data layers over the last 20 years has provided a wealth of information that can potentially be used to predict soil organic carbon (SOC) concentrations and bulk density (BD), the two primary inputs for calculating SOC stocks. Simultaneous increases in computing power have enabled these data layers to be statistically combined using simple and complex algorithms to improve predictions. Growth in academic, peer-reviewed modeling literature has mirrored these increased capabilities, resulting in a huge array of layer combinations and statistical methods, each with varying accuracies across agricultural systems.

Despite this progress, spatial modeling of SOC stocks is still constrained by availability of ground-truth data which are needed to calibrate and validate each model's efficacy. Therefore, while these approaches are being incrementally improved and tested, there will still be a need for fieldwork to collect concrete, trusted measurements directly from farms and ranches in order to generate Scope 1 credits or Scope 3 assets.

In landscapes targeted for fieldwork, spatial SOC stocks typically co-vary with underlying pedological, geological, climatic, and topographic controls. Many of these factors vary on smaller spatial scales than the farm boundary. As a result, soil sampling efforts that take advantage of these patterns through the use of stratification as compared to random sampling will, in most cases, result in greater accuracy, lower sample size requirements, and consequent reductions in cost.

In this report, we first outline the relationships between SOC accumulation and landscape controls in order to better interpret the correspondence between existing geospatial data layers (including data gaps) and SOC/BD. This understanding serves to ground our review of academic modeling literature, in which we assess geospatial data layers and algorithms used for prediction. Next, we evaluate the efficiency of different stratification approaches, most of which use geospatial data layers as an input. Finally, we provide recommendations on the use of modeling algorithms and stratification sampling techniques as a way to estimate SOC stocks at the field scale.

Landscape Controls on SOC Stocks

In the majority of landscapes, the magnitude of SOC stocks and stock accumulations is likely to be determined by a range of interacting soil, topographic, biotic, and climatic factors. To help unpack these complex influences, the following questions can assist our exploration:

1. Where are conditions amenable to the stabilization of soil organic carbon?
2. Where is there sufficient net primary productivity (NPP) such that there are also sufficient organic matter inputs into soils?

Understanding these processes and how they might be reflected in readily available spatial data may help to provide tools for mapping soil carbon at the field scale and for selecting adequate sampling locations for direct measurement.

Stabilization of SOC

Accumulation of SOC at the field scale is synonymous with increasing SOC stocks. To enhance accumulation, SOC must not only be added to the soil system; it must also be retained and prevented from leaving the system via respiration, leaching and erosion. This process is referred to as SOC stabilization.

Recent research suggests that accessibility of SOC to microbes in soil pores in combination with the correct conditions for decomposition (i.e., sufficient water, oxygen, etc.) determines the stability of SOC[1]. Six et al., (2002)[2] suggested two pathways of SOC stabilization that are useful to this theory: 1) mineral protection, and 2) protection in microaggregates. These pathways are not mutually exclusive, and in fact they frequently overlap because clay minerals are key to the formation of aggregates.

Mineral protection happens as SOC particles form associations with clay mineral surfaces via a variety of different mechanisms, including H-bonding, Van der Waals forces, and polyvalent cation bridges[3,2,4,5,6]. As SOC is integrated into these organo-mineral complexes, they become more likely to be located on mineral surfaces and pore spaces that render them less accessible to being metabolized by microbes and lost as CO_2. SOC protected via mineral association is often some of the oldest in soils and is generally composed of low molecular weight carbon compounds. Various metal cations, such as Ca, Al, and Fe, and their oxidized forms, appear to play an important role in the formation of organo-mineral complexes and this form of SOC protection, as does the clay content of soils.

Aggregation is the process by which smaller particles in soils bind together to form larger particles. This binding is often facilitated by the presence of the organo-mineral complexes described above and the

[1] Dungait *et al.*, (2012)
[2] Six *et al.*, (2002)
[3] Kleber, Sollins, and Sutton (2007)
[4] Feng *et al.*, (2019)
[5] Cotrufo *et al.*, (2013)
[6] Lützow *et al.*, (2006)

activity of soil microbes which produce organic compounds that can act as binding/cementing agents[7,8,9]. As aggregates form, they can trap organic matter of a variety of size classes and states of decomposition in low oxygen, low moisture environments, occluding them from microbial attack and decomposition. However, disturbance can often break aggregates apart, exposing previously protected SOC to microbes, meaning aggregate-protected SOC is less stable than mineral-protected SOC. However, metal cations and clay particles seem to play a similarly important role in the formation of aggregates[2].

Given the central role of clay and related minerals in both of these stabilization pathways, spatial data corresponding to soil mineral composition and concentration may be the most useful in mapping, modeling, and sample stratification efforts.

Sufficient NPP

New SOC is derived primarily from photosynthetic carbon, meaning sufficient net primary productivity is a key determinant of whether or not SOC will accumulate in a given area regardless of how amenable soil conditions may be to stabilizing carbon. Globally, circumpolar tundra has the highest belowground carbon accumulation because wet, cold conditions limit decomposition. But outside of those regions, high NPP tropical and temperate rainforests along with grasslands have the highest concentrations of soil carbon[10].

At more regional scales, aboveground NPP can be an important predictor of soil carbon storage within similar ecosystems[11,12]. In addition, precipitation patterns and physical landscape features such as topography that dictate where precipitation accumulates have been demonstrated to be predictive of where SOC accumulates[13].

However, these variables are not universally predictive of where SOC accumulates and their predictive power does not always translate from broader scales (i.e., global, regional) to the field or watershed scale. For example, Peterson and Lajtha (2013)[14] found that in an experimental forest in Oregon, USA, while ANPP (aboveground NPP) was related to total litter fall and dissolved organic carbon content, it was not related to soil carbon content or areal stocks, suggesting that additional landscape and biogeochemical factors undermined the accumulation of SOC despite high levels of new organic C entering the system.

Instead, NPP appears to be most useful as a predictor of SOC when used in combination with other important data that is more specifically predictive of where SOC is likely to be stabilized. For example,

---

[7] Six *et al*., (2004)
[8] Six, Elliot, and Paustian (2002)
[9] Tisdall and Oades (1982)
[10] Amundsen (2001)
[11] Chen *et al*., (2018)
[12] Roman-Sanchez *et al*., (2018)
[13] Mondal *et al*., (2017)
[14] Peterson and Lajtha (2013)

Burke et al. (1984)[15] found that precipitation and clay combined were most predictive of where SOC accumulated in a series of samples taken from semi-arid landscapes.

Useful Predictors of SOC

Given the evidence presented above, below is a suggested list of variables that are likely to be predictive of SOC concentrations and stocks at landscape scales and therefore useful in both modeling and sample stratification efforts.

1. *Soil clay content*: Higher levels of clay would provide higher reactive mineral surface area for the formation of organo-mineral complexes, as well as an increased capacity for aggregation.
2. *Soil taxonomy*: Soil taxa are broadly based on weathering status and/or unique soil forming factors, both of which are related to potential soil carbon content and NPP.
3. *NDVI*: Normalized differential vegetation index is a unitless measure of relative plant productivity derived from remote sensing data. It is broadly correlated to ANPP and is globally available at high resolution. Other similar vegetation indices may also suffice.
4. Topographic *data*: Various spatial variables derived from digital elevation models can be predictive of multiple landscape processes that could dictate where SOC accumulates, such as erosion, accumulation of rainfall, etc.
5. *Soil pH*: Soil pH is largely related to the weathering status of soils - older, more highly weathered soils, such as ultisols, are more acidic. As such, pH is also predictive of the relative importance of different minerals and cation exchange complexes for protecting carbon in soils. Rasmussen et al., (2018)[16] suggest it as a variable for initial assessment to determine the relative importance of other variables in predicting soil carbon accumulation at a given location. Briefly, at pH < 7 Al and Fe organo-mineral complexes are the most predictive of soil carbon concentrations, while at pH > 7 Ca and clay are more predictive.

[15] Burke *et al.*, (1989)
[16] Rasmussen *et al.*, (2018)

Modeling

As outlined above, accumulation of SOC is an emergent biogeochemical process. Topographic, pedological, and ecological features of a landscape dictate where SOC might accumulate and how much. Based on this knowledge, relevant environmental covariates can be selected and combined within a variety of empirical models to estimate SOC concentrations, stocks and/or stock changes. These models can estimate values on a point basis but are frequently extended using geospatial data inputs to provide spatially explicit estimates. The list of statistical algorithms which are capable of making spatial estimates has greatly expanded in recent years, particularly with the development of machine learning and other advanced computational techniques. These algorithms and environmental covariates are the subject of this section.

NOTE: These models are distinct from process-based biogeochemical models (e.g., DNDC, DAYCENT) which attempt to model and forecast changes in SOC pools over time. Instead, the models discussed herein are focused on static estimation of SOC on a point or areal basis.

Objective

In this section, we reviewed literature from the past 20 years in order to identify algorithms and environmental covariates that have been successful in estimating soil organic carbon (SOC) concentration and stocks. It is intended that this review will assist ESMC by identifying existing knowledge clusters and gaps, thereby guiding development of the technological roadmap and pointing towards future research needs. To achieve this aim, we evaluate and provide recommendations on appropriate algorithms, accuracy metrics, and optimum sets of environmental covariates for mapping SOC, primarily at the field scale.

Search Process Methodology

In order to collect a comprehensive list of predictive algorithms and environmental covariates, we performed a deep literature search within the Scopus and Web of Science databases using the search expressions listed in Appendix A.

Search strings were selected to maximize inclusion of articles of interest. Multiple expressions were separated by the 'OR' operator while searching the databases. All search expressions included the string 'soil' OR 'carbon' OR 'organic matter' OR 'bulk density'. Strings such as 'predict' and 'explanatory' were used to focus the search on research that could be used for extrapolation.

The primary search was limited to peer-reviewed papers published from 2000 to present and written in English. Studies focused on process-based biogeochemical models and strictly geostatistical models (i.e., interpolating solely based on field measurements) were left out. Papers were required to adhere to the following inclusion criteria:

- Focused on modeling, mapping, or prediction
- Focused on measuring SOC, SOC stocks, soil organic matter (SOM), or Bulk Density

- Focused on covariates, explanatory variables, or auxiliary variables to explain spatial variability of SOC parameters

A full list of search terms is listed in Appendix A. For the final set of literature, we extracted detailed information on study location, year of publication, scale, main variable/s assessed, prediction algorithms/ models, input data type and quantity (e.g., number of samples collected), covariates, accuracy metrics, and performance. Literature that was previously reviewed in the remote sensing section of this project was also omitted from this analysis.

Results: Literature Overview and Statistics

The resulting set of literature consisted of 30 papers, which was reduced to 17 after filtering for those focused on the local or field scale. We further reduced this set by only retaining papers which reported a medium to high modeling accuracy, lowering the final number to 12. Most of the studies focused on estimating SOC via multiple linear regression (MLR), ordinary kriging (OK), random forest (RF), artificial neural network (ANN) and boosted regression trees (BRT) prediction models. The reviewed papers are cited in Appendix B.

- 8 studies focused on SOC; 2 focused on SOM and 1 focused on SOM and Bulk Density
- 8 studies used ML; 5 studies used krigings; and 3 studies used regressions to estimate spatial variability of SOC, SOM, or Bulk Density.

Main Covariates for SOC at the Field Scale

A full list of covariates is detailed in Appendix C. Of the 12 papers that were reviewed at the local scale, the largest concentration of studies was located in China, Germany, and Iran (Table 1). One of the published studies was a review paper, and therefore was not associated with a specific country or region.

Table 1. Study locations of reviewed modeling literature.

Country	# Studies
Germany	2
China	2
Iran	2
Australia	1
Zimbabwe	1
Argentina	1
Brazil	1
Sri Lanka	1

The covariates listed in Appendix C were grouped into six primary categories: Topographic, Land use/ Land Cover (LULC), Hydrologic, Topographic, Climatic, and Pedologic.

Across all these studies, the most **frequently assessed** covariates were:

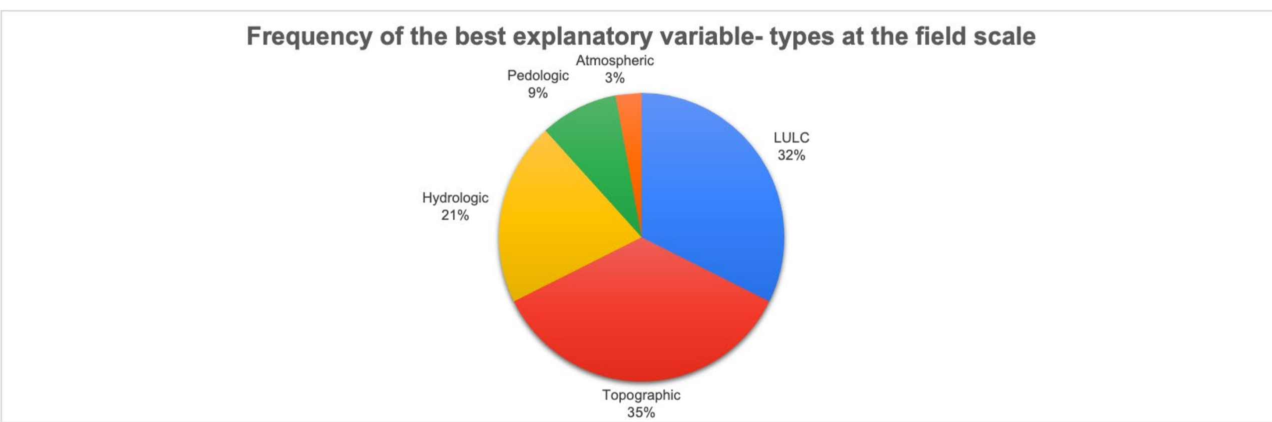
- Climatic Covariates: Temperature and Precipitation
- Hydrologic Covariates: Topographic Wetness Index (TWI) and Stream Power Index (SPI)
- Topographic Covariates: Elevation, Slope, Plan Curvature (Plcu), Profile Curvature and Aspect
- LULC Covariates: spectral bands of satellites (e.g., Landsat), and NDVI.

Given the wide range of research locations, methodologies, and spatial data inputs, comparing models on the basis of quantitative accuracy metrics would have the potential to be misleading. Therefore, in order to assess the performance of the different covariates, those that were most relevant (i.e., with **high explanatory power**) were identified from the results from each study and then the frequency of occurrence was quantified for (A) each individual covariate and (B) the covariate groups.

The most relevant results from this analysis are listed below:

- The *types* of covariates that most frequently showed a high explanatory power at the field scale were the Topographic variables, the ones related to Land Use and/ Land Cover (LULC), and the Hydrologic variables, respectively (Figure 1).
- The *Topographic covariates* that most frequently performed well were elevation and slope, followed by aspect, erosion, terrain ruggedness Index (TRI) and the multi-resolution valley bottom flatness index (MrVBF) (Figure 2).
- The most relevant and frequent covariates related to *LULC* were satellite spectral bands (e.g., Landsat TM), NDVI and land use types, followed by the leaf area index, the tillage erosion pattern, and yield (Figure 2).
- The most relevant and frequent *Hydrologic variables* were the topographic wetness index (TWI), followed by the catchment area and the stream power index (Figure 2).

  NOTE: All the topographic variables assessed in these studies can be derived from Digital Elevation Models (DEMs), as well as most of the hydrologic variables like the TWI.



Our results align with the findings from a review carried out by Minasny et al. (2013)[17], in that at more localized scales, land use and topographic factors appear to increase in importance in predicting SOC.

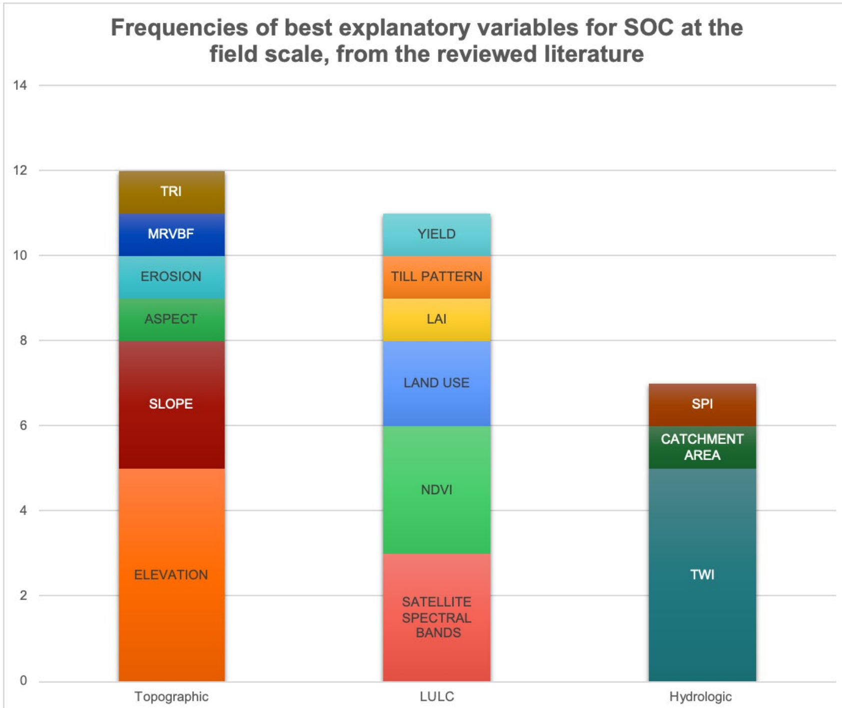


Figure 2. Frequencies of the individual covariates that showed high explanatory power in the assessed studies at the local scale.

Models for Linking SOC to Covariates

Only the modeling approaches that showed moderate to high accuracy (i.e., $R^2 > 0.5$) for estimating SOC at the field scale are listed in Table 3. From the complete list of modeling approaches, the ones that were successful in more than just one use case were Multiple Linear Regression (MLR), Ordinary Kriging (OK), Random Forest (RF), and Artificial Neural Network (ANNS) with a frequency of 3 use cases each, and the Boosted Regression Tree (BRT) and Regression Kriging with a frequency of 2.

When comparing to the results from the review carried out by Lamichhane *et al*. (2019[18]; Fig. 3) there is a clear correspondence in three of the main models that performed best at regional scales, i.e., MLR, RF and NN. Ordinary kriging appears to perform well, but only at the field scale.

[17] Minasny *et al.* (2013)
[18] Lamicchane *et al.*, (2019)

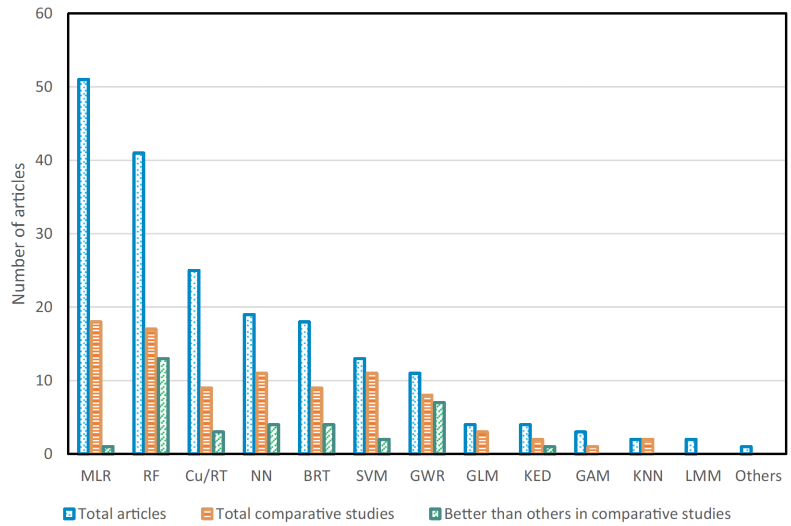


Figure 3. Results from the review carried out by Lamichhane *et al*., (2019), for the regional scale. Bars show the frequency of the use and the comparative performance of algorithms for predicting soil organic carbon (MLR, Multiple Linear Regression; RF, Random Forest; Cu/RT, Cubist/Regression Tree; NN, Neural network; BRT, Boosted Regression Tree; SVM, Support Vector Machine; GWR, Geographically Weighted Regression; KED, Kriging with External Drift; GAM, Generalized Additive Model; KNN, K-Nearest Neighbour; LMM, Linear Mixed Model).

For model validation, no clear preference was displayed for either k-fold cross validation (6 studies) or splitting of data into calibration and validation datasets (5 studies). None of the studies at the local scale evaluated for this report used external independent data sets for cross validation.

Table 3. Best predictive models (medium to high accuracy) for modelling SOC at the local scale, and the frequency of use within the assessed literature.

Predictive Models	Frequency
"mixed model over continuous depth" (MMCD)	1
a 2D kriging with external drift (KED),	1
a 3D kriging with external drift (KED 3D),	1
Artificial Neural Networks (ANNS)	3
Boosted regression tree (BRT)	2
Co-kriging (CK)	1
Cubist model	1
Depth interval-based multiple linear regression	1
General linear model (GLM)	1
Geographically weighted regression (GWR)	1

Linear regression	1
Linear mixed models (LMM)	1
Mixed-scale Geographically weighted regression (mGWR)	1
Multiple Linear Regression (MLR)	3
Ordinary Kriging (OK)	3
Partial Least Square Regression (PLSR)	1
Random Forest (RF)	3
Regression Kriging (RK)	2
Support Vector Machines (SVM)	1
Universal Kriging (UK)	1

Accuracy

The most commonly used metrics for evaluating the performance of different predictive models were the coefficient of determination (R^2), the root-mean-squared-error (RMSE), the mean error (ME) and the mean absolute estimation error (MAEE) (Table 4). As previously mentioned, these metrics were not provided due to the lack of comparability across studies.

Table 4. Most frequent accuracy metrics reported for field scale studies.

Accuracy Metrics	Frequency
Model Efficiency (EF)	1
Mean prediction error (MPE)	1
Mean standardized squared deviation ratio (MSDR)	1
Normalized root mean square error (NRMSE)	1
Root median square error (RMedSE)	1
Root mean square prediction error (RSMPE)	1
Lin's concordance correlation coefficient (LCCC)	2
Mean square error (MSE)	2
Mean absolute estimation error (MAEE)	4
Mean error (ME)	5
Coefficient of determination (R^2)	8
Root Mean Square Error (RMSE)	8

Covariates at a Regional Scale

As established in the review by Lamichhane (2019)[17] of over 120 papers on modeling and prediction of SOC stocks, remote sensing-based spectral bands and indices were found to be useful covariates in locations where vegetation was correlated with SOC levels. At both local and regional scales where diverse terrain features were present, topographic covariates also appeared to improve predictions of topsoil SOC variability.

Additionally, Lamichhane (2019)[17] found that covariates representing organisms/biological activity were among the most frequent of the top five covariates, followed by the variables representing climate and topography (Figure 3). Climate was reported to be influential in determining SOC variation at regional scales, followed by parent materials, topography and land use. However, for mapping at a resolution that represents smaller areas, such as the farm- or plot-scale, land use and vegetation indices were more influential in predicting SOC. Similar results were previously found by Minasny et al. (2013)[17].

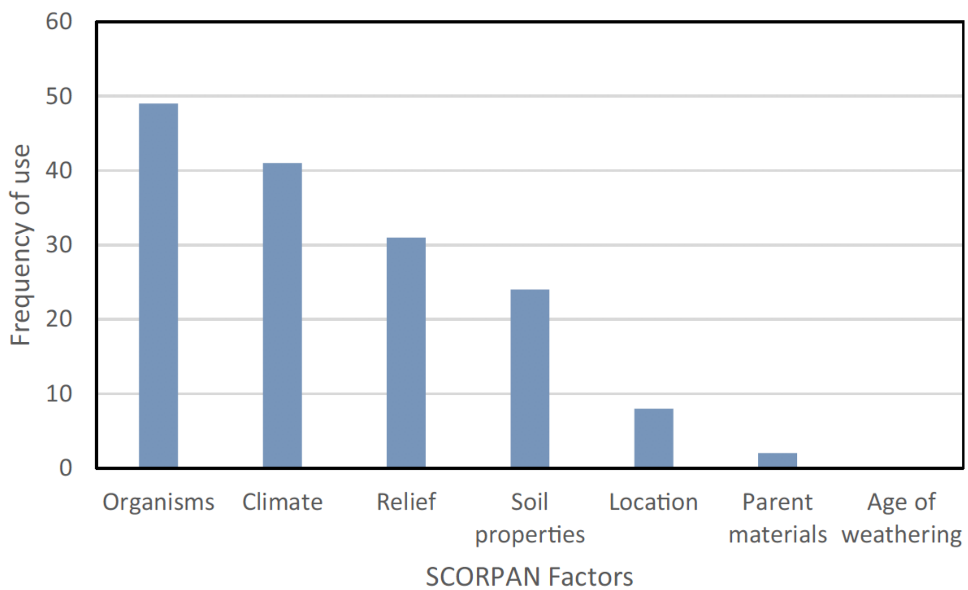


Figure 4. Frequency of the most important environmental predictors assessed based on Random Forest technique. At the regional scale. *Source: Lamichhane et al 2019*[17].

Climates, Models, and Covariates

It was our intention to search for any patterns that could potentially relate covariates and model performance to specific climates or climatic patterns. To accomplish this, the Koppen climates were identified for each study along with previously noted metrics. Unfortunately, there were too few papers to find conclusive patterns, even when narrowing down the Koppen climate types into broader classes (i.e., Temperate, arid/ semi-arid, tropical and boreal), the covariates into main categories, and the models into the three types - Machine Learning, Regression and Kriging.

We acknowledge a remaining need to understand the impact of climate (in addition to other factors such as cropping system) on model performance. In lieu of that, below are some general observations gathered from the existing literature:

- For arid and semi-arid climates, the top prediction algorithms were all Machine Learning models. In contrast, in temperate climates kriging and regression models more frequently succeeded in predicting the variability of SOC at the field scale than ML algorithms.
- In all the climates assessed (arid/semi-arid ,tropical and temperate), Topographic and LULC category covariates performed better as predictors, although the specific covariates varied across climates.

Current Gaps

We have identified two primary gaps within existing literature: 1) a lack of relevant research at different scales and for geospatially modeling BD at the field scale, and 2) the use of field sampling to validate predictive models. The vast majority of the current literature focuses on predicting SOC variability at the regional or country scale. Out of the 120 papers included in the Lamichhane et al., (2019)[17] review, only 12 papers were at the field or parcel scale. Additionally, we were unable to assess any literature on modeling BD at the field scale (while we did find papers at the regional scale, they were not included in this review). With so few studies, drawing comparisons between results, and making site specific recommendations (predictive powers of models, important covariates, relationship with climate and other factors, etc.) is difficult, as mentioned in the sections above. Finally, all of the models still require direct soil samples for model validation, which is, for obvious reasons, impractical for the purposes of farm-scale soil carbon auditing. As more field data are gathered, it is hoped that this need for ground sampling will diminish, enabling modeling approaches to gain broader use and applicability.

Stratification

Context

Research to improve spatial soil sampling strategies extends back to the early 20th century[19]. This field is built on a foundation of classical statistics that emphasized randomization and stratification and was later expanded with the introduction of geostatistical methods aimed at mapping and description of spatial structure. In this section, we limit our analysis to a small subset of this domain focused on quantifying change in SOC stocks for Scope 1 credits and Scope 3 assets (as specified by ESMC).

[19] Lawrence *et al.*, (2020)

Specifically, we define our scope using the following sampling scheme objective outline as described by de Gruijter et al. (2006)[20]:

Objective

- Target universe & domain: the topsoil of the field, farm, or ranch seeking to quantify SOC stock change over the duration of the auditing period
- Target variable: SOC stock (mass/area) across the entire domain at a single point in time. Methods for estimating this variable can be based on modeling or direct sampling.
- Target parameter: Difference in SOC stocks between two points in time over the auditing period.
- Target quantity: SOC stock change for the entire domain over the auditing period.

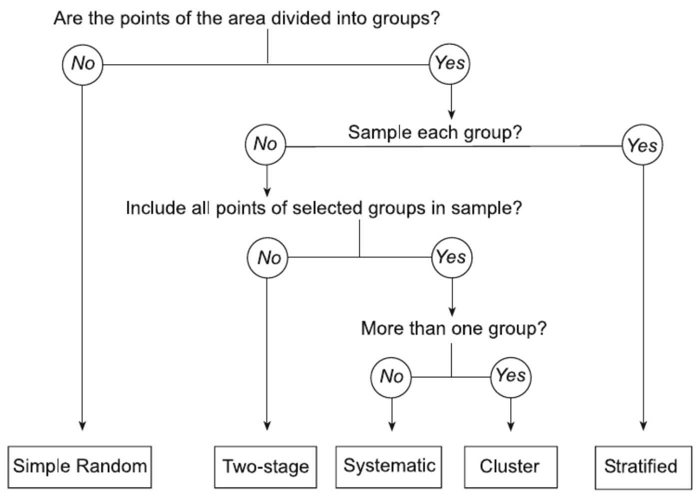
Beyond this basic objective description, it is also important to specify additional requirements that are relevant for Scope 1 credits and Scope 3 assets. First, the sampling scheme should be as objective and free of assumptions as possible. In statistical terms, it should be design-unbiased and correct on average over many repetitions of the same sampling process. In addition, we will have varying levels of prior information on the structure of spatial variation within the domain, and therefore should strive to avoid making assumptions about the actual nature of spatial variation. Following from this, we will also have varying levels of information on the strength of spatial autocorrelation, therefore should plan for some locations to have low levels and some areas to have high levels.

The above requirements all point toward the use of design-based (e.g., stratified random sampling) sampling versus model-based (e.g., grid) sampling[17,21,22,16]. If the objective of sampling included characterization of the model of spatial variation and mapping of the target variable, model-based methods (i.e., purposive/grid sampling) would likely be prescribed. However, objectively quantifying SOC stock change without bias and under differing levels of prior information clearly prescribes design-based sampling. As a result, for the remainder of this section, our analysis will be focused on design-based sampling strategies, of which stratification is a key component.

Design-based Sampling

The two foundational pillars of design-based sampling are randomization and stratification, the latter of which is the focus for this report. Alternative sampling designs such as cluster, systematic, and two-stage sampling are valid for specific objectives, but the need to achieve adequate representation of the entire domain precludes their use (Figure 5).

[20] de Gruijter *et al.*, (2006)
[21] Brus *et al.*, (2011)
[22] Wang *et al.*, (2012)



Stratification Methods

A wide range of stratification methods exists and incorporate ancillary data to varying degrees, all with the objective of minimizing the sampling variance of the spatial mean. On one end of the spectrum, no data are available that are correlated with SOC or BD. In such situations, the domain is typically stratified with the objective of optimizing coverage in geographic space, which is achieved by spatial coverage sampling. Methods in this category include k-means[23] (in this context applied without ancillary data) and balanced acceptance sampling[24], among others. Within the scope of ESMC's intended market area (the US), it is difficult to conceive of situations in which no relevant ancillary data are available.

On the other end of the spectrum, there may be a profusion of ancillary data, which will be correlated with SOC or BD to varying degrees. When these data sources are available, the typical aim of stratification is to adequately sample and divide the feature space covered by the covariates as well as the geographic space. Methods to accomplish this that exclude the costs of sampling include conditioned latin hypercube sampling (cLHS)[25] equal-range stratification[26], and k-means[20] (here using ancillary data) or fuzzy k-means. These function by ingesting ancillary data into a stratification algorithm that optimizes sample placement to arrive at a set of points representative of the multivariate feature or geographic space. For a more thorough review of variations on these methodologies, we refer the reader to Biswas and Zhang (2018)[27].

---

[23] Brus *et al.*, (1999)
[24] Robertson *et al.*, (2013)
[25] Minasny and McBratney, (2006)
[26] Hengl *et al.*, (2003)
[27] Biswas and Zhang, (2018)

Where ancillary data layers serve as the input to the stratification algorithms, each layer is weighted equally and therefore has an equal impact on stratification. Alternatively, a map of target variable predictions can be used, which enables ancillary data to be combined and "weighted" according to each layer's predictive capacity in a linear or non-linear manner. Where ground-collected data on SOC and BD are available to calibrate these "weights" (technically regression coefficients or parameters in a machine learning model), it is preferable to non-weighted stratification.

Optimization of Sample Number

For the application being evaluated here, the goal is to detect a change in SOC stock for the lowest possible cost. For many stratification approaches, optimization of sample number is not implicit and is instead user-defined. In these cases, the optimum sample number must be determined in a separate, preceding step. With more recent methods, optimization of sample size for a target outcome, such as a specific magnitude of SOC stock change, is integrated with stratification and sample allocation steps. Here we explore different methods for optimizing sample numbers to detect changes in SOC stocks and how they are integrated into different stratification methods.

An early example of this type of optimization is the use of the minimum detectable difference (MDD)[28]. When prior data or an estimate of spatial variability is available, the MDD may be inverted and used to estimate the required sample sizes for simple random sampling (Eq. 1):

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{MDD} \right)^2 \tag{1}$$

where n is the estimated sample size, $z_{1-\alpha/2}$ is the z-statistic for the desired precision level alpha and degrees of freedom equal to the input sample size minus one, and σ is the standard deviation. This equation may be extended to stratified random sampling via Neyman allocation[29]. For approaches in which the user is required to define the number of samples allocated, such as cLHS, estimating the appropriate number of samples based on MDD prior to stratification is recommended.

When sampling costs (which can include fixed costs and variable costs such as travel time) are factored into the optimization process, stratification is referred to as "Optimal stratification". Design-based methods that account for cost have been available for a long time[25] (Eq. 2), and are described by the following equation,

[28] Garten and Wullschleger, (1999)
[29] Cochran, W. G. 1977. Sampling Techniques. Wiley, New York.

$$n_h = n \frac{\frac{N_h \sigma_h^2(y)}{\sqrt{\bar{c}_h}}}{\sum_{h=1}^{L} \frac{N_h \sigma_h^2(y)}{\sqrt{\bar{c}_h}}} \qquad (2)$$

where n_h is the sample size in stratum h, N_h is the population size in stratum h, σ_h^2 is the variance in stratum h, and $\underline{c}_h$ is the mean cost-per-point in stratum h. In this equation, the sample size in a stratum is inversely proportional to the mean costs in that stratum. This equation can be simplified if prior information on within-stratum variances is not known.

The sampling algorithm used in the Ospats software package[30] extended the use of the sampling variance under Neyman allocation by introducing information on the uncertainty of the stratification layer. Ospats incorporates rasterized spatial error maps for SOC predictions into the optimization process, directing greater sampling effort to locations where the target variable predictions are less accurate[26]. This algorithm was extended under the name of Ospats+[31] to account for costs (assumed to be equal across strata) and explicitly address SOC stock change auditing for which baseline and follow-up sampling is conducted. Under this new method, farmer profit is the goal of optimization and is realized by utilizing a value of information (VOI) approach. Value accrued to the farmer by increasingly precise estimates of SOC stock changes (associated with increased sampling) is weighed against sampling costs and uncertainty.

Two primary drawbacks to Ospats+ exist. The first arises if the practitioner chooses to use information from the baseline sampling to update the stratification used for follow-up sampling, as recommended in the original method publication[27]. This may be preferable for auditing carbon stocks where site tampering may be an issue[32], but it may substantially increase the sampling requirements over that of paired sampling[33]. However, we note that the choice to re-stratify is not necessarily a requirement for using Ospats+, and the issue of decreased statistical power would manifest under any system where paired sampling is deemed unacceptable.

The second drawback is that Ospats+ assumes the costs of sampling are equal across strata and sampling units. Brus *et al*. (2019)[34] recognized this omission and proposed an alternative stratification methodology that accounted for differential sampling costs (e.g., for hard-to-access terrain) and determined the stratum boundaries via simulated annealing, similar to the method of cLHS. Code

---

[30] de Gruijter *et al.*, (2015)
[31] de Gruijter *et al*, (2016)
[32] Malone *et al.*, (2018)
[33] Lark, (2009)
[34] Brus *et al.*, (2019)

associated with this methodology has yet to be published and it has yet to be integrated with a VOI approach.

Performance

Few studies have examined the relative performance of different stratification methodologies with respect to SOC stocks and SOC stock changes, and only a few published papers have actually implemented recent methods such as Ospats. This paucity of research is probably due to the relatively recent emergence of these methods. Singh et al. (2012)[35] used a gridded approach to collect data on SOC stocks from a single crop field to fit variograms, then used a kriged surface to evaluate the efficacy of different design-based sampling strategies in Australia. Stratification that was based only on geographic coordinates or that incorporated ancillary data only provided a very weak improvement over simple random sampling. However, the use of variograms (with attendant error), and the lack of correlation between ancillary data and SOC stocks may have caused this lack of improvement.

In contrast, Viscarra Rossel and Brus (2018)[36], found that k-means stratification based on proximal sensing and terrain data was 1.2 to 2.1 times more cost-efficient than compositing samples that were collected across geographically compact strata. The stratification layers (landform versus proximal sensing and terrain), and the means of testing stratification efficacy were different, nevertheless the discrepancy illustrates that stratification may be appropriate in some situations but not for all.

Sherpa *et al.* (2016)[37] collected data from an equilateral triangle grid (triangular grids are technically the most efficient[38]) in New York, and compared simple random sampling, stratified random sampling, and systematic sampling. Systematic sampling schemes were optimal for estimating population parameters, experimental variograms, and achieving high accuracy, especially when low sample sizes were used (n=83). Stratified random sampling performed well at moderate-to-high sample sizes (n=160). The stratification was performed across a variety of land uses including forest, cropping and pasture, which may not be representative of the ESMC use case where land under one use is evaluated in isolation.

At the landscape scale, a number of studies have evaluated the efficiency of cLHS[39], but none have done so at the field-scale or with SOC stocks.

One study did evaluate the use of Ospats to estimate SOC stocks in Australia and in New Zealand at the field-scale[28]. It used a method developed by Malone *et al*. (2017)[40] to downscale national-scale SOC stock maps via calibration to proximal sensing data, which subsequently served as the input for Ospats. Due to issues with bias and uncertainty specification, stratification using Ospats performed no better than simple random sampling. In New Zealand, the opposite was true and stratification was superior.

---

[35] Singh *et al.*, (2012)
[36] Viscarra Rossel and Brus, (2018)
[37] Sherpa *et al.*, (2016)
[38] Yfantis et al., (1987)
[39] Yang *et al*., (2020)
[40] Malone *et al*., (2017)

However, the stratification was not a true probability sample (as is the case with cLHS)[41], therefore the conclusion may carry less weight.

[41] Brus *et al.*, (2019))

Conclusion and Recommendations

The use of modeling for prediction and stratification for sampling are both predicated on a solid understanding of the landscape controls that dictate soil carbon accumulation. For both methods of estimating SOC stocks, knowledge of first principles will guide this understanding and help identify the covariates that are likely to be strongly associated with SOC and BD. However, the utility of these covariates is constrained by the availability and quality of representative geospatial data layers that can be used as inputs for modeling or stratification algorithms.

Among the covariates for which geospatial data is commonly available, fine-scale soil taxonomic information should be strongly considered for incorporation strictly based on first principles. Soil map taxonomic units are an integrative representation of a myriad of soil forming factors and are likely to be associated with soil and biotic properties that promote carbon accumulation. In our analysis, performance of layers representing soil taxonomy was lower than that of other covariates, yet this may have been a result of differing data qualities across study areas. DEMs, DEM derivatives (e.g., TWI) and NDVI also tend to be strongly associated with soil carbon accumulation and are globally available at moderate to high resolution. Clay content has the potential to be predictive of SOC stocks and may potentially be approximated by the DEM derivatives. pH is not typically available as a geospatial data layer; however, it is worthy of exploration in future research.

Of these covariates, elevation, slope, NDVI, and TWI all tended to perform well for predicting SOC and SOC stocks at the field scale. Closely related layers such as aspect, catchment area, and combinations of individual satellite spectral bands also improved model performance. The ability for any of these layers to predict SOC is likely to vary by location, so including variables that are closely related to these primary layers should be considered as long as information redundancy and dimensionality reduction are taken into account.

As with the previous report focused on remote sensing, a plethora of potential statistical models and algorithms were used for ingesting spatial data layers and creating predictions. Given the diversity of agricultural systems evaluated, the variety of analysis methods, and the differing accuracy metrics, it is difficult to recommend one model or algorithm over another (although, based on our analysis, MLR, RF, and NN should receive top consideration). Therefore, we continue to recommend automated implementation of multiple models to assess performance within and across agroecosystems.

The same data layers used for modeling SOC, BD, and SOC stocks can also be used as inputs for sample stratification and sample planning. If ancillary data are not available, compact geographic stratification via k-means is advised. Where high quality, relevant ancillary data are available, we recommend Ospats or spatially balanced sampling since costs and uncertainty are explicitly taken into account. This approach is only practical where a high-resolution, reasonably accurate, and unbiased map of SOC stocks is available. In the absence of such a map, cLHS or fuzzy k-means can be used.

To implement cLHS or fuzzy k-means, the coefficients of variation (CV) for SOC and BD in similar agroecosystems should be compared. In some landscapes, SOC is more variable than BD at the field scale[31], and in other landscapes the opposite holds true[42]. In general, if high-quality SOC stock prediction maps are available, then stratification should be performed on the variable with the highest expected CV.

[42] Goidts et al., 2009

Appendix A: Search Terms and Classification

Search Terms

Method / Tool		*Property*		*Variable*
Model* OR "Spatial prediction"	AND	"Soil carbon" OR "Soil organic carbon" OR SOC	AND	explanatory OR covariate

Appendix B: List of Citations for Studies at the Field Scale Resulting in Medium to High Accuracy

Amirian Chakan, Alireza & Taghizadeh-Mehrjardi, Ruhollah & Kerry, Ruth & Kumar, Sandeep & Khordehbin, S.Khordehbin & Yousefi, Shahram. (2017). Spatial 3D distribution of soil organic carbon under different land use types. Environmental Monitoring and Assessment. 189. 10.1007/s10661-017-5830-9.

Castro Franco, Mauricio & Costa, José & Peralta, Nahuel & Aparicio, Virginia. (2015). Prediction of Soil Properties at Farm Scale Using a Model-Based Soil Sampling Scheme and Random Forest. Soil Science. 180. 74-85. 10.1097/SS.0000000000000115.

Dlugoss, Verena & Fiener, Peter & Schneider, Karl. (2010). Layer-Specific Analysis and Spatial Prediction of Soil Organic Carbon Using Terrain Attributes and Erosion Modeling. Soil Science Society of America Journal. 74. 922-935. 10.2136/sssaj2009.0325.

Duarte de Menezes, Michele & Silva, Sérgio & de Mello, Carlos & Owens, Phillip & Curi, Nilton. (2016). Spatial prediction of soil properties in two contrasting physiographic regions in Brazil. Scientia Agricola. 73. 274-285. 10.1590/0103-9016-2015-0071.

George, Justin & Kumar, Suresh & Arya Raj, R. (2018). Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environmental Earth Sciences. 77. 10.1007/s12665-018-7367-9.

Laub, Moritz & Blagodatsky, Sergey & Lang, Rong & Yang, Xueqing & Cadisch, Georg. (2018). A mixed model for landscape soil organic carbon prediction across continuous profile depth in the mountainous subtropics. Geoderma. 330. 10.1016/j.geoderma.2018.05.020.

Mosleh, Zohreh & Salehi, Mohammad & Jafari, Azam & Borujeni, Isa & Mehnatkesh, Abdolmohammad. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental monitoring and assessment. 188. 195. 10.1007/s10661-016-5204-8.

Ratnayake, Renuka & Karunaratne, Senani & Lessels, J. & Yogenthiran, N. & Rajapaksha, R.P.S.K. & Gnanavelrajah, Nalina. (2016). Digital soil mapping of organic carbon concentration in paddy growing soils of Northern Sri Lanka. Geoderma Regional. 7. 10.1016/j.geodrs.2016.03.002.

Van Apeldoorn, Dirk & Kempen, B. & Bartholomeus, Harm & Rusinamhodzi, Leonard & Zingore, Shamie & Sonneveld, M.P.W. & Kok, Kasper & Giller, Ken. (2014). Analysing soil organic C gradients in a smallholder farming village of East Zimbabwe. Geoderma Regional. 2-3. 10.1016/j.geodrs.2014.09.006.

Wang, Bin & Waters, C. & Orgill, Susan & Gray, Jonathan & Cowie, Annette & Clark, Anthony. (2018). High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Science of The Total Environment. 630. 367-378. 10.1016/j.scitotenv.2018.02.204.

Zeng, Canying & Yang, Lin & Zhu, A-Xing & Rossiter, David & Liu, Jing & Liu, Junzhi & Qin, Cheng-Zhi & Wang, Desheng. (2016). Mapping soil organic matter concentration at different scales using a mixed geographically weighted regression method. Geoderma. 281. 10.1016/j.geoderma.2016.06.033.

Appendix C: List of All Assessed Covariates in Studies at the Field Scale

Covariates Assessed	Frequency	Main type
Valley distance/ some distance metric	1	Topographic
Topography (doesn't specify)	1	Topographic
Patterns of tillage erosion	1	LULC
Patterns of water	1	Hydrology
DEM (doesn't specify)	1	Topographic
Channel network base level	1	Hydrology
Convergence index	1	Hydrology
Valley Index	1	Topographic
Terrain characterization index (TCI)	1	Topographic
Relative Position Index	1	Topographic
Landscape position (toeslope, backslope, etc.)	1	Topographic
Hillside	1	Topographic
Soil loss factor	1	Pedologic
Flow accumulation	1	Hydrologic
Vertical distance to channel	1	Hydrologic
Multi-resolution of ridge top flatness index (MrRTF)	1	Topographic
TRI	1	Topographic
Height above nearest drainage	1	Hydrologic
Distance to nearest drainage	1	Hydrologic
Modified catchment area	1	Hydrologic
TaC tangential curvature	1	Topographic
Yield	1	LULC
Vegetation cover	1	LULC
Perpendicular vegetation index (PVI)	1	LULC
Enhanced Veg Index (EVI)	1	LULC
Land use	2	LULC
Electrical conductivity	1	Pedologic
Clay Index, clay content	1	Pedologic
Carbonate index	1	Pedologic
Parent material	1	Pedologic
Natural gamma emission of potassium	1	Pedologic
% Silt	1	Pedologic
Number of cattle	1	LULC
Soil composition (silica, illite, Kaolin, smectite, uranium, etc.)	1	Pedologic
LS factor	1	Topographic
Horizon depth	1	Other
Total erosion	2	Topographic
Catchment area	2	Hydrologic
Catchment index	2	Hydrologic
Flow path length	2	Hydrologic
Length-slope factor	2	Topographic
TPI Topographic position index	2	Topographic

Multi-resolution Valley Bottom Flatness Index (MrVBF)	2	Topographic
Temperature	3	Climatic
Precipitation	3	Climatic
Stream power index (SPI)	4	Hydrologic
SAGA wetness index (WI)	4	Hydrologic
Wetness index	4	Hydrologic
Topographic wetness index (TWI)	4	Hydrologic
Aspect	5	Topographic
Profile curvature	5	Topographic
NDVI	5	LULC
Landsat tm bands, SFC seasonal fractional cover data	5	LULC
Plan curvature (Plcu)	6	Topographic
Elevation	8	Topographic
Slope	8	Topographic