How to estimate soil organic carbon stocks of agricultural fields? Perspectives using ex-ante evaluation

Eric Potash[1,2], Kaiyu Guan[1,2,3], Andrew Margenot[1,4], DoKyoung Lee[1,4], Evan DeLucia[1,5], Sheng Wang[1,2], Chunhwa Jang[1,4]*

1. Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
2. Department of Natural Resource and Environmental Sciences, College of Agricultural, Consumer and Environmental Sciences, University of Illinois at Urbana Champaign, Urbana, IL 61801, USA
3. National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
4. Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
5. Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

**Corresponding author: Eric Potash (epotash@illinois.edu)*

Abstract

Estimating soil organic carbon (SOC) stocks of agricultural fields has a range of important applications from development of sustainable management practices to monitoring carbon stocks. There are many estimation strategies with the potential for more reliable estimates of SOC stock and more efficient use of soil sampling and analysis resources, especially by leveraging readily available auxiliary information such as remote sensing. However, concrete guidance for strategy selection is lacking. This study narrows this gap with a comparison of strategies for estimating deep SOC stock (0-60cm) in a prototypical field. Using high density

SOC stock measurements and simulation, we built on past studies by 1) ex-ante evaluating a large number of strategy options, 2) using a Bayesian approach to quantify the uncertainty of the comparison, and 3) considering multiple Bayesian models to assess sensitivity to this modeling choice. We found that, using readily available auxiliary information, both balanced and stratified sampling offer substantial improvements over simple random sampling. The auxiliary information most important for this improvement is a Sentinel-2 SOC index = *blue / (green x red)*, followed by the topographic wetness index. We found that these results are robust to the choice of mapping method, but that there is uncertainty in the magnitude of improvement. We recommend future studies implement this Bayesian approach for simulated ex-ante evaluation of SOC stock estimation strategies across more fields to investigate the generalizability of these findings.

Keywords: soil carbon stocks; sampling; estimation; evaluation; geostatistics; Bayesian

1. Introduction

Estimating soil organic carbon (SOC) stock in agriculturally managed soils at the field scale has a range of important applications from development of sustainable management practices to monitoring carbon stocks. Such an estimation strategy entails two statistical steps: (1) a sampling design selects locations at which to take measurements, and (2) an estimator combines those sample measurements to estimate mean SOC stock across the field. Which strategy should we use? In this study we focus on probability-based sampling designs (e.g. stratified sampling) with design-unbiased estimators (e.g. inverse probability-weighted mean) since these are preferred for spatial mean estimation (Brus and de Gruijter, 1997; Brus, 2021) and are required by various

SOC stock monitoring protocols (Oldfield et al., 2021). The baseline estimation strategy is simple random sampling with the sample mean estimator.

Stratified sampling, i.e. dividing the field into areas of similar characteristics, is often recommended because it can lead to more efficient estimation of mean SOC stock (de Gruijter et al., 2006; Oldfield et al., 2021). However, several choices must be made to design a stratification, e.g. which variables to stratify and into how many strata. Guidance for these choices remains qualitative and quantitative evidence for the benefits of stratified sampling and how these benefits might depend on these choices is lacking (Oldfield et al., 2021). A promising probability sampling design that also takes advantage of auxiliary information is balanced sampling (Deville and Tille, 2004). Balanced sampling does not require designing an intermediate stratification which can both improve performance and reduce the number of choices requiring guidance.

However, there is a knowledge gap about the performance of these strategies for estimating mean SOC stock in agricultural fields. Recently, Lawrence et al. (2020) identified just one study (Mallarino and Wittry, 2004) evaluating stratified sampling for estimating mean soil organic matter (SOM) in agricultural fields, and zero studies for mean SOC stock. De Gruijter et al. (2016) validated a stratified sampling design for mean SOC stock. However, because their study site was a 2083 ha farm and their stratification relied on a previous SOC stock evaluation with soil sampling, their findings are not directly relevant to us. Another study validating strategies for mean SOC stock estimation is Brus (2015), though it was at the district level in Ethiopia. Altogether, data on the performance of strategies to estimate mean SOC stock in agricultural fields is lacking.

To fill this gap, we need to evaluate these estimation strategies in agricultural fields by estimating and comparing their performance. One conventional approach to evaluating

estimation strategies is to implement each one in the field and estimate its performance *ex post* using variance formulas. A potentially more versatile and efficient method evaluates performance *ex ante* using simulation. First, a field is intensively sampled to create an SOC stock map. Then different estimation strategies are simulated against the map and their estimates compared with the map's mean SOC stock (Figure 1). Uncertainty in the SOC stock map can be incorporated by repeating this process using many such maps.

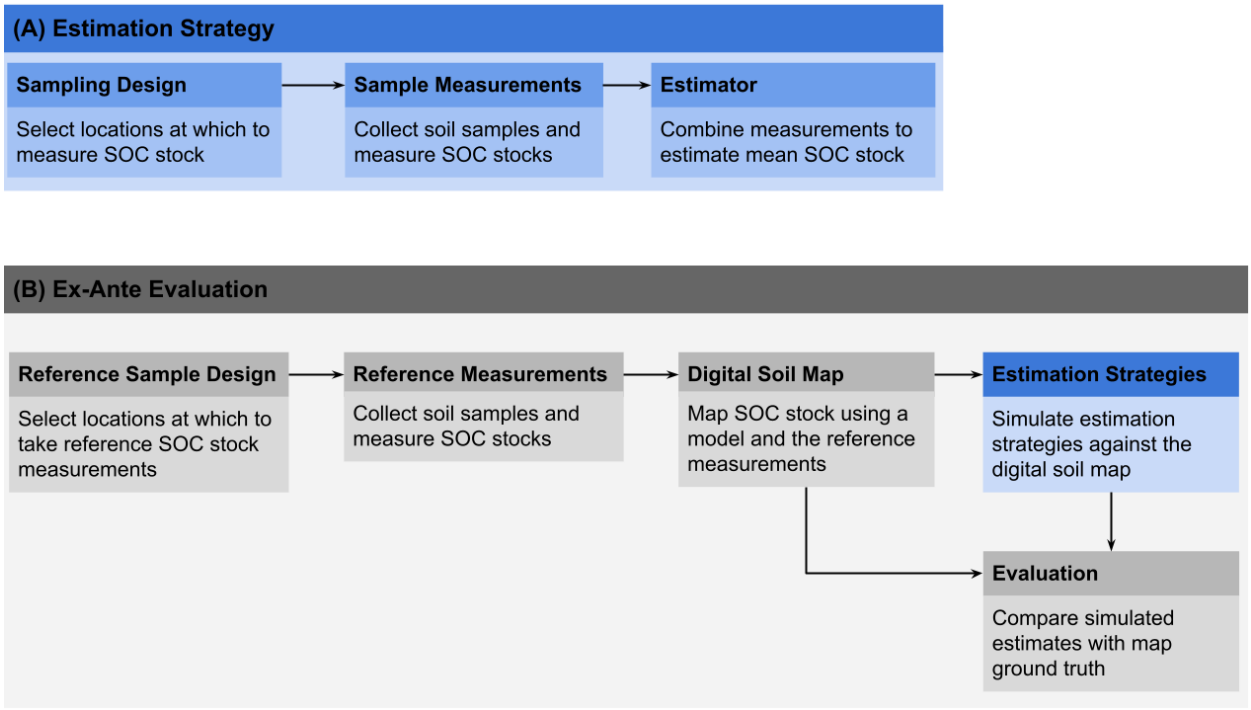


Figure 1: Flowcharts of (A) mean SOC stock estimation strategy and (B) ex-ante evaluation of these strategies.

While ex-ante evaluation using simulation has proved a useful tool for evaluating mean SOC stock estimation strategies, most applications have ignored two important technical considerations. The first consideration is propagating uncertainty in the reference map through the evaluation procedure to quantify uncertainty in the performance of quantification strategies

and their comparison. This consideration has been previously addressed in the context of estimating mean nitrate content by using Bayesian methods (Hofman and Brus, 2021). The second consideration is the sensitivity of the evaluation to the predictive mapping method used to generate the map. To address this consideration, our approach expands on Hofman and Brus (2021) by employing and comparing both geostatistical and machine learning methods for predictive mapping of SOC stock.

The objective of this study is to demonstrate the use of ex-ante evaluation to compare different estimation strategies (simple random sampling, stratified sampling, and balanced sampling) in a prototypical agricultural field to fill the above knowledge gap and address the technical considerations. Specifically, we aim to answer the following two questions: (1) Which estimation strategy would perform best and which auxiliary information is most beneficial? (2) How much uncertainty and sensitivity is there in the evaluation? We draw on high-density soil sampling and SOC stock measurement at a commercial field in central Illinois to address these questions. Importantly, we estimate deep (0 – 60 cm) SOC stocks because of evidence that lower depths play an important role in SOC stock dynamics (Tautges et al., 2019). We discuss how future studies can build on our evaluation results to develop a knowledge base for guiding efforts to estimate mean SOC stock in agricultural fields.

2. Review of estimation strategies and evaluation methods

In this section we review strategies for estimating mean SOC stock (section 2.1) and methods for evaluating these strategies (section 2.2). Compared to other reviews of these topics (e.g., de Gruijter et al., 2006), ours has two distinctions. First, while de Gruijter et al. (2006) refer to the combined stages of a sampling design and estimator as a sampling strategy, we prefer the term

estimation strategy to emphasize that the sampling design does not completely determine the estimator. Our review highlights these as discrete choices. Second, we devote significant attention to what we call ex-ante evaluation. This method for evaluating estimation strategies has not to our knowledge received a careful review in the soil science literature, nor a direct comparison to the traditional alternative which we term ex-post evaluation.

2.1 Estimation strategies

We define an estimation strategy as the combination of two statistical steps: a sampling design and an estimator.

2.1.1 Sampling designs

A probability sampling design is one in which each point in the study area has a known and non-zero probability of being selected for measurement. Probability sampling has the benefit of supporting robust estimation of the population mean (i.e. mean SOC stock) as described in the next subsection. For regulatory applications, an auxiliary benefit of randomized sampling locations is mitigation of fraud (de Gruijter et al., 2016; Lawrence et al. 2020). We consider three probability sampling designs: simple random sampling (SRS), stratified sampling, and balanced sampling. SRS serves as our baseline.

Stratified and balanced sampling have the potential to improve on SRS by incorporating auxiliary information (covariates) such as topography and remote sensing into the selection of sample locations. In addition to choosing which auxiliary information to include, stratified sampling requires several further choices including: rescaling these covariates to make them comparable, an allocation of samples among the strata, and the number of strata (de Gruijter et al., 2006). While the traditional k-means approach to constructing a stratification only supports

continuous covariates, there are other clustering algorithms that accommodate categorical covariates (Huang, 1998).

Balanced sampling (Deville and Tillé, 2004; Brus, 2015) selects samples that are representative in the sense that the (inverse probability weighted) mean value of a covariate (e.g. slope) at the sample locations is equal to the mean value across the field. Balanced sampling has several advantages over stratified sampling. First, it can naturally incorporate categorical covariates. Second, we need not make the somewhat arbitrary choices listed above for constructing a stratification (Grafström and Schelin, 2014). One disadvantage of balanced samping is it may lead to less robust uncertainty quantification than simple or stratified sampling (see next section).

2.1.2 Estimators

Probability sampling designs yield a natural unbiased estimate of mean SOC stock, called the Horvitz-Thompson (HT) estimator in its most general formulation, which averages the measurements weighting each by the inverse of probability of inclusion in the sample. In the case of SRS and stratified sampling, the HT estimator is the usual sample mean and weighted sample mean, respectively. The HT estimator is design-unbiased so that the average estimate across many random samples of a given design is equal to the true mean SOC stock.

One disadvantage of the HT estimator is that it does not take into account auxiliary information beyond what was used to inform the sampling design. Most monitoring protocols require estimators to be design-unbiased, so that model-based estimators accounting for auxiliary information are not permitted. An alternative to this is the so-called model-assisted estimators (Brus, 2000).

In addition to providing a point (i.e. single-number) estimate of mean SOC stock, it is

important for both scientific and regulatory applications to quantify the uncertainty of this estimate via a confidence interval (CI). For the probability design-based strategies we are considering, CIs are constructed by estimating the variance of the estimator and then assuming a normal or Student-t distribution to calculate the CI. For simple and stratified sampling, this assumption is justified by the central limit theorem and variance estimation is design-unbiased. However, for balanced sampling it is not possible to have a design-based unbiased variance estimate (Grafström and Schelin, 2014) and so uncertainty quantification may be less robust.

2.1.3 Measures of estimation strategy performance

There are several ways of quantifying the performance of an estimation strategy. For a given point estimate of mean SOC stock, the error is commonly quantified in terms of squared error, absolute error, and relative error. Since probability sampling designs are randomized, the estimate is also random and so are these error quantities. Thus, each estimation strategy has a corresponding *distribution* of squared error, relative error, etc. These are commonly summarized using a single number, e.g. mean squared error is the mean (or expected) squared error across many random samples.

In this study our primary performance measure is the 95th percentile of the relative error distribution, which we simply call the *relative error bound* because with high probability (95%), the relative error of the estimate will be less than (bounded by) this number. This is a version of expanded measurement uncertainty as defined by ISO Guide 98 (ISO, 2009) to be “a quantity defining an interval about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand” (see also Hofman and Brus, 2021).

We can also measure the performance of the estimated CI. A simple but important

measure of CI performance is coverage, i.e. the proportion of CIs which contain the true value. Ideally the coverage of the 95% CI is 95%. Another measure of CI performance that is common in SOC stock monitoring protocols (Oldfield et al., 2021) is the relative width of the 95% CI. We note that when the point and variance estimates are design-unbiased (see section 2.1.2) then the expected relative width of the 95% CI is equal to the relative error bound.[1] We prefer the relative error bound measure because its meaning does not rely on the design-unbiasedness of the estimate nor its variance.

2.2 Evaluation methods

Here we review two different approaches to validating estimation strategies, i.e. measuring and comparing their performance. The conventional approach is ex-post evaluation , in which each strategy is implemented in the field to estimate its performance. In this study we opt for ex-ante evaluation, in which SOC stock maps are created and then different strategies are simulated against these maps.

2.2.1 Ex-post evaluation

One way to evaluate an estimation strategy is by implementing it and estimating its estimation variance. There are standard formulas for estimating the variance of SRS and stratified sampling (de Gruijter et al., 2006). Moreover, after stratified sampling we can estimate the precision that would have been obtained with SRS using the law of total variance (equation 7.16 of de Gruijter et al., 2006). These formulas for simple and stratified sampling are expected to be quite robust (owing to the central limit theorem) for large sample sizes. For example, de Gruijter et al. (2016)

[1] In this case they are both equal to $t_{0.975}\sigma/SOC$ where $t_{0.975}$ is the 0.975 quantile of the Student-t distribution with $n-1$ degrees of freedom, n is the sample size, σ is the standard deviation of the sampling distribution of the estimator, and SOC is the mean SOC stock.

quantified SOC stocks in surface soils (0–7.5cm depth) in vertisols and alfisols across the 2083 ha University of Sydney Holtsbaum Agricultural Research ("Nowley") Farm in Australia (Stockmann et al. 2016) and estimated that their stratification had a standard error of 0.62 Mg ha^{-1}. Using the law of total variance, they estimated that SRS would have a standard error of 0.87 Mg ha^{-1}, meaning that the stratification improved the relative error of the estimation by 29%.

However, this ex-post approach to evaluation of simple and stratified estimation strategies has three important limitations. First, the maximum number of strategies that can be compared by implementing a single sampling design is two (e.g., the previous example): (1) implementing a stratified design and (2) comparing it to SRS. We are unable to compare the implemented stratification with alternatives arising from different auxiliary data or even the same auxiliary data but some different stratifications (e.g. a different number of k-means clusters). Second, ex-post evaluation does not fully apply to strategies besides SRS and stratified sampling. The formulas for estimating the variance of balanced sampling estimates are not as firmly grounded as those for simple or stratified sampling so we do not wish to rely on them for evaluation. Moreover, we are primarily interested in the relative error bound, which is only directly related to the estimator variance for normally distributed estimators. Third, ex-post evaluation does not quantify the uncertainty of the performance estimate or any comparison. For example, the standard error of 0.62 Mg ha^{-1} estimated for the stratification of de Gruijter et al. (2016) is not accompanied by an uncertainty interval. One could be constructed by assuming a normal distribution of SOC stock within each stratum and constructing CIs on the chi-squared distribution. This would give a very wide 95% CI of 0.37 to 1.78 Mg ha^{-1}. However, unlike the normality assumption used to justify the variance estimate itself, which is supported by the central limit theorem, a normality assumption here is less plausible.

2.2.2 Ex-ante evaluation

An alternative to ex-post evaluation is ex-ante evaluation in which an estimation strategy is simulated rather than implemented. If we had knowledge of the SOC stock at every location in the field then we could simulate an estimation strategy by repeatedly generating random locations according to the sampling design, looking up the corresponding SOC stocks, and evaluating the estimator. Since we cannot (with current measurement technology) measure SOC stock at every location in the field, we approximate it using a digital SOC stock map.

The fidelity of the SOC stock map is essential to the validity of ex-ante evaluation so we review several approaches to digital soil mapping and their consequences for ex-ante evaluation. One approach is to measure SOC stock at each pixel of a map. This was the approach of Mallarino and Wittry (2004), who used 0.2 ha pixel maps to ex-ante evaluate mean SOM estimation strategies in eight Iowa, USA fields. In each pixel they randomly selected an 80 m^2 subplot from which they collected 20-24 vertical cores to a depth of 15 cm, which they composited and analyzed using the Walkley-Black method. The major limitation of this approach to digital soil mapping is that it does not capture any variability within each pixel, e.g. in this case on a scale less than 45 m.

Short range variation can be incorporated into the SOC stock map using geostatistical simulation (Chilès and Delfiner, 2012). For example, Brus (2015) used a random forest model to generate their map from SOC stock measurements at convenient sample locations in three districts of Ethiopia. Importantly, independent predictions of SOC stock at each point in the field produced an SOC stock map with unrealistically low variability, and so normally distributed noise was added. An important limitation of both of these simulation approaches is that they do not account for uncertainty in the underlying measurements or predictions.

Uncertainty in the SOC stock map can be incorporated into ex-ante evaluation by

using a Bayesian model, as shown by Hofman and Brus (2021) in the context of nitrate estimation strategies. Instead of generating a single digital soil map, many maps are drawn from the posterior distribution of the Bayesian model. This collection of maps captures the uncertainty in the SOC stock map according to the model. For each map, we perform ex-ante evaluation of the estimation strategies under consideration. The result is that for each map we have a measure of estimation performance, e.g. relative error bound (section 2.3.3). Combining the maps, we obtain samples from the posterior distribution for the performance measure. These samples express our uncertainty in the performance of the estimation strategy due to our uncertainty in the SOC stock map. We can then summarize this distribution in various ways (e.g. the median and 95% CI).

The Bayesian approach allows us to quantify the uncertainty in the performance measures of the estimation strategies. However, the uncertainty is limited to the scope of the model. For example, if we model the relationship between SOC stock and topographic wetness index (TWI) as linear, the Bayesian approach only captures our uncertainty in the slope of this linear relationship, not in the possibility that the relationship is non-linear. In other words, the uncertainty may be mis-stated because the model is wrong. In order to investigate the sensitivity of our results to this latter possibility, we suggest simply performing ex-ante evaluation with multiple Bayesian models.

3. Materials and methods

3.1 Study site

The Bondville site is a 34 ha field located in Champaign county, in central Illinois, USA. The

field is mapped as five closely related soil series classified as Mollisols (USDA Soil Taxonomy) with textural classes of silt loam to silty clay loam (Figure 2). According to SSURGO, the A horizon depths of these soils generally range 20 to 36cm, with a pH 5.6–7.4, and SOC stock in the 0–50cm profile of 98 (Wyanet) - 195 (Drummer) Mg ha^{-1} (Soil Survey Staff).

The field at Bondville has been managed using a soybean-maize rotation cropping system for 12+ years with no-till after soybean and conservation tillage after maize. This is a rain-fed agricultural system. In 2020, soybean was planted and fertilized with 168 kg-N ha^{-1} as monoammonium phosphate 11-52-0 fertilizer, 168 kg-K ha^{-1} as potassium chloride (0-0-60), and 4.5 t ha^{-1} of soft lime before planting. Weeds were controlled using herbicides according to regional recommendations (Illinois Agronomy Handbook, 2017). There was a heavy presence of tall fescue (*Festuca* sp.) at the edges of the site and the grassed waterway in the middle of the field. The average monthly precipitation, maximum and minimum temperatures for the nearby (7 km) Champaign-Urbana Willard Airport station (USW00094870) in 2020 were 7.7cm, 17.3 and 5.8 °C, respectively (NOAA). Growing-season precipitation (April-Sept) in 2020 was 10.1 cm.

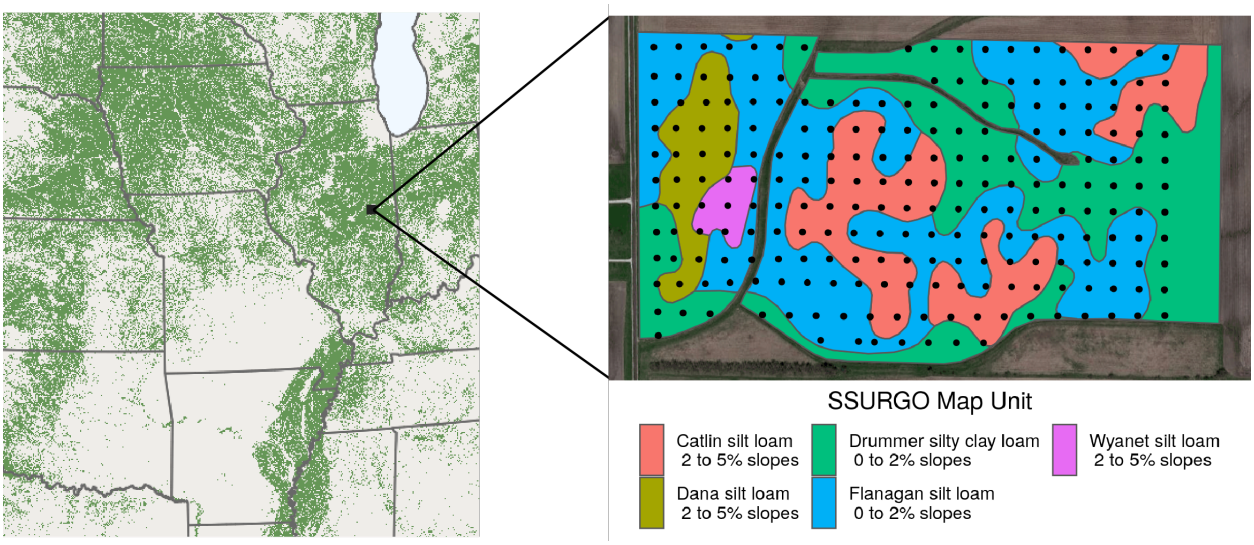


Figure 2: Regional map (left) of North Central USA showing cropland (USDA 2020 National Cultivated Layer) in green and the Bondville site location and field map (right) with 225 sample locations (black dots) used for SOC stock measurement, overlying and soil map units (SSURGO).

3.2 SOC stocks and auxiliary data

A set of reference SOC stock measurements were made as follows. On April 22, 2020, vertical core samples were taken to a depth of at least 60 cm (and up to 100 cm) using a Giddings probe (Giddings machine company: Winsor, CO) mounted on an all terrain vehicle. Soils were sampled on a 35 m x 35 m grid, yielding 225 sampling locations within the cultivated area (Figure 2). The cores were split into 0-30cm and 30-60cm depths and homogenized by hand crumbling. Gravimetric water content was measured by drying 5-7g of subsample at 100 °C for 24 hours. The bulk density (BD) was obtained from dry weight of soil from each section (g) over volume of the segmented portion of the soil core (cm^3). The samples were prepared and sent to a third-party lab to measure total soil carbon concentration by dry combustion method in a LECO CN828 . For soils with pH > 7.2, inorganic carbon was estimated gravimetrically after addition

of 1% HCl (Walthert et al., 2010).

We collected the following auxiliary information (covariates) for the site (Figure 3). From SSURGO we collected the map units and the gSSURGO estimate of 0-60 cm SOC stock for each map unit (obtained by linear interpolation of the 0-50cm and 50-100cm estimates). From the National Elevation Dataset we collected elevation from which we derived three topographic covariates: slope, aspect, andTWI. We used northing and easting geographic coordinates (measured in meters from the SW corner of the site). Finally, we used an SOC Index (SOCI) defined as *blue / (green x red)* (Thaler et al., 2019). We computed the index from a Sentinel-2 image retrieved on February 11, 2020, the most recent cloud free image available prior to planting. All auxiliary information was stored on a 10 m x 10 m raster grid (100 pixels ha^{-1}) so that the cultivated area of the field contained 3,085 pixels. These covariates were chosen because of their 1) potential to predict SOC, 2) recommendation in SOC monitoring protocol guidance (Oldfield et al., 2021), and 3) availability in public databases for every point of the field, a requirement for those sampling designs (stratified and balanced) and estimation methods (model-assisted) that use covariates (section 2.1.2).

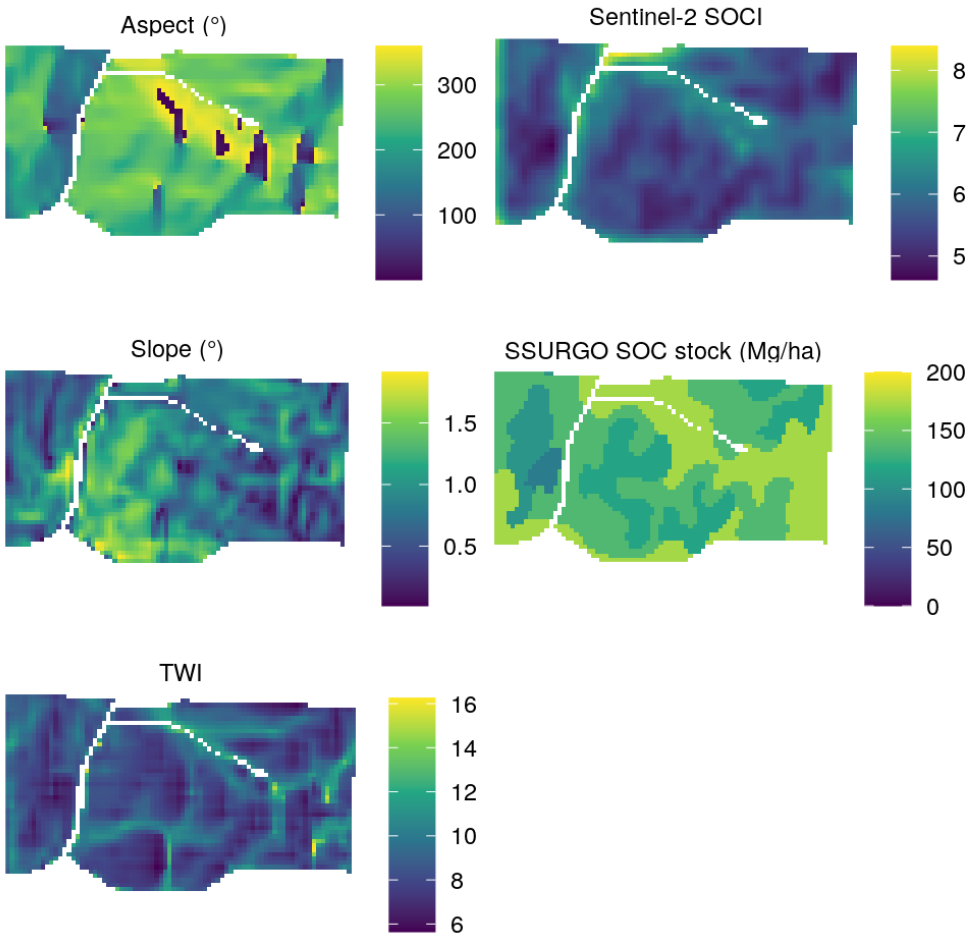


Figure 3: Spatial patterns of covariates at Bondville site used for SOC stock estimation. Abbreviations: Topographic wetness index (TWI), soil organic carbon index (SOCI), Soil Survey Geographic database (SSURGO).

3.3 Evaluation Methods

We evaluated the three sampling designs (simple random sampling, stratified sampling, and balanced sampling) using ex-ante evaluation (section 2.2.2). For this purpose the study site was represented by the points at the centers of the raster pixels described in section 3.2, i.e. 3,085 points on a 10 m x 10 m grid. A Bayesian model of SOC stock (described next) was used to simulate 200 SOC stock maps. For each sampling design and sample size (15, 20, 25, …, 50) we

generated 200 samples. Each of the 200 x 200 combinations of an SOC stock map and sample led to a point estimate and CI for mean SOC stock.

The relative error was calculated for each of these estimates relative to the mean SOC stock of the corresponding map. For each map and sample design and sample size there were thus 200 relative errors, one for each sample. The relative error bound (see section 2.1.3) for this map was then calculated as the 95th percentile of these 200 values. There is thus a relative error bound for each of the 200 posterior maps. For each map and sample the estimated CI either does or does not cover the true mean SOC stock. For each map, CI coverage is calculated as the proportion of the 200 estimated CIs (one for each sample) that covers the true mean SOC stock.

Our primary model of SOC stock was kriging with external drift (KED), also known as universal kriging or regression kriging (Pebesma, 2006). Because the SSURGO SOC stock is constant within each map unit, including both the map units and SSURGO SOC stock would lead to a singular regression design matrix, so we omitted the map unit in the KED model. We used an exponential variogram for the KED model. After standardizing the outcome and covariates, the following non-informative prior distributions were put on the egression coefficients β, variogram scale α, variogram nugget σ, and variogram range ρ:

$\beta \sim Normal(0, 2.5)$, $\alpha \sim Exponential(1), \sigma \sim Exponential(1)$, $\rho \sim Uniform(a, b)$, where a=22 m and b=788 m are the minimum and maximum distances, respectively, between sample points in the reference SOC stock design (Figure 2). The KED model was fit using the Markov Chain Monte Carlo software Stan (Carpenter et al., 2017). We generated 4 chains with 1000 iterations each, saving the last 500 to produce 2000 samples from the joint posterior parameter distribution. We assessed mixing using the criteria $\hat{R}<1.05$ and n_{eff} /N > .001 where $\hat{R}$ is the Gelman-Rubin convergence statistic and n_{eff} is the effective sample size (Gelman et al., 2013).

For the purpose of sensitivity analysis, we considered an alternative SOC stock model

using Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). BART was chosen because, compared to KED, its modeling approach is substantially different which is desirable for the sensitivity analysis. The BART model consists of an ensemble of regression trees which are non-linear compared to the linear regression term of KED. The trees are constrained to be weak learners but unlike related machine learning methods such as Gradient Boosted Trees (Hastie et al., 2009), this is accomplished using a prior and likelihood to obtain a Bayesian statistical model. Unlike KED, which uses a spatially autocorrelated error term, BART has a spatially independent Gaussian error term. We included all of the available covariates (section 3.2) in the BART model and fit the model in R using the `BART` package (Sparapani et al., 2021).

For both KED and BART models, we included a measurement error term. Following Hofman and Brus (2021), we assumed a normally distributed measurement error informed by the prior literature. Specifically, we assumed a measurement standard deviation of 0.15 g cm^{-3} for bulk density and 0.16% for SOC concentration. We assumed these errors were independent so that the measurement standard deviation for SOC stock was 1.44 Mg ha^{-1}. These errors were incorporated into the simulations by subtracting the corresponding variance from the nugget of the KED model and the Gaussian error of the BART model.

For stratified designs we used the standard k-means clustering algorithm (de Gruijter et al., 2006). As mentioned above (section 2.1.1) this does not accommodate categorical covariates so the SSURGO map unit was not included. For rescaling, our default method was z-score standardization, though we also considered percentile rank and min-max. In the absence of prior information on the variability of SOC stock within each stratum, we used proportional allocation of samples (de Gruijter et al., 2006). Since uncertainty quantification is essential, each stratum must have at least 2 samples. Thus the number of strata was set such that, under proportional allocation, each stratum received at least two samples. We also considered this with 3, 4, or 5

samples per stratum. For balanced sampling we included all covariates and generated samples in R using the `BalancedSampling` package (Grafström and Lisic, 2019). We also considered a model-assisted estimator in conjunction with SRS using generalized regression in the `mase` R package (McConville et al., 2018).

4. Results

The mean SOC stock at the 225 locations was 101.8 Mg ha^{-1} with a standard deviation of 26.0 Mg ha^{-1} (Figure S1). Before fitting models (section 4.1), we examined the relationship between these measurements and each of the covariates (Figure 4). We found stronger linear relationships between measured SOC stock and SOCI (R^2=0.31), TWI (R^2=0.21), SSURGO Map Unit (R^2=0.17), and SSURGO SOC stock (R^2=0.16).

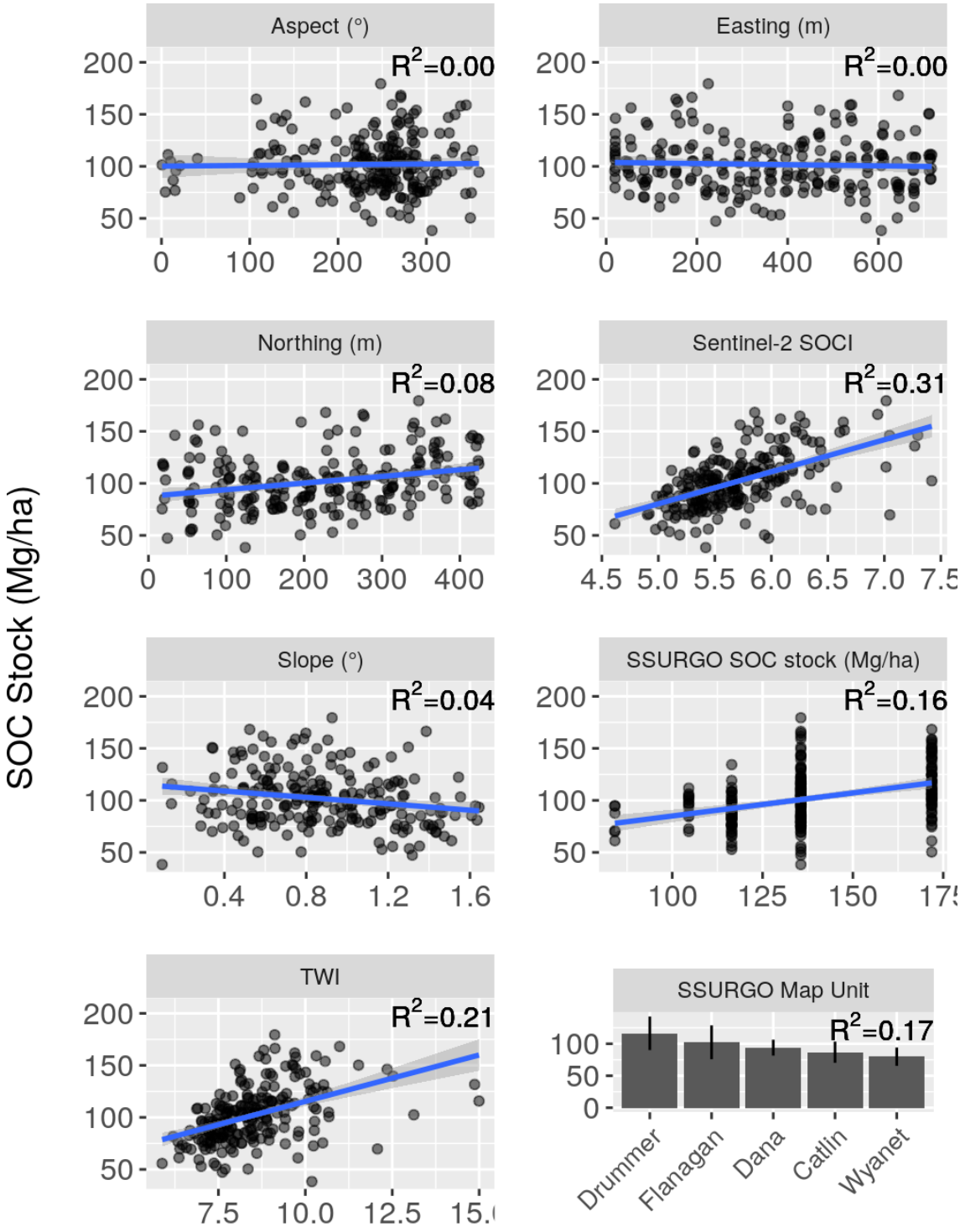


Figure 4: Relationship between SOC stock and each covariate. Abbreviations: topographic wetness index (TWI), soil organic carbon (SOC), SOC index (SOCI), Soil Survey Geographic database (SSURGO).

4.1 Bayesian SOC maps

The Bayesian KED model was fit to the 225 measurements and their associated covariates. Both TWI and SOCI had significant relationships with SOC stock in the model (Figure 5). The estimated spatial autocorrelation structure has a posterior nugget-to-sill ratio 0.83 (95% CI 0.39 to 1.0) and range 430 m (95% CI 53 to 768). The median Bayesian R^2 (Gelman et al., 2019) of the KED model was 0.46. Note that we used a linear KED model as opposed to log-linear because the linear model outperformed the log-linear model in terms of mean absolute error under 10-fold cross validation (Figure S2).

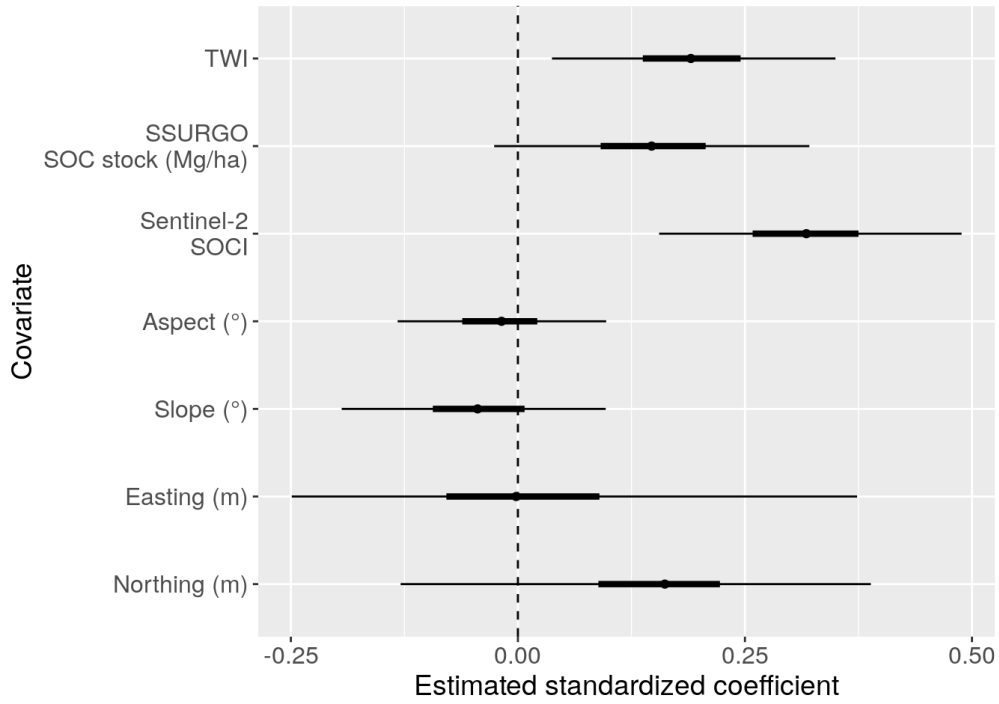
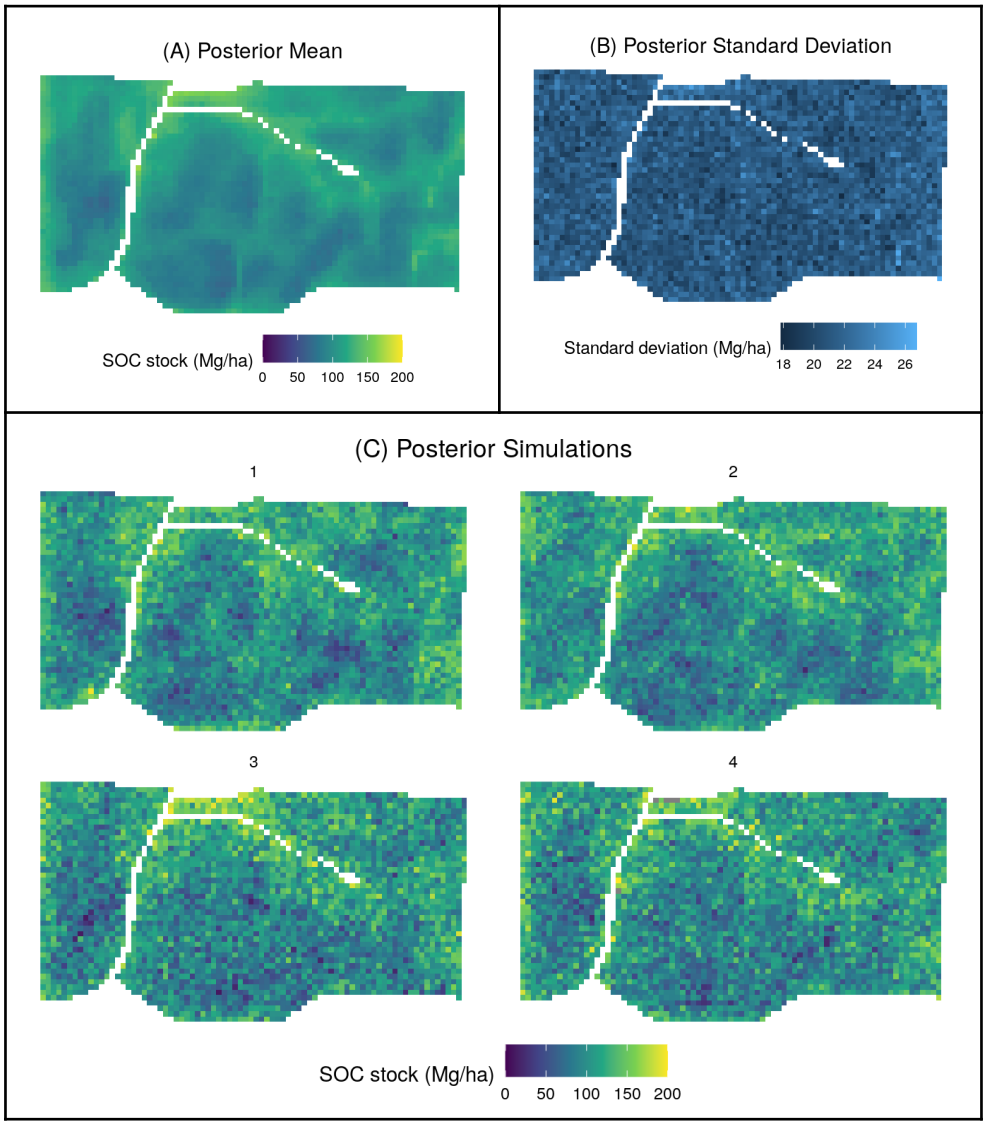


Figure 5: Estimated coefficients in the Bayesian Kriging with External Drift model, after standardizing predictors and outcome. Dots are posterior medians and error bars span posterior 50% and 95% intervals. Abbreviations: topographic wetness index (TWI), soil organic carbon (SOC), SOC index (SOCI), Soil Survey Geographic database (SSURGO).

Maps of the posterior mean and standard deviation (summarizing our uncertainty in the SOC stock at each point) are shown in Figure 6, along with posterior simulations of SOC stock which will be used to perform the ex-ante evaluation in the next section. Based on these posterior simulations, we estimated the mean SOC stock to 60 cm depth to be 103.4 Mg ha^{-1} (95% CI 100.8 to 106.6 Mg ha^{-1}). For comparison, previous studies of agricultural fields in the region have estimated mean SOC stock to 60 cm depth ranging from 91.0 Mg ha^{-1} (Zuber et al., 2015) to 172.6 Mg ha^{-1} (Johnson et al., 2011) and the SSURGO estimate for the site (obtained by weighting an estimate for each map unit) is 140.8 Mg ha^{-1}. The within-field standard deviation of SOC stock was 26.8 Mg ha^{-1} (95% CI 24.9 to 29.6) for a coefficient of variation of 26% (95% CI 24% to 29%). Thus the assumed measurement standard deviation of 1.44 Mg ha^{-1} (section 3.3) is very small compared to the within-field SOC stock standard deviation.



417

418 **Figure 6:** Modeled SOC stock map (A) posterior mean, (B) standard deviation, and (C) 4
419 (of 200) randomly chosen simulations used for ex-ante evaluation.

420

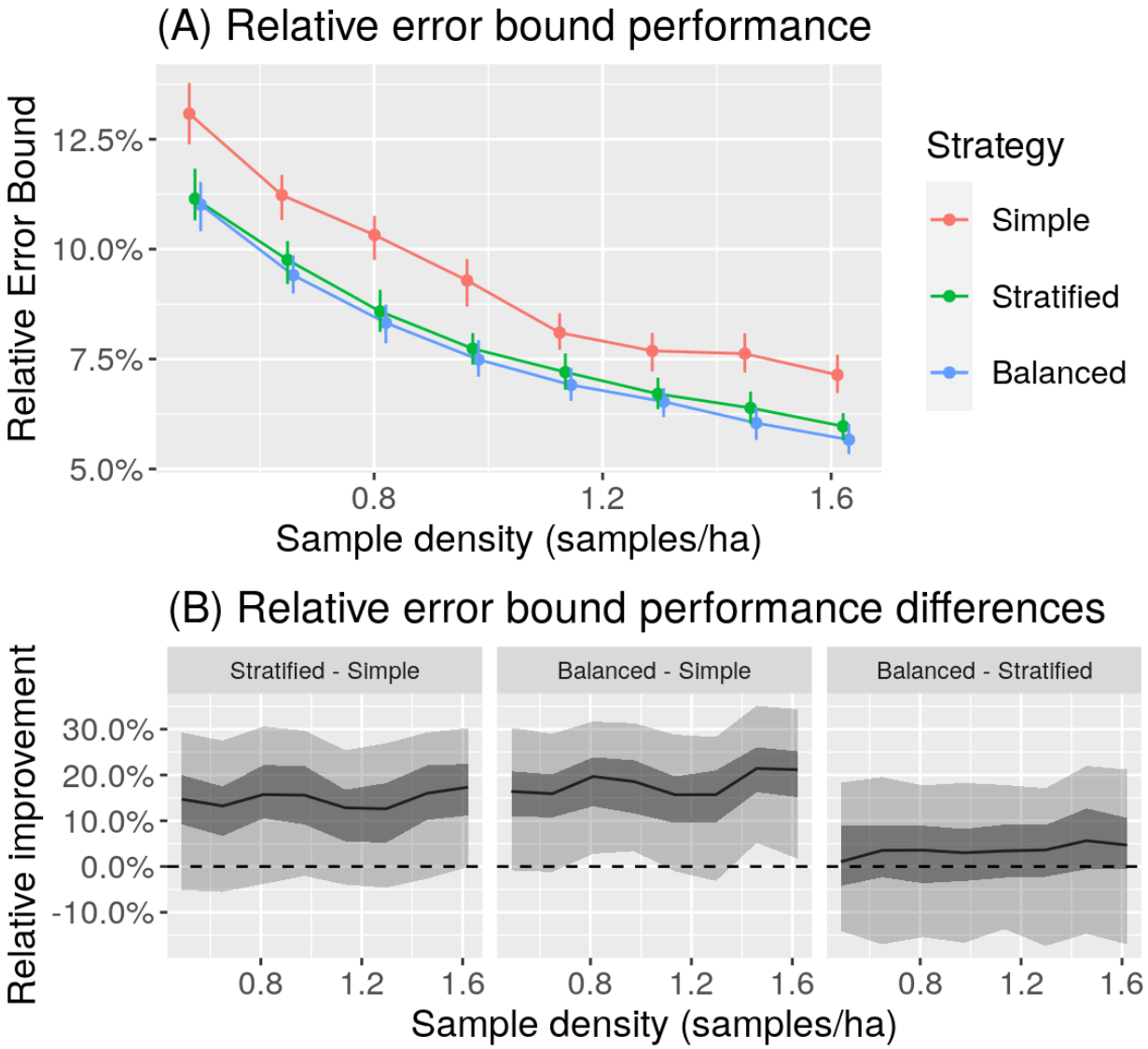
421 To examine the sensitivity of the ex-ante evaluation to this choice of mapping model we
422 also considered the BART model. While the BART model produced very similar estimates of
423 mean SOC stock (Figure S3) and explained a similar proportion variance (R^2 =0.48), we
424 observed a non-linear relationship between the BART and KED within-field predictions (Figure
425 S4). This suggests that BART and KED give qualitatively different SOC stock maps so that

comparing the results of ex-ante evaluation using the two models is a substantive test of sensitivity.

4.2 ex-ante evaluation

The stratifications produced for various sample sizes are displayed in Figure S5. Estimates of the relative error performance of the three primary estimation strategies are displayed in Figure 7A. For a given strategy and sample size, our evaluation technique produces samples of the distribution of relative error bound (section 2.1.3), one for each posterior map (Figure 6). We use the median of this distribution as a point estimate (i.e. a single number summary). Across the range of sample sizes, these point estimates show that balanced sampling outperforms stratified sampling outperforms simple random sampling. For each posterior map we can also calculated the confidence interval coverage rate, and we found that the 95% intervals for all three strategies obtain very nearly the nominal 95% rate (Figure S6). ex-ante evaluation results are qualitatively similar between the KED or BART SOC stock models (Figure 8), suggesting little sensitivity to this choice.

To quantify the difference in performance between any two strategies at a given sample size, our evaluation again produces a distribution-- now of the *difference* in relative error between the strategies. For each of the three pairs of comparisons between our three strategies, these distributions are shown in Figure 7B using the median, 50% and 95% intervals. We see that while there is little uncertainty that balanced sampling outperforms SRS, there is more uncertainty in the comparison of stratified sampling and SRS, and greater uncertainty comparing balanced and stratified sampling.



448

449 **Figure 7:** (A) Relative error performance of estimation strategies with dots and vertical

450 lines showing posterior medians and 50% CIs, respectively, and (B) relative difference in

451 relative error between strategies, showing posterior median (black), 50% CI (dark gray),

452 and 95% CI (light gray).

453

454

455

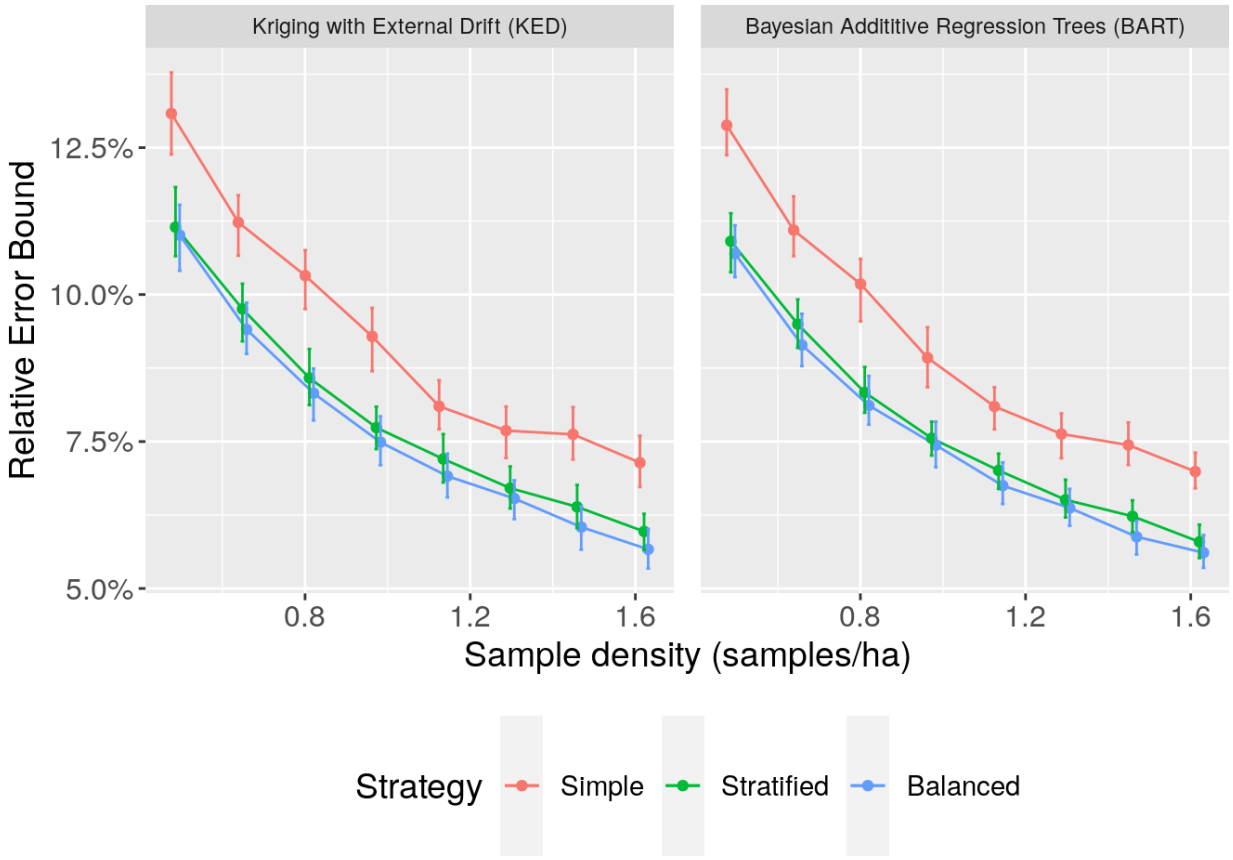


Figure 8: Sensitivity of relative error ex-ante evaluation results to choice of SOC stock map model. Dots and vertical lines show posterior median and 50% CI, respectively.

To assess the relative benefit of each of the covariates to the estimation performance, we considered designs including just one of the covariates. Stratifying on the Sentinel-2 SOCI covariate alone performed about as well as stratifying on all of the covariates together (Figure 9). At the same time, stratifying on easting performed about as well as no stratification, i.e. SRS. We also evaluated compact geographic stratification, which performed better than SRS but fell short of the full stratification (Figure S7).

The performance of the stratified estimation strategy was not sensitive to the minimum number of samples per stratum or the distance measure used on the covariates (Figures S8-S9).

Covariates were also incorporated into an estimation strategy with simple random sampling with a generalized regression model-assisted estimator. Compared to simple random sampling with the Horvitz-Thompson estimator, the model-assisted estimator with all covariates was an improvement, and using lasso to select covariates improved performance further (Figure S10). However, neither of these performed as well as the strategies with stratified or balanced sampling strategies.

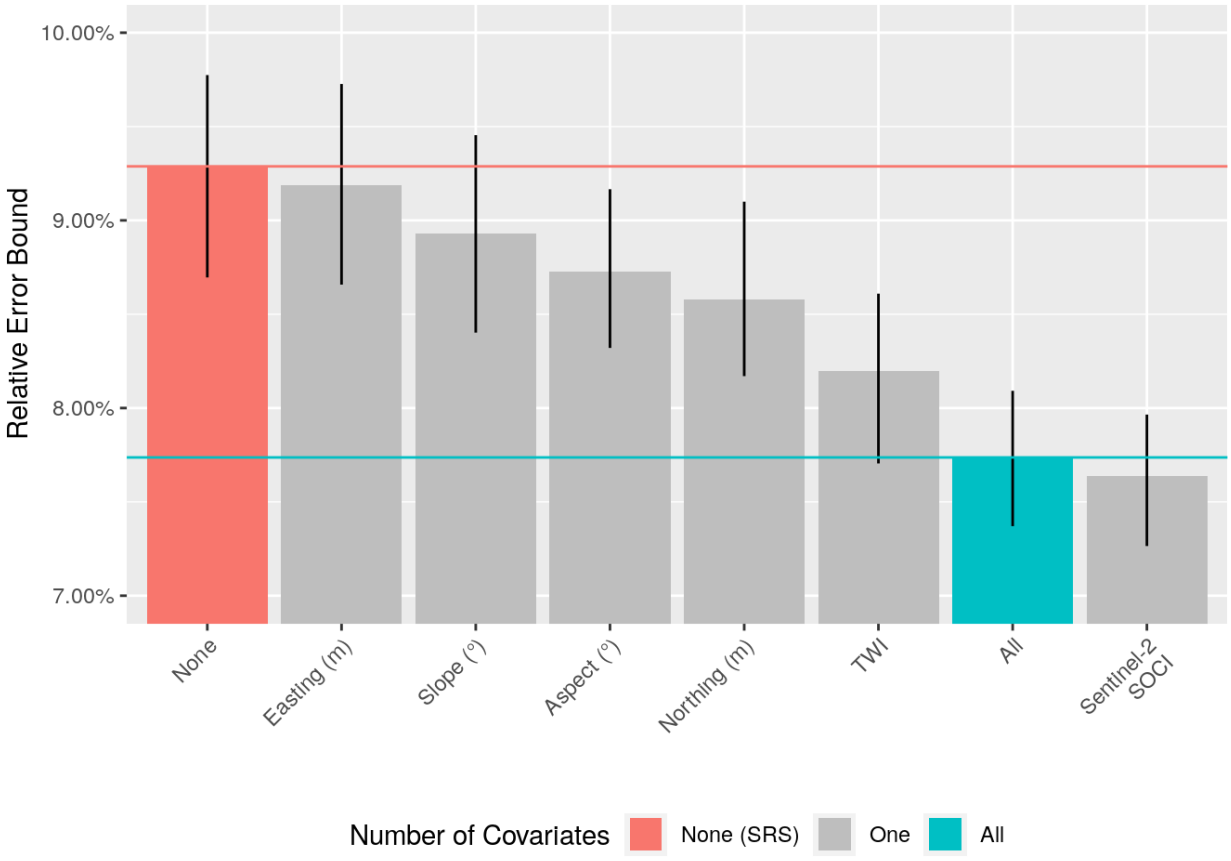


Figure 9: Performance of stratified sampling with various single covariates compared to no covariates, i.e. simple random sampling (SRS), and all covariates. Each design uses 30 samples (1.0 samples ha^{-1}). Bars and lines display posterior median and standard deviation, respectively.

5. Discussion and conclusions

We found that both stratified and balanced sampling strategies offer potentially substantial improvements in relative error over the SRS baseline. This result is promising because our implementation of these strategies only used auxiliary information that was already collected in public databases (SSURGO, NED) and so can easily be adopted for future mean SOC stock estimation studies. Model-assisted estimation did not show as much of an improvement over the baseline, suggesting that auxiliary information is most effectively incorporated in the sampling design stage. Our stratification, which requires no preliminary field work, likely achieves a 15% improvement over SRS across a range of sample sizes (Figure 9). As a function of the sample size n, relative error declines $\sqrt{n}$. This 15% improvement for a fixed sample size is thus equivalent to 28% fewer samples needed to achieve a given relative error.

Our sensitivity analysis found the results to be robust to the choice of SOC stock model, including both non-linear spatial autocorrelation (KED) and non-linear regression (BART). Because balanced sampling is predicated on a linear regression model, it is encouraging that balanced sampling performed well here. The Bayesian approach found substantial uncertainty in the performance of the estimation strategies and their comparison. This uncertainty stemmed from uncertainty in the SOC map models. To reduce the uncertainty about the performance benefits of these estimation strategies we would need to reduce the uncertainty of the SOC stock

maps used in evaluation. There are several ways to achieve this. Incorporating additional auxiliary information (e.g. proximal or remote sensing) may be helpful. Increasing the sample size of the reference sampling design, and improving the reference sampling design (e.g. an optimized model-based design instead of a grid). In particular, better mapping of short-range variation would be possible with more measurements made on distances less than the 35 m grid used here. In addition to reducing the uncertainty of the SOC stock maps it would be useful to use an auxiliary probability sample to obtain an independent estimate of the mean SOC stock and validate the maps themselves (Brus et al., 2011; Wadoux and Brus, 2021).

To compare our results to the literature, note that we found that using SRS approximately 1.0 samples per hectare would be needed to achieve a relative error bound of 10% (Figure 7). This matches prior estimates for SOM and SOC variability in similarly sized agricultural fields (Figure 2 of Lawrence et al., 2020), suggesting that those estimates could be used successfully to select a sample size for SRS in other fields. As described in the introduction there is a dearth of prior literature on the performance of stratified or balanced sampling in agricultural fields. The closest comparison is the stratification of an Australian farm (de Gruijter et al., 2016; section 2.2.1) which achieved a 29% improvement over SRS (section 2.4.1) compared to our 15% improvement, though the former relied on an initial reconnaissance sampling effort with measurements of SOC stocks to build the stratification.

The reliance on an imperfect ground truth map of SOC stock is a notable challenge to any ex-ante evaluation (section 2.2). Building on the methodology of past studies, we mitigated this challenge by examining the uncertainty and sensitivity of our results. However, both models as well as the stratified and balanced strategies shared the same set of covariates. This may have led to overestimating the performance of these strategies. We took steps to minimize potential overestimation of performance, including selecting a parsimonious set of covariates identified in

the literature and using stochastic models so that the simulated maps were not deterministic functions of the covariates. Our evaluation of strategies employing a single covariate (Figure 9) also shows that the performance benefit is not dependent on a complete coincidence of covariates. One way to avoid the issue is to use a ground truth model that does not employ covariates at all. We considered such a model (ordinary kriging) but it was a poor fit to the SOC stock measurements (Figure S2), undermining its utility for SOC stock mapping. With many more SOC stock measurements, the reliance on covariates for mapping SOC stock could be removed.

Our results can be used to inform future studies or monitoring projects. Where there are insufficient resources for reconnaissance prior to constructing a sample design, our results suggest that the use of publicly available information in stratified or balanced sampling can still offer substantial benefits over SRS. These findings also apply indirectly to quantifying the change in mean SOC stock over time when using unpaired samples (i.e. different sampling locations at different time points). However, the magnitude of the benefits of these sampling designs may vary across sites and may be related to factors such as soil type. In order to test the external validity of our findings, they will need to be replicated in other fields.

Acknowledgements

The authors acknowledge financial support from the DOE ARPA-E SMARTFARM program and the NSF Signal-in-Soil program.

References

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* 80 (1–2), 1–44.

Brus, D. J. "Using regression models in design-based estimation of spatial means of soil properties." *European Journal of Soil Science* 51.1 (2000): 159-172.

Brus, D.J., B. Kempen, and G. B. M. Heuvelink. "Sampling for validation of digital soil maps." *European Journal of Soil Science* 62.3 (2011): 394-407.

Brus, D.J. "Balanced sampling: a versatile sampling approach for statistical soil surveys." Geoderma 253 (2015): 111-121.

Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science* 72 (2), 686–703.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76 (1)

Chilès, J. P., and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty* (2nd ed.). John Wiley & Sons, 2012.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4.1 (2010): 266-298.

Deville, J.-C., Tillé, Y., 2004. Efficient balanced sampling: the cube method. Biometrika 91, 893–912.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*.

Grafström, Anton, and Lina Schelin. "How to select representative samples." *Scandinavian Journal of Statistics* 41.2 (2014): 277-290.

Grafström, Anton and Lisic, Jonathan (2019). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.5. https://CRAN.R-project.org/package=BalancedSampling

de Gruijter, Jaap, Dick J Brus, Marc FP Bierkens, and Martin Knotters. 2006. Sampling for Natural Resource Monitoring. Springer Science & Business Media.

de Gruijter, J. J., et al. "Farm-scale soil carbon auditing." Geoderma (2016).

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D, Bayesian Data Analysis, Chapman and Hall/CRC 2013.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer.

Hofman, Siegfried CK, and D. J. Brus. "How many sampling points are needed to estimate the mean nitrate-N content of agricultural fields? A geostatistical simulation approach with uncertain variograms." *Geoderma* 385 (2021): 114816.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical variables. Data Mining and Knowledege Discovery, 2:283–304.

ISO, Uncertainty of measurement — part 1: Introduction to the expression of uncertainty in measurement ISO Geneva, CH 2009.

Johnson, J. M. F., Archer, D. W., Weyers, S. L., & Barbour, N. W. (2011). Do mitigation strategies reduce global warming potential in the northern US corn belt?.

Lawrence, Patrick G., et al. "Guiding soil sampling strategies using classical and spatial statistics: A review." Agronomy Journal 112.1 (2020): 493-510.

Mallarino, Antonio P., and David J. Wittry. "Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter." Precision Agriculture 5.2 (2004): 131-144.

McConville, Kelly, Becky Tang, George Zhu, Shirley Cheung, and Sida Li (2018). Mase: Model-Assisted Survey Estimation. R package version 0.1.2 https://cran.r-project.org/package=mase

Oldfield, E.E., A.J. Eagle, R.L Rubin, J. Rudek, J. Sanderman, D.R. Gordon. 2021. Agricultural soil carbon credits: Making sense of protocols for carbon sequestration and net greenhouse gas removals. Environmental Defense Fund, New York, New York. edf.org/sites/default/files/content/agricultural-soil-carbon-credits-protocol-synthesis.pdf

Pebesma E.J., The role of external variables and GIS databases in geostatistical analysis, *Transactions in GIS*, 2006 615-632.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at the following link: http://websoilsurvey.sc.egov.usda.gov/. Accessed 2021-09-01.

Sparapani R, Spanbauer C, McCulloch R (2021). “Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package.” Journal of Statistical Software, 97(1), 1-66. doi: 10.18637/jss.v097.i01

Stockmann, U., et al. "Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis." Catena 139 (2016): 220-231.

Tautges, Nicole E., et al. "Deep soil inventories reveal that impacts of cover crops and compost

on soil carbon sequestration differ in surface and subsurface soils." Global change biology 25.11 (2019): 3753-3766.

Thaler, Evan A., Isaac J. Larsen, and Qian Yu. "A new index for remote sensing of soil organic carbon based solely on visible wavelengths." Soil Science Society of America Journal 83.5 (2019): 1443-1450.

Wadoux, Alexandre MJ-C., and Dick J. Brus. "How to compare sampling designs for mapping?." *European Journal of Soil Science* 72.1 (2021): 35-46.

Walthert et al. "Determination of organic and inorganic carbon, δ13C, and nitrogen in soils containing carbonates after acid fumigation with HCl." Journal of Plant Nutrition and Soil Science 173.2 (2010): 207-216.

Zuber, S. M., Behnke, G. D., Nafziger, E. D., & Villamil, M. B. (2015). Crop rotation and tillage effects on soil physical and chemical properties in Illinois. *Agronomy Journal*, *107*(3), 971-978.