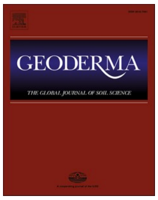
Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



Valid inferences about soil carbon in heterogeneous landscapes

Paige Stanley [a,*], Jacob Spertus [b], Jessica Chiartas [c], Philip B. Stark [b], Timothy Bowles [a]

[a] *Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA*
[b] *Department of Statistics, University of California, Berkeley, Berkeley, CA, USA*
[c] *Department of Land, Air, and Water Resources, University of California, Davis, Davis, CA, USA*

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil carbon
Spatial heterogeneity
Valid inference
Measurement
Carbon markets

ABSTRACT

Using soil organic carbon (SOC) to generate carbon offsets requires reliably quantifying SOC sequestration. However, accuracy of SOC measurement is limited by inherent spatial heterogeneity, variability of laboratory assays, unmet statistical assumptions, and the relatively small magnitude of SOC changes over time, among other things. Most SOC measurement protocols currently used to generate offsets for C markets do not adequately address these issues, threatening to undermine climate change mitigation efforts. Using analyses and simulations from 1,117 soil samples collected from California crop and rangelands, we quantified measurement errors and sources of uncertainty to optimize SOC measurement. We demonstrate that (1) spatial heterogeneity is a primary driver of uncertainty; (2) dry combustion assays contribute little to uncertainty, although inorganic C can increase error; (3) common statistical methods—Student's *t*-test and its relatives—can be unreliable for SOC (e.g. at low to medium sample sizes or when the distribution of SOC is skewed), which can lead to incorrect interpretations of SOC sequestration; and (4) common sample sizes (10–30 cores) are insufficiently powered to detect the modest SOC changes expected from management in heterogeneous agricultural landscapes. To reduce error and improve the reliability of future SOC offsets, protocols should: (1) require power analyses that include spatial heterogeneity to determine minimum sample sizes, rather than allowing arbitrarily small sample sizes; (2) minimize the use of compositing; (3) require dry combustion analysis, by the same lab for all assays; and (4) use nonparametric statistical tests and confidence intervals to control Type I error rates. While these changes might increase costs, they will make SOC estimates more accurate and more reliable, adding credibility to soil management as a climate change mitigation strategy.

1. Introduction

Interest in measuring soil organic carbon (SOC) is expanding dramatically because agricultural interventions that sequester C in soil may help to mitigate climate change. Recent policy initiatives and emerging soil C markets designed to accelerate management transitions require practical methods to measure SOC with low uncertainty or they may often reward false positives and fail to reward genuine sequestration. Indeed, the high uncertainty of SOC measurements likely contributed to the 2011 collapse of the Chicago Climate Exchange, the only prior U.S. voluntary C market (Gosnell et al., 2011).

In practice, the accuracy of SOC measurements is limited by spatial heterogeneity, sampling design, variability in bulk density, and variation in soil processing methods and laboratory assays. Reliably detecting and accurately quantifying changes in SOC stocks is challenging because, compared to these sources of variation and uncertainty, the annual changes produced by agricultural management interventions are often small (Bai et al., 2019; Minasny et al., 2017), for instance ranging from<0.1 % absolute change for conversion to no-till (Franzluebbers, 2005) to approximately 0.5 % with biochar application (Jones et al., 2012; Majumder et al., 2019). Methods for estimating SOC must be precise and powerful enough to detect such small changes in a heterogeneous medium (Ellert et al., 2002; Homann et al., 1998; Lehmann et al., 2007; Robertson et al., 1997). Minimizing the errors that arise in each of the many steps in SOC measurement (see SI 1 and SI Table 1 for a full description) is especially important in the context of C offset markets. Only accurate estimates of SOC sequestration with transparent levels of uncertainty should be used for generating credits and allowing governments and industries to offset greenhouse gas (GHG) emissions.

Yet protocols currently being used by C markets for measurement, reporting and verification (MRV) of SOC sequestration may be inadequate (Necpalova et al., 2014; Oldfield et al., 2021). Importantly, many

[*] Corresponding author.
E-mail address: paige.stanley@berkeley.edu (P. Stanley).

https://doi.org/10.1016/j.geoderma.2022.116323
Received 13 October 2022; Received in revised form 29 December 2022; Accepted 30 December 2022
Available online 4 January 2023

MRV protocols recommend but do not require—or else make no mention of—powering measurement campaigns using representative spatial heterogeneity information. Agricultural soils used to generate C credits have varying degrees of spatial heterogeneity and require different sample sizes to detect a given absolute or relative change in SOC. For example, spatial heterogeneity is typically higher on rangelands than croplands due to diverse topography, rocky soil horizons, low and patchy soil fertility, and patchiness of grazing and manure deposition. Rather than tailoring sample size requirements to expected levels of heterogeneity, many protocols (including the Climate Action Reserve Soil Enrichment Protocol, Australian Carbon Methodology, and Verra VM0021) simply set a minimum sample size within designated areas (e.g., 3 samples per stratum). If the MRV protocol does not require determining the number of samples necessary to detect a reasonable level of SOC sequestration, it could fail to reward legitimate sequestration or have a large chance of erroneously rewarding nonexistent sequestration.

Addressing knowledge gaps associated with sampling design—including sample placement, stratification, and compositing—could further reduce the measurement uncertainties of SOC offsets. For example, C market protocols often encourage the use of systematic sampling, but samples collected by simple or stratified random sampling are less likely to bias SOC estimates and allow more rigorous statistical analysis. While stratifying soil sampling into more homogeneous land subunits can increase the power to detect SOC sequestration, many protocols lack quantitative guidance for defining strata and some do not require field sampling at all, relying instead on model output (Oldfield et al., 2021), with notable exceptions (e.g. Australian Carbon Methodology). Compositing—combining samples to reduce analysis costs—is a common practice allowed in MRV protocols (e.g. Climate Action Reserve Soil Enrichment Protocol, Australian Carbon Methodology), though the impact on measurement error is often unknown (de Gruijter et al., 2016). At one extreme, all samples collected within an experimental unit can be combined into a single sample for analysis (i.e., full compositing) (Carey et al., 2020; Tautges et al., 2019), making it impossible to estimate spatial heterogeneity and substantially increasing measurement error (Spertus, 2021).

The impact of compositing on measurement error depends in part on the error of laboratory analyses. The extent to which dry combustion assays contribute to overall error in measuring SOC from either intra-lab (replicated measurements on the same instrument) or inter-lab (measurements on different instruments) analytical variability is not well known (Chatterjee et al., 2009; O' Rourke and Holden, 2011), limiting the ability to optimize sampling campaigns and the reliability of estimates and inferences. Compositing and subsequent laboratory analyses can be optimized to minimize contribution to error within a given budget, given estimates of spatial heterogeneity, analytical error, and laboratory costs (Spertus, 2021). To our knowledge, such an analysis has never been done to inform soil-sampling campaigns.

Lastly, the choice of statistical methods for data analysis also influences the likelihood of false positives (Type I errors)—generating C offsets when SOC wasn't sequestered—and false negatives (Type II errors)—failing to generate C offsets when SOC was sequestered. In a C market, Type I error can lead to allocation of payments without actual SOC sequestration, and possibly even increase net C emissions; while a Type II error can fail to generate C offsets when SOC is sequestered (Sanderman and Baldock, 2010). Both types of error undermine the utility of C markets, leading to missed opportunities for climate change mitigation. When the assumptions required of common statistical methods are not met (e.g., SOC is not normally distributed), standard hypothesis tests can have Type I error rates that greatly exceed their nominal significance level (e.g., 5 %), and confidence intervals can have coverage probabilities far lower than nominal (e.g., 95 %) (Lehmann and Romano, 2010). For example, the two-sample Student *t*-test is often used to assess changes in SOC stocks (Brus and de Gruijter, 2011; deGruijter et al., 2016; Kravchenko and Robertson, 2011). Student's *t*-test assumes that SOC at both measurement times is normally distributed with the same variance. Since SOC generally does not have a normal distribution (Yan et al., 2011) and because agricultural management interventions can redistribute SOC without changing the total (Chappell et al., 2012), Student's t confidence intervals can have true coverage probabilities far lower than the nominal confidence level (e.g., 95 %), and Student's t-tests can have true Type I error rates that greatly exceed the nominal significance level (e.g., 5 %) (Lehmann and Romano, 2010). This undermines the validity of many standard methods for inference—including ANOVA, mixed effects models, geostatistical models, bootstrapping, Wilcoxon rank-sum tests, permutation tests, and Bayesian models. Quantifying the chance of false conclusions about whether and how much SOC has been sequestered is crucial for SOC offsets.

Lesser-known statistical methods can strictly limit the Type 1 error rate and increase reliability. For example, there are nonparametric tests and confidence intervals that are valid for any SOC distribution (Anderson, 1969; Learned-Miller and Thomas, 2019; Romano and Wolf, 2000; Stark, 2009, 2023; Waudby-Smith and Ramdas, 2020). These methods are conservative or exact: the probability of Type I errors is not larger than the nominal significance level (e.g., 5 %), and the chance that confidence intervals include the true amount of SOC sequestered is not less than the nominal confidence level (e.g., 95 %). Suitable nonparametric tests and confidence intervals can produce reliable inferences about SOC stocks and changes, though their widespread adoption has been hindered by their relatively low power.

Below, we investigate these uncertainties and knowledge gaps and how they affect the cost and reliability of SOC sequestration measurements. Using new, on-farm data from California crop and rangelands, we 1) evaluate the relative impact of spatial heterogeneity, analytical variability, and compositing on measurement precision and power; 2) use simulations to examine the validity and power of common statistical tests to detect SOC sequestration using different sampling designs on high and low heterogeneity agricultural landscapes; and 3) compare the validity and power of the *t*-test to those of a new nonparametric method across a range of sample sizes and SOC changes. Based on our findings, we make straightforward recommendations, targeted toward SOC markets, to improve the accuracy and reliability of SOC sequestration measurements, yield more trustworthy C credits, and support progress towards climate change mitigation goals.

2. Methods

2.1. Collecting SOC data: Rangeland and cropland sampling and laboratory analysis

We leverage new data collected from two intensive field sampling campaigns on California crop and rangelands. While these samples were originally collected for other purposes, we use them to study field-level spatial heterogeneity and to provide an empirical basis for simulations. We outline our sampling methods briefly below, with more details **SI 2.1**.

Rangeland samples were collected in December 2019 from a ranch in Paicines, California. The data were collected to quantify spatial heterogeneity of SOC in a constrained, field-scale setting, controlling for soil type, catenal position, slope aspect, and vegetation—not to quantify SOC stock for the whole ranch. We used soil survey information within the ranch boundaries (SSURGO; Soil Survey Staff et al., n.d.) to identify Auberry Fine Sandy Loam soils. Samples were collected using a stratified transect design with five 100 m transects on two adjacent hillslopes stratified by slope position: summit/shoulder (1 transect), backslope (2 transects), and footslope (2 transects). Soils were sampled down to 100 cm, or the point of refusal, and divided into 5 depth ranges (0–10 cm, 10–30 cm, 30–50 cm, 50–75 cm and 75–100 cm). We attempted to collect 33 samples along each transect, but time constraints limited us to 25 samples at one transect. In all, we attempted to collect 785 samples, but bedrock or rock obstructions limited the depth of sampling at some locations (mostly along the summit position), resulting in 662 total samples. Each sample was air-dried and sieved to 2 mm. Visible plant materials were removed, and soils were ground using a ball mill (**SI 2.1**; Retsch, Newtown, PA).

Cropland soil samples were collected in September and October 2019, from seven farms across Southern California (SI Fig. 1) representing various soil types and cropping systems, including two orchards, a vineyard, two intensive cropping systems, and two diversified farms (full soil taxonomy in SI Table 2). Samples were collected along 50 m transects. At each site, transect locations were selected based on the dominant soil type (Soil Survey Geographic (SSURGO) Database, United States Department of Agriculture, Natural Resource Conservation Service), consistent historic and current management, and cropping system. The number of transects ranged from two at the small, diversified farms to six at one of the larger cropland sites. Depth ranges were defined slightly differently at different sites based on tillage depths (0–10, 10–20 cm vs 0–15, 15–30 cm) and genetic horizon in the subsurface. In all, 455 samples were collected from the seven farms. Samples were air-dried and sieved to 2 mm; visible plant materials were removed; and then soils were oven-dried at 60̊C and ground using a ball mill.

Bulk density (BD) samples for croplands and rangelands were collected using the pit method (Walter et al., 2016). Cores were collected from the center of each depth increment used for bulk soil samples. For sampling depths greater than 10 cm, multiple cores were collected to ensure samples were representative. At the rangeland site, three 1.5 m deep soil pits were dug along each transect (one at each end, and one in the center at 50 m) using an excavator, a total of 15 pits. At the cropland sites, one soil pit was dug at the central location of each transect to 1.5 m or the point of refusal. Bulk density samples were oven-dried at 105̊C until their weight no longer decreased. Visible rock fragments were removed before weighing the samples and submerged in water to measure their volume. Rock volume was subtracted from core volume in estimating soil density. We used bulk density and TC% to calculate SOC stocks for each depth increment.

Two different dry combustion analyzers were used to measure C concentrations (TC%) of prepared samples. All cropland soils were analyzed with a Costech ECS 4010 elemental analyzer (Costech, Valencia, CA)—a widely used instrument for dry combustion analysis. Rangeland samples were analyzed on an Elementar soliTOC cube (Elementar, Ronkonkoma, NY; see (Natali et al., 2020)), a relatively new instrument designed to improve precision by combusting higher sample masses (up to 3 g of soil vs ~ 50 mg) while separating total organic C (TOC), residual organic C (ROC), and total inorganic C (TIC) via a temperature ramping method, DIN19539. The ECS 4010 measures only the mass of total C (TC), and thus we compare only TC% between the two instruments.

To quantify the precision and bias of each instrument, we reanalyzed 15 rangeland and 22 cropland samples that had the minimum, median, and maximum TC% for each depth and land-use category (SI Fig. 2). Five analytical replicates of each sample were measured on each instrument. Samples with high TIC (greater than0.1 %), as measured by the soliTOC, were treated with HCl to remove TIC and re-assayed on the ECS 4010 (SI Figs. 2 and 3**)**. Finally, we ran 25 additional replicates of two soil standards with known TC% on each instrument (SI Fig. 4).

2.2. *Assessing spatial heterogeneity of SOC and bulk density*

To visualize the relative heterogeneity of TC% by land use, depth, and transect, we used histograms, sample means, and coefficients of variation (CV). While TC% of the non-rocky component of the soil is the focus of this study, we also examined the variability of BD measurements by comparing the CV across depths for rangeland and cropland site 7, which had substantially more BD samples than other cropland sites**.**

To assess differences in spatial heterogeneity across land uses, depth, strata (transects), and sites, we tested the hypotheses that population TC% distributions were equal across depths and transects on rangeland soils, or depths and sites on cropland soils using a nonparametric test called *permutation ANOVA*, a way of calibrating the ANOVA test statistic to control the rate of false rejections without any assumption about the distribution of SOC (Pesarin and Salmaso, 2010). Details of how the permutation ANOVA was performed are in **SI 2.2**. Code is available at: github.com/spertus/soil-carbon-statistics.

2.3. *Evaluating analytical variability*

We repeated analyses of the same samples to estimate the variability of laboratory assays. For each sample and instrument, analytical error was quantified by the *estimated relative error* (see **SI 2.3** for the formula), which is approximately the CV. We report the median estimated relative error for each instrument. (The estimated relative error measures variability but not bias; we estimated bias using measurements of known standards.) To evaluate whether there were systematic differences in measurements between the two instruments (soliTOC and ECS 4010), we used permutation tests for the two-sample problem, which asks whether the difference between two samples would be unlikely if the samples were created by randomly partitioning their pooled values into the two groups. We used the difference in means as the test statistic and simulated 10,000 draws from the permutation distribution using the R package **permuter** (see **SI 2.3** for more details).

2.4. *Quantifying relative uncertainty from spatial and analytical heterogeneity*

We quantified the contributions of analytical variability and spatial heterogeneity to uncertainty in estimates of the population mean TC% using the delta method (Goidts et al., 2009), which decomposes the total uncertainty into a sum of the contributions from analytical variability and spatial heterogeneity (**SI 2.4).** If the ratio of the contribution from spatial heterogeneity to total uncertainty is close to 1, spatial heterogeneity contributes more than analytical variability to overall uncertainty, vice versa if the ratio is closer to 0. If the ratio is 0.5, analytical and spatial heterogeneity contribute equally to total uncertainty. To assess how compositing affects the relative contributions to uncertainty, we computed the proportion of total uncertainty due to spatial heterogeneity without compositing, and the corresponding proportion when 90 cores are composited to one analytic sample (an extreme degree of compositing). We computed these ratios within depths for both land use types and both instruments.

2.5. *Comparing how sources of error affect statistical power*

We studied how spatial heterogeneity, assay variability, and compositing affect the ability to detect changes in average TC%. Specifically, we approximated the power of the unpaired two-sample *t*-test when samples are drawn by simple random sampling and there is no compositing, optimal compositing (derived in (Spertus,2021), or full compositing. We only examined the power for relatively large sample sizes ($n \geq 90$ cores) because Student's *t*-test is especially unreliable for small sample sizes (see below).

Comparing compositing strategies requires a budget; if money were no object, assaying every sample separately (i.e., not compositing) minimizes error. Compositing involves a tradeoff between various costs and errors. To explore the tradeoff, we took the marginal cost of collecting a single soil core in the field to be $20 USD and the cost of laboratory analysis (including sample preparation) in an elemental analyzer to be $13.60 USD per sample, the average price charged by five commercial labs for TC% analysis. Given these unit costs, the cost to collect, prepare, and analyze 90 cores (without compositing) is $3,024 USD. Using the same total budget, we explored what the uncertainty would have been had the money been used to take more cores and composite some of them (*optimal compositing,* which maximizes power within the budget) or all of them (full compositing) instead of assaying them individually. The power calculations use the estimates of land-use-specific TC% average and spatial heterogeneity (averaged across sites for cropland) and instrument-specific median relative error to approximate the power to detect a change of a given magnitude.

2.6. *Power and validity of tests for detecting TC% change*

We performed two simulations to estimate the true significance level and power of different two-sample hypothesis tests. The “validity simulation” estimated the significance level (i.e., Type 1 error rate) of two tests—the chance a test erroneously rejects the null hypothesis when there is no change in total SOC—in four scenarios. The two tests were the usual two-sample Student *t*-test and a nonparametric test that uses a pre-specified upper bound on TC concentration (See SI 1.6 for details). We set this bound at 10 % or 20 % TC, established TC ranges in mineral soils. In each of the four scenarios, SOC means were set exactly equal, but the shapes of the SOC distributions could differ in ways that might plausibly result from agricultural interventions, inferred from our empirical crop and rangeland TC data. In the “unchanged normal distribution” scenario, both distributions were normal with SDs of 0.5 %; in the “tilled cropland” scenario, the distribution at the first time was the actual topsoil TC% samples from cropland site 5 (right-skewed, Fig. 1) and the distribution at the second time was normal with SD 0.5 %; in the “change in skew” scenario, the distribution at time 1 was rangeland topsoil samples (right-skewed) and the distribution at time 2 was the same but multiplied by −1 (left-skewed); in the “extreme hotspot” scenario, the distribution at time 1 had 99 % of its mass in a normal

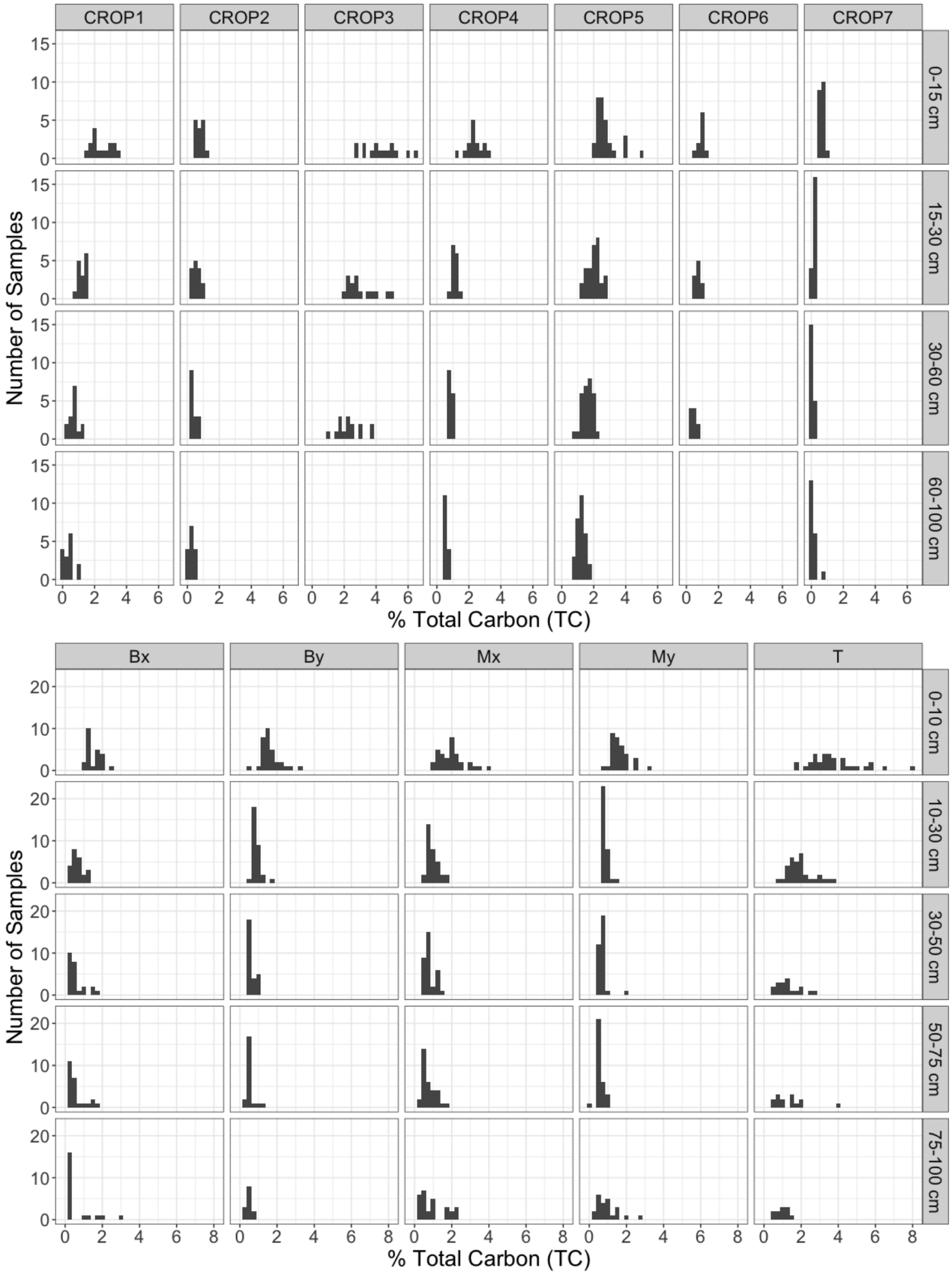


Fig. 1. Histograms TC% by (a) transect and depth in rangeland soils and (b) site and depth in cropland soils. Transect labels in (a) refer to catental positions and replicates: Bx (footslope, replicate X), By (footslope, replicate Y), Mx (backslope, replicate X), My (backslope, replicate Y), and T (summit/shoulder, no replication). Site labels in (a) refer to cropland sites: CROP1 is a diversified farming system, CROP2 is a vineyard, CROP3 is an orchard, CROP4 and CROP 5 use conventional cropping, CROP6 is a diversified farming system, and CROP7 is an orchard. Depth increments differed between rangeland and cropland sampling schemes. Plotted values are TC%, which is equal to TOC% in samples with zero TIC.

distribution centered at 2.8 % TC (SD: 0.05 %) and 1 % as a point-mass centered at about 20 % TC (Beem-Miller et al., 2016; Miller et al., 2016; Mishra and Riley, 2015). At the second time, the distribution was normal (SD: 0.05 %). We ran both tests at a nominal 5 % level 5000 times with sample sizes ranging from 5 to 150 at each epoch and recorded the rate of (false) rejections. We compared these simulated significance levels to the nominal 5 % significance level (Fig. 6).

The "power simulation" estimated the chance of detecting increases in SOC of various magnitudes with sample sizes 10, 30, 90, and 200 using the Student's *t*-test with unstratified sampling, Student's *t*-test with stratified sampling, and the nonparametric test with unstratified sampling. (Stratified nonparametric tests are in development.) The reference population distributions (at time 1) were taken to be the empirical distributions of samples from the rangeland site or from cropland site 5, which had median spatial heterogeneity and the most samples among the cropland sites. The hypothetical change in TC% was an additive shift of the reference distribution, with shifts ranging from 0 % (no change) to 60 % of baseline. For example, the baseline average TC % across our cropland sites was 2.7 % TC, so the simulated TC% at time 2 ranged from 2.7 % to 4.32 %. The stratified Student *t*-test was used only on rangeland samples because the cropland transects were not stratified and there were few samples per transect. For the purpose of this simulation, we treated each of the 5 transects as if it were a random sample from a distinct stratum. Under this assumption, samples from a transect are representative of the distribution of %TC within the corresponding stratum. This assumption is probably false in a way that favors the stratified Student *t*-test—within-transect heterogeneity is likely lower and between-transect variation higher than the corresponding quantities in an actual stratified random sampling design. The simulations sampled independently with replacement from each distribution (either pooled or stratified by transect), conducted the tests at nominal significance level 5 %, and recorded whether the null was rejected. The nonparametric test requires the user to specify an upper bound on the concentration: smaller bounds leads to more powerful tests, but misspecification can make the test invalid. We ran nonparametric tests with upper bounds of 10 %, which exceeds the maximum in any of our data (7.8 % TC), and 20 %, the established bound on TC in mineral soils. We also ran the nonparametric tests at a significance level of 10 % to examine how raising the significance level increases power. We ran each simulation 500 times, with 10, 30, 90, or 200 samples drawn from the population at each epoch. For stratified sampling, sample sizes were allocated proportional to "size," measured by the number of samples in the original transect.

All statistical analyses were conducted in R (version 3.6.1). Code is available at: https://github.com/spertus/soil-carbon-statistics.

3. Results

3.1. Spatial heterogeneity of SOC and bulk density

In both rangeland and cropland soils, TC% generally decreased with depth (Fig. 1). In rangeland soils, mean TC% varied from 3.77 % in topsoils (0–10 cm) of the summit/shoulder transect to 0.47 % at 75–100 cm of the footslope transect. Mean TC% in cropland soils varied from 4.31 % at 0–15 cm (at CROP3) to 0.10 % at 60–100 cm (at CROP7). Permutation ANOVA found that variations in mean TC% were statistically significant across transects (p < 1e-4) and depth (p < 1e-4) in rangeland soils and across sites (p < 1e-4) and depth (p < 1e-4) in cropland soils. Mean TOC% at the rangeland site was similar to TC% (SI Fig. 10): most samples had low TIC%.

The spatial heterogeneity of TC% varied with land use, depth, and geographic location (transect and site; Table 1). Heterogeneity of TC%, as measured by the coefficient of variation (CV), was higher in the rangeland site than in the cropland sites. The CV increased with depth in every rangeland transect and in cropland sites with diversified or perennial farming systems (vineyards and orchards), but not in conventionally managed croplands.

Bulk density was highly variable with land use and across sites but generally not with depth. Heterogeneity was particularly high in the rangeland soils, where CV ranged from 0.08 at 30–50 cm and 75–100 cm to 0.16 at 0–10 cm and 50–75 cm. Heterogeneity within the cropland sites was lower with CV ranging from 0.04 at 0–15 cm and 15–30 cm to 0.07 at 60–100 cm. Within a given depth, BD varied substantially across rangelands (15 soil pits) and the CROP7 site (16 soil pits), but no consistent patterns emerged (Fig. 2). BD for the six other cropland sites combined is plotted in SI Fig. 11.

Rangeland SOC stocks were 30.3, 31.6, 22.5, 25.9, and 30.4 Mg C/ha at 0–10, 10–30, 30–50, 50–75, and 75–100 cm, respectively. Whole profile stocks (0–100 cm) were 141.7 Mg C/ha, SE: 6.7 (SI Table 3). Like most SOC stock estimates, the estimated SE does not reflect uncertainty and variability of bulk density (although we argue below that those uncertainties should be taken into account). In croplands, whole profile SOC stocks varied by site from 32.6 Mg C/ha (0–100 cm for CROP7) to 230.0 Mg/ha (0–70 cm for CROP3) (SI Table 4).

3.2. Analytical variability

We compared measurements of 25 analytical replicates of two standard soils on the soliTOC and ECS 4010 dry combustion analyzers. Both instruments showed low variance and a small but consistent positive bias (SI Fig. 4). Based on analytical replicates of 36 samples measured on both instruments, the estimated median relative errors of

Table 1
Estimates of TC% means and coefficients of variation (CV; in parentheses) for cropland and rangeland. Mean and CV for rangeland sites are listed by transect. Mean and CV for croplands are listed by site. Cropland depths were not always consistent by site. For example, the second sampling depth ranged from 15–30, 15–35, and 15–40 in some cases. We used the most common depth increments here.

CROPLAND

Depth (cm)	CROP1	CROP2	CROP3	CROP4	CROP5	CROP6	CROP7	Total
	DFS	Vineyard	Orchard	Crop	Crop	DFS	Orchard	
0–15	2.45 (0.26)	0.82 (0.26)	4.31 (0.26)	2.37 (0.21)	2.74 (0.25)	0.94 (0.20)	0.64 (0.21)	2.04 (0.24)
15–30	1.21 (0.19)	0.56 (0.41)	3.05 (0.31)	1.14 (0.15)	2.03 (0.21)	0.73 (0.27)	0.17 (0.36)	1.27 (0.25)
30–60	0.73 (0.39)	0.36 (0.58)	2.36 (0.34)	0.88 (0.13)	1.60 (0.22)	0.50 (0.42)	0.10 (0.44)	0.93 (0.31)
60–100	0.42 (0.77)	0.25 (0.62)	–	0.56 (0.21)	1.22 (0.21)	–	0.12 (1.14)	0.52 (0.38)

RANGELAND

Depth (cm)	Bx	By	Mx	My	T	Total
0–10	1.55 (0.27)	1.63 (0.32)	2.02 (0.36)	1.63 (0.32)	3.77 (0.36)	2.16 (0.54)
10–30	0.67 (0.48)	0.90 (0.27)	0.99 (0.32)	0.86 (0.20)	1.99 (0.36)	1.11 (0.56)
30–50	0.60 (0.66)	0.64 (0.28)	0.82 (0.33)	0.71 (0.39)	1.32 (0.48)	0.78 (0.53)
50–75	0.59 (0.75)	0.53 (0.41)	0.75 (0.51)	0.59 (0.36)	1.41 (0.63)	0.71 (0.70)
75–100	0.65 (1.17)	0.47 (0.27)	1.01 (0.71)	0.94 (0.61)	0.96 (0.34)	0.84 (0.75)

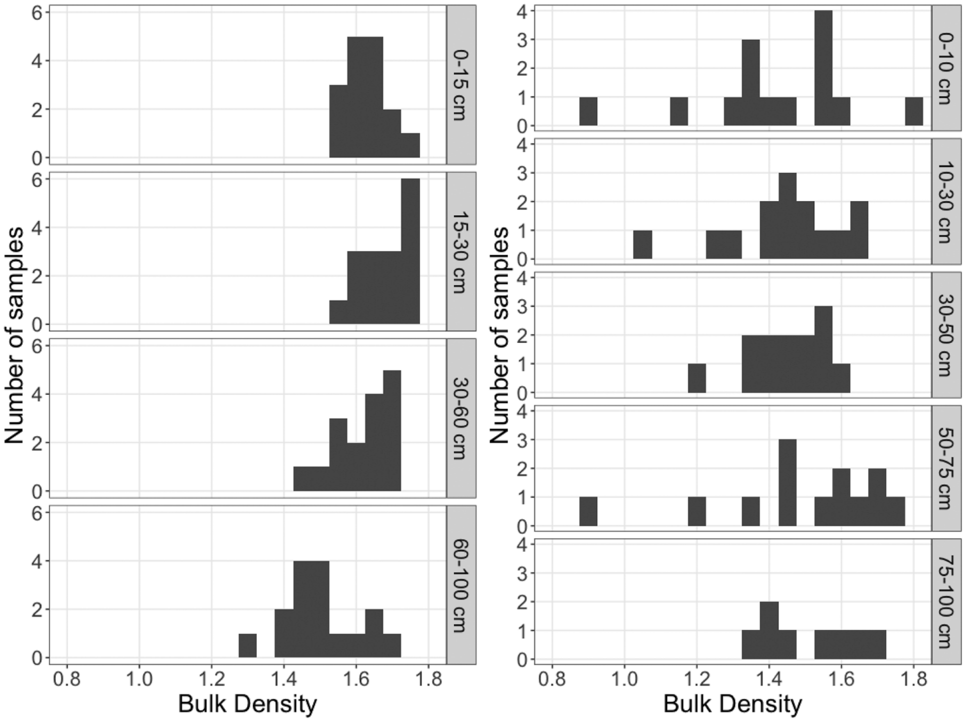


Fig. 2. Empirical distributions of bulk density (BD) samples across 16 soil pits on CROP7 (a) and 15 rangeland soil pits (b) by depth (in rows).

the measurements were 0.024 for the soliTOC and 0.061 for the ECS 4010 (Fig. 3). Permutation tests generally found little evidence of systematic differences between the instruments in replicated TC% measurements, except for samples with TIC% greater than 10 % of TC%. In the most extreme case, average replicated TC% measured on the soliTOC was nearly triple that of ECS 4010 for a rangeland sample with ~ 90 % of TC% as TIC%. Removing inorganic C with HCl improved the agreement of measured TOC% between the two instruments (SI Fig. 2).

3.3. Sources of uncertainty and their effects on statistical power

In general, spatial heterogeneity contributes much more uncertainty than analytical variability does, both for rangelands and croplands (Fig. 4). However, compositing can mitigate or exacerbate the relative contributions to uncertainty from spatial heterogeneity and analytical variability. With no compositing, analytical variability contributes little to the overall uncertainty (Fig. 4). If all $n = 90$ cores are composited to $k = 1$ analytic sample ("full compositing"), analytical error becomes a major component of the uncertainty in estimates of TC% for cropland soils, especially for the less precise ECS 4010 analyzer (Fig. 4). The theoretical power of Student's *t*-test under various compositing schemes reflects this tradeoff (Fig. 5). The power of Student's *t*-test to detect TC% change generally depends more on spatial heterogeneity than analytical variability for both instruments, except for full compositing with the ECS 4010, which had much less power (Fig. 5).

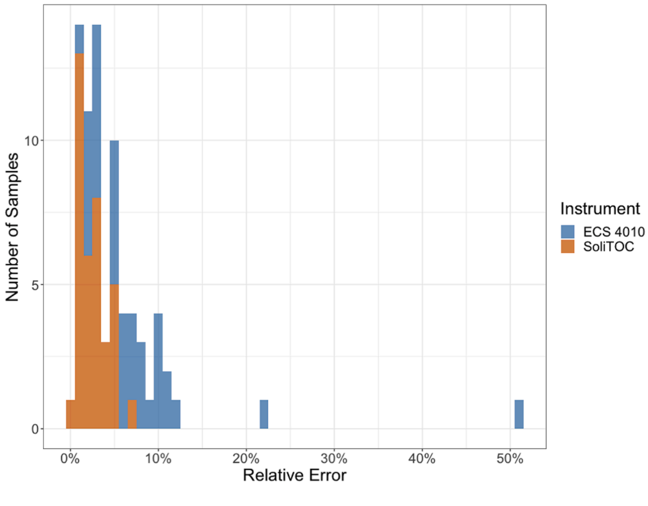


Fig. 3. Histogram of relative error of replicated assays computed for each sample run on soliTOC (blue) and ECS 4010 (orange). Histograms bins are 1% relative error wide and stacked. The samples with relative error above 20% on ECS 4010 had high proportions of inorganic C.

There was relatively little difference in power between optimal compositing and no compositing for every land use and analytical instrument. When spatial heterogeneity is high (e.g., in rangeland) and lab analysis is precise (e.g., with soliTOC), power is maximized by allocating more of the budget to sampling and using some compositing to reduce the number of assays. On the other hand, when spatial heterogeneity is low (e.g., in cropland) and lab analysis is imprecise (e.g., ECS 4010), accuracy is maximized by allocating more of the budget to assays and reducing or avoiding compositing.

3.4. Power and validity of tests for detecting TC% change

The nominal significance level of Student's *t*-test can greatly understate its actual chance of making a Type I error, i.e., of erroneously rejecting the null hypothesis when the hypothesis is true (Fig. 6). In the validity simulations, the true significance level of Student's *t*-test was always larger than its nominal level, except when the distributions were both normal. In the "tilled cropland" and "change in skew" scenarios, the level was close to 10 % at very small sample sizes, but approached its nominal 5 % at larger sizes. In the "extreme hotspot" scenario, the true significance level was always many times higher than the nominal significance level, and remained above 20 % for a sample size of 150. In contrast, the nonparametric test never erroneously rejected the null hypothesis.

Our "power" simulation compared the power of the unstratified Student *t*-test, stratified Student *t*-test, and a nonparametric test for

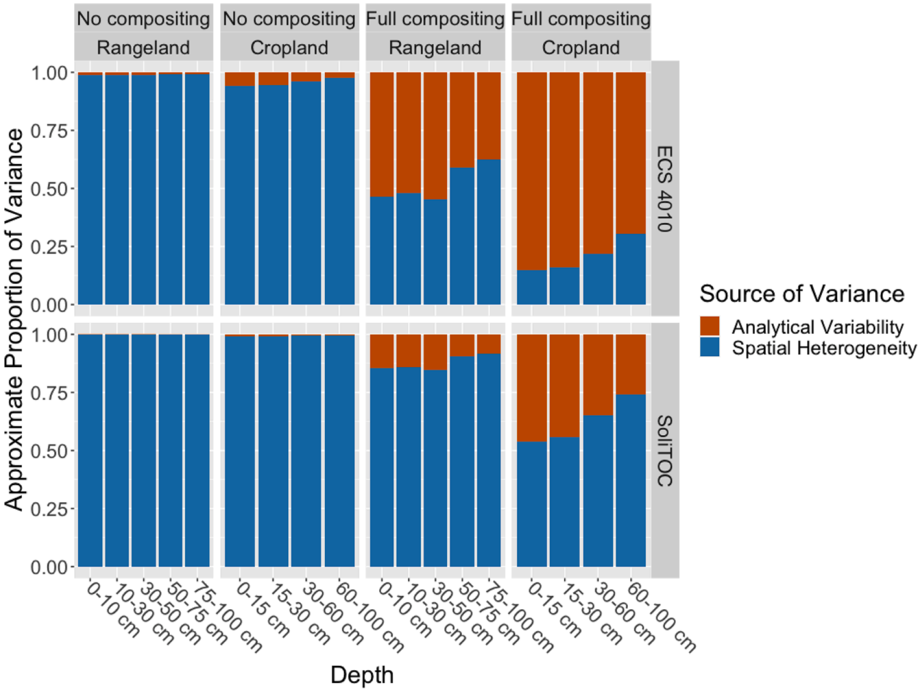


Fig. 4. Contributions to the variance of the sample mean from assay uncertainty and compositing. Proportion of variance (y-axis) reflects assaying either all field samples individually ("No Compositing" panels) or 90 field samples together ("Full Compositing"). Different panels correspond to different land uses (in columns) and instruments (in rows). Field heterogeneity is estimated using data from the rangeland site or averages across cropland sites, at various depths (x-axis). Assay variability is estimated either on ECS 4010 (top panels) or soliTOC (bottom panels) elemental analyzers. Rangeland depths: a:0–10 cm, b:10– 30 cm, c:30–50 cm, d:50–75 cm and e:75–100 cm. Cropland depths vary slightly by site (see Fig. 1).

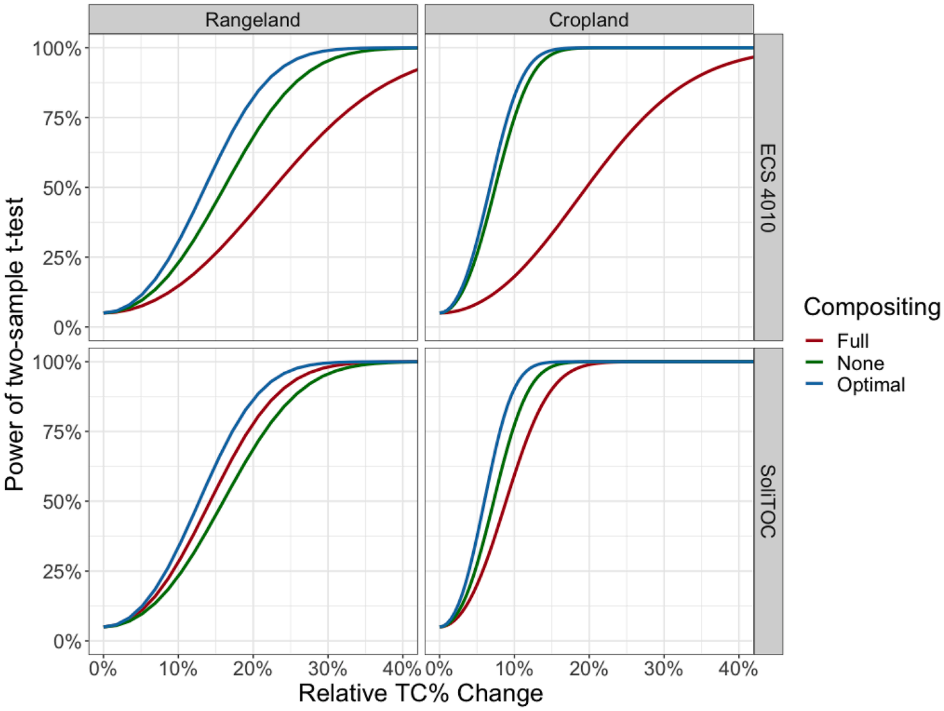


Fig. 5. Theoretical power of Student's *t*-test to detect changes in topsoil TC% (if TC% is normally distributed) for two levels of compositing, for a budget that covers the cost of 90 cores and 90 laboratory analyses without compositing, or 150 cores and one laboratory analysis for full compositing. The X-axis shows the relative change in average TC% from baseline (2.16 TC% for rangeland or 2.04 TC% for cropland); the Y-axis is power. Different panels correspond to different land types (in columns) and instruments (in rows). Colors correspond to the compositing scheme. Optimal compositing for the same budget uses 140 cores composited to 19 analytic samples (SI 1.4).

detecting SOC shifts, for sample sizes of 10, 30, 90, and 200 from each time (Fig. 7). Both Student t-tests appear to have more power than the nonparametric test to detect shifts in TC% at all sample sizes; the stratified Student *t*-test was more powerful than the unstratified test at the same level. (Stratified nonparametric tests would presumably have higher power than the unstratified nonparametric test; they are the subject of ongoing research.) However, comparing Student's *t*-test and the nonparametric test can be misleasing: Student's *t*-test rejects more often than the nonparametric test when the null hypothesis is false, but it also rejects more often than it should when the null hypothesis is true. Student's *t*-test at (nominal) significance level 5 % does not limit the true Type I error rate to 5 % unless the population has a normal distribution. In general, when the population distribution is not normal, the true Type I error rate of Student's *t*-test rate cannot be determined unless the population distribution is known. The power of the nonparametric test improved as the population bounds were tightened and as the level was relaxed: the nonparametric test with 10 % max TC and significance level 0.10 was the most powerful among the nonparametric tests. For example, to have 80 % power to detect a relative change of 20 % from baseline average TC% on rangeland soils requires about 30 samples with

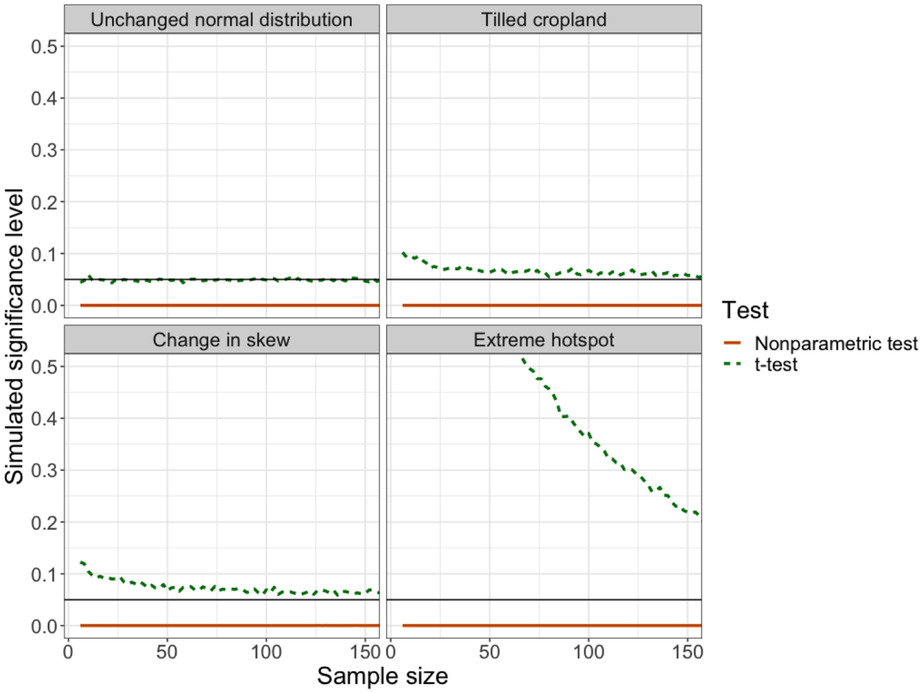


Fig. 6. Simulated significance levels of two nominal 5% level tests: the two-sample Student *t*-test and a nonparametric test. Each panel reflects 5000 simulations at each sample size (x-axis); random samples were drawn independently from each of two distributions that had identical means. The y-axis plots the rate of false rejections of the null hypothesis (the Type I error rate). The solid black line is the nominal 0.05 significance level, which both curves should be at or below.

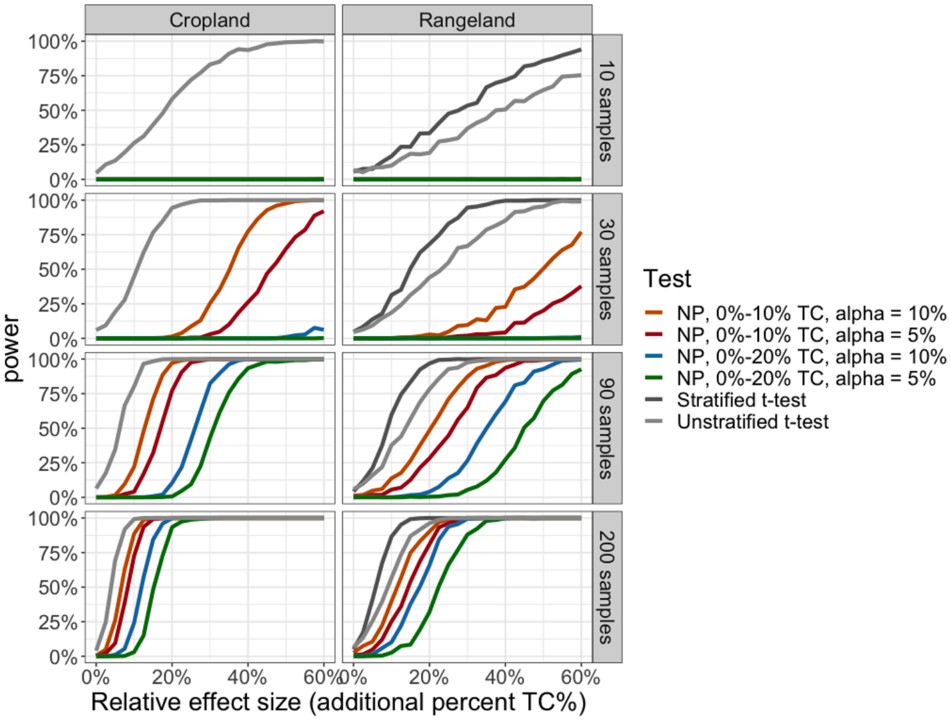


Fig. 7. Simulated power of three two-sample hypothesis tests (Student's *t*-test for unstratified and stratified samples and a nonparametric test for unstratified samples based on (Learned-Miller and Thomas, 2019) to detect relative changes in TC% with sample sizes 10, 30, 90, or 200 from each of two populations. The first population distribution is the empirical distribution of TC% measurements for CROP5 (left column) or the rangeland site (right column) topsoil. The second population distribution is the same as the first, but each value was shifted by 0 % to 60 % of the mean of the first population, 2.7 % TC for the cropland samples and 2.2 % TC for the rangeland samples. Unstratified samples were simple random samples with replacement from the populations. Stratified samples from rangeland were simple random samples with replacement within transects, independent across transects, with sample sizes proportional to the original number of data in each transect. Stratified samples from cropland were not explored because there were no natural strata in the original data. Four curves are presented for the nonparametric test by varying pre-specified upper bounds on the population (10 % TC or 20 % TC) and the nominal significance level (5 % or 10 %); both the Student t-tests used a nominal significance of 5 %, which may understate the chance of false positives. NP = nonparametric.

the stratified Student *t*-test, 90 with the unstratified Student *t*-test, and 200 with the nonparametric test with 10 % max TC and/or a relaxed significance level of 0.1. All tests had more power to detect small changes in cropland soils than in rangeland soils due to lower spatial heterogeneity in croplands. For example, the power of the unstratified Student *t*-test to detect a 10 % change with 90 samples was 80 % for cropland soils but only about 30 % for rangelands.

4. Discussion

4.1. *Crop and rangelands are spatially heterogeneous*

Given the rapid development of C markets, accurate detection and quantification of the impact of management interventions on SOC changes are more important than ever. Our study demonstrates how tailoring sampling and analytical decisions to the high spatial heterogeneity often found in managed lands could improve the reliability and efficiency of SOC sequestration estimates and their associated C credits. As expected, SOC at the rangeland site was more heterogeneous than at the seven cropland sites, with roughly twice as large a CV at every depth (Table 1). This is consistent with other surveys of California rangelands, though differences in depths, spatial scales, and measures of variability limit quantitative comparisons (Carey et al., 2020; Devine et al., 2020; Silver et al., 2010). This is also consistent with the general notion that rangeland SOC is typically more heterogeneous than croplands because of variation in topography, presence of rock fragments, and patchiness of grazing and manure deposition. This makes accurately estimating SOC change on rangelands more challenging.

Deep sampling is critical for making reliable conclusions about C sequestration and greenhouse gas mitigation (Kravchenko and Robertson, 2011; Kuzyakov and Blagodatskaya,2015), especially since SOC gains near the surface may be offset by losses at depth (Poffenbarger et al., 2020; Slessarev et al., 2021; Syswerda et al., 2011; Tautges et al., 2019). The CV of SOC in our study tended to increase with depth, while standard deviations decreased. Hence, a given relative change (e.g., 10 % gain from baseline TC%) is harder to detect in subsurface soils than in topsoil, but a given absolute change (e.g., 0.5 TC% gain) may be easier to detect. Since detecting an equivalent absolute change in the subsurface requires fewer samples, topsoil heterogeneity should generally guide decisions around sample size.

Though we have emphasized TC% measurements, high variability of BD within sites (Fig. 2), especially in rangelands, contributes additional uncertainty to SOC stock estimates (Walter et al., 2016). Even where TC % is relatively homogeneous, variability in BD could lead to large uncertainties in estimates of SOC stocks and of SOC stock changes, and ultimately prevent the reliable detection of these changes (Slessarev et al., 2021). Failing to account for BD variability (e.g., treating BD estimates as fixed) underestimates uncertainty and may lead to erroneous conclusions about SOC stock change. See **SI 4.1.** for further discussion on combining BD and TOC% uncertainties for SOC stock estimates.

4.2. *Analytical variability contributes little to measurement error*

Variability in assay measurements contributed far less to measurement error than spatial heterogeneity of SOC, except when samples were highly composited (Fig. 4). Replicated measurements show that both the soliTOC and ECS 4010 analyzers have estimated median relative error below 0.07.

TC%, however, differed substantially between instruments for samples with high TIC%, (see also SI Fig. 3). re-analysis on the ECS 4010 after TIC removal improved agreement between TC% measurements on the ECS 4010 and TOC% on the soliTOC, indicating that TIC should be removed when using elemental analyzers like the ECS 4010. Larger sample masses (~10–20x the mass of traditional dry combustion instruments) may explain the higher precision of the soliTOC, which had about one-third the median relative error of the ECS 4010 (Fig. 3). Larger analytical subsamples should better represent the entire sample and reduce variability inherent to small subsamples. In the case of the ECS 4010, increased analytical replication may be necessary. SOC monitoring schemes could mitigate analytical error by using the same instrument, ideally in the same lab, for repeated analysis, and by including standards with comparable amounts of TIC, when analyzing samples known to contain TIC.

4.3. *Measurement protocol recommendations to reduce uncertainty*

Spatial heterogeneity is likely to dominate measurement uncertainty in many scenarios. We recommend three ways to for measurement protocols to reduce uncertainty in SOC estimates and increase the reliability of C credits: provide stratified sampling guidance, minimize compositing, and, most importantly, require larger sample sizes.

Stratification on variables such as catenal position, soil type, topography and historical management can increase the power of detecting SOC sequestration and generally reduces uncertainty for a given total sample size on heterogeneous landscapes (Devine et al., 2020; deGruijter et al., 2016). Our simulations provide further evidence that stratification can be a useful sampling strategy: stratified sampling had higher power to detect increases in TC% at the rangeland site than simple random sampling (Fig. 7). Without stratification, far larger sample sizes are required to reliably detect and quantify SOC changes. While current protocols such as Climate Action Reserve's (CAR) Soil Enrichment Protocol and Verra's VM0021 allow and encourage stratification, they do not provide straightforward and quantitative stratification guidance. Preliminary field surveys, geospatial information regarding soil and landscape features, and expert pedological knowledge are useful for defining strata in research settings (Post et al., 2001). C market protocols should look to incorporate algorithmic stratification (Devine et al., 2020; deGruijter et al., 2016), digital soil mapping, and user-friendly software tools (e.g. Stratifi; https://www.quickcarbon.org/tools), to help standardize and ease barriers to stratification for SOC measurement.

Compositing can be optimized to minimize uncertainty within a cost budget, given estimates of the analytical precision, spatial heterogeneity, and the (marginal) unit cost of collecting, preparing, and analyzing a sample (Spertus,2021). Without such estimates, it is best to avoid compositing (especially when collecting baseline samples), because it reduces information on spatial heterogeneity, complicates sampling designs and analyses, and increases the contribution of analytical error (Fig. 5). Compositing also tends to reduce power by decreasing the effective sample size. Compositing is most helpful when SOC is highly heterogeneous, the cost of each laboratory assay is high, and the budget is small. In such cases, investigators should consider optimal, rather than full compositing (Spertus,2021). We've developed a web app for investigators (including for use in soil C measurement protocols) to help determine optimal compositing schemes, which is accessible at: https://scf.berkeley.edu/shiny/bosf/soil-carbon-statistics/.

Finally, many current sampling designs for the sale of C credits use sample sizes that are too small to allow any statistical test to have a reasonable chance of detecting moderate changes in SOC (Necpalova et al., 2014) or quantifying SOC changes on heterogeneous landscapes on relevant timescales. To illustrate, assume that compost application on rangelands increases relative TC% by 20 % (as per (Ryals et al., 2014) after 3 years of application). Based on the spatial heterogeneity we observed in rangeland soils, in order to have 80 % power to detect such an increase (using stratified sampling and Student's *t*-test) would require collecting and analyzing nearly 100 soil samples at baseline and another 100 samples after the compost was applied, with no compositing (Fig. 7).

Most rangeland management interventions, however, such as improved grazing practices, are expected to produce much smaller C gains. For instance, (Conant et al., 2017) found a relative increase of ~ 10 % from grazing improvements. The smaller the anticipated change in SOC, the larger the sample size must be to reliably detect and quantify the change. Similarly, using nonparametric tests—which may be needed to properly control the false positive rate—require larger sample sizes. For instance, it would require more than 200 samples to have an 80 % chance of detecting a 10 % relative increase in SOC using either the unstratified Student *t*-test or the nonparametric test (Fig. 7). No matter the sampling design or statistical test, the sample sizes typical in current campaigns and protocols (e.g., 8 samples for USDA GRACEnet; 9 samples composited to 1 for CDFA Healthy Soils Program; minimum of 3 samples for the Australian Carbon Credits Methodology; Davis et al., 2017) are far too small to have sufficient power to detect and quantify changes in rangeland SOC (Fig. 7).

Our simulations suggest that detecting SOC changes in croplands may be easier than rangelands, but common sample sizes are still inadequate. Nonparametric tests have little chance of detecting reasonable changes with only 10 samples, and Student's *t*-test is likely to be misleading for such small samples and to lack sufficient power to detect realistic changes. With only 10 cropland samples, a relative increase of 30 %—a very large change—would be needed for Student's *t*-test to have 80 % power (Saby et al., 2008). At the CROP5 site, Student's *t*-test required about 90 samples to have an 80 % chance of detecting a 10 % relative change in TC%.

Given that sampling campaigns are routinely underpowered, we suggest a priori power analyses to determine site-specific minimum sample sizes (Kravchenko and Robertson, 2011). This could include conducting a power analysis either by collecting and analyzing reconnaissance samples, or with regional and relevant spatial heterogeneity information (e.g., from prior studies or soil survey information). We also suggest routinely conducting post hoc power analyses to determine whether studies that find no effect of management on SOC had sufficient power to detect expected differences.

4.4. Tests must be valid to provide credible evidence of carbon change

Even when sampling is well-designed and executed, statistical analysis matters. Student's *t*-test and its relatives may erroneously conclude C was sequestered when it was not, at a much higher rate than the nominal significance level. As shown in Fig. 6, this occurs when even one of the TC distributions is skewed. The false positive rate for Student's *t*-test is particularly high when there are SOC hotspots or the distribution of TC (but not its mean) changes over time. Some management interventions redistribute SOC and create or destroy C hotspots (Baker et al., 2007; Kuzyakov and Blagodatskaya, 2015; Marin-Spiotta et al., 2014). For example, establishing perennial intercrops or hedgerows and spreading high-C inputs such as biochar and compost can create SOC hotspots. Valid inference is crucial to measure SOC sequestration credibly; Student's *t*-test and related tests and confidence intervals likely often understate the chance of false positives and have an inordinately large chance of false negatives.

How can monitoring and verification campaigns ensure that estimates and inferences are reliable? An important consideration is whether the soil population of interest might have skewed SOC, including from SOC hotspots. If so, it might be possible to stratify the sample so that SOC distributions within strata are not severely skewed. Skewness in the population distribution makes Student's *t*-test behave particularly poorly. While transformations (e.g. logarithmic) are possible, skewness in the population that can undermine parametric statistical inferences may not be evident for realistic sample sizes. Larger sample sizes improve the approximations Student's *t*-test relies on, but in general, it is not possible to determine how large the sample must be for the approximation to have a particular level of accuracy (Cochran, 1977).

If hotspots might exist but their locations are unknown prior to sampling, Student's *t*-test should not be used. In our simulations, nonparametric tests were less powerful than Student's *t*-test, but they control the false positive rate for every SOC distribution (Figs. 6 and 7), while Student's *t*-test can fail for some SOC distributions. Thus, Student's *t*-test may *appear* more powerful, but it is wrong more often. Our simulations show that using prior geochemical knowledge to bound TC more tightly (e.g. 10 % instead of 20 %) can increase the power of nonparametric tests, as can testing at a higher significance level (e.g. 10 % instead of 5 %). Deriving more powerful nonparametric tests is an active research area in Statistics (Romano and Wolf, 2000; Waudby-Smith and Ramdas, 2020), which we hope to extend to stratified soil samples (e.g., (Wendell and Schmee, 1996)). We have written an R package to facilitate wider use of nonparametric tests, which can be installed from the R console by running devtools::install_github("spertus/nptests").

4.5. Study limitations and future research

Our analyses relied on soil samples that were collected using common approaches, rather than the sampling protocols we recommend here (systematic rather than random samples). This could understate overall spatial heterogeneity, making our findings conservative, if SOC is spatially autocorrelated. However, we found little evidence of spatial autocorrelation in our rangeland samples (SI Figs. 5–9). These simulations are a starting point; other changes to SOC distributions and deeper soil depths should be examined. The geographic extent of the soil sampling was also limited and thus does not fully represent the heterogeneity of croplands and rangelands worldwide, but we expect the qualitative differences in heterogeneity between them will be more broadly applicable.

4.6. Broader implications for research, C markets, and policy

There have been numerous calls to standardize protocols for measuring SOC (Bispo et al., 2017; Davis et al., 2017; Jandl et al., 2014), but complete standardization may not be practical given differences among project needs and budgets, landscape heterogeneity, and lab constraints. In particular, given the large contribution of spatial heterogeneity to the uncertainty of SOC estimates, protocols that require fixed sample sizes or generate C credits on the basis of a fixed, small, minimum number of samples are not appropriate. For instance, sampling designs optimized to detect SOC changes for croplands may have little chance of detecting similar changes on rangelands, which typically require larger samples because they are more heterogeneous. Instead, *sampling design processes* should be standardized, such as the use of algorithmic stratification and *a priori* power analyses to select sample sizes adequate to detect plausible changes. In the case of C markets, verifiers should ensure that the sample size was adequate to detect and quantify SOC sequestration prior to generating and selling C offsets.

The consequences of inaccurate estimates of SOC for C markets are large. Current verification sampling protocols used to quantify and generate C credits for C markets cannot reliably estimate SOC sequestration, especially on heterogeneous agricultural lands. This could result in SOC offsets having little connection to the true extent of sequestration (Jackson Hammond et al., 2021). Verification protocols for croplands include Climate Action Reserve's (CAR) Soil Enrichment Protocol, Verra's VM0021, Australia's Carbon Credits Methodology (ACCM), and the Food and Agriculture Organization's (FAO) Global Soil Organic Carbon (GSOC). All four protocols require a minimum of only three or more samples per stratum, far fewer than required to estimate the impact of management changes on a timescale of years. Some—though not all—protocols also lack details on how to stratify and analyze the resulting data to estimate SOC stocks and stock changes. Especially because they sanction such small sample sizes, these protocols may often reward "false positives" and fail to reward genuine sequestration. We recommend revising each of these protocols to require substantially more samples tailored to land-specific spatial heterogeneity, and, following ACCM as an example, provide much more detailed and useful guidelines for participants on when, where, and how to sample to minimize uncertainties. While governments, companies, and society must decide what level of confidence suffices to demonstrate SOC sequestration (e.g. the ACCM accepts SOC sequestration with only 60 % confidence, to encourage participation), protocols must actually be able to deliver that level of confidence.

For some purposes, instead of estimating SOC stock changes on *each* participating farm or ranch, it might suffice to estimate the aggregate change across many farms/ranches, collecting few samples from each, to minimize costs. Alternatively, one might conduct intensive sampling on a random sample of sites or a network of regional research monitoring sites (which could be supported by the development of funding programs like AgARDA or through increases in funding to LTERs or Climate Hub networks). Limiting sampling efforts to a smaller number of dedicated sites representing a range of climates, soil types, and cropping systems could allow for more intensive sampling—with higher power to detect SOC stock changes. This intensive sampling could then be used to calibrate, validate, and improve models such as MEMS 2.0 (Microbial Efficiency-Matrix Stabilization) (Zhang et al., 2021) that can estimate SOC change across broader landscapes and generate SOC credits for similar farms and ranches. This may be a more efficient use of resources and could drive more accurate verification in the long-term. However, both strategies represent a shift from paying for *results* to paying for *practices* that are expected—but not guaranteed—to produce results.

5. Conclusions

Spatial heterogeneity of SOC is a primary obstacle to accurately measuring changes in SOC stocks, even with careful sampling design and execution, accurate assays, and rigorous statistical analysis. Attempting to measure or verify SOC sequestration using too few samples, poor sampling design, imprecise laboratory instruments, or inappropriate statistical analysis can undermine climate change mitigation goals. We highlighted errors, quantified uncertainties, and demonstrated potential improvements in design and analysis, using data from California croplands and rangelands. There are several straightforward ways that sampling schemes can be improved, especially for C markets. Collecting information on the degree and pattern of heterogeneity before a comprehensive sampling campaign can make it possible to use stratified sampling to advantage. Such information also makes it possible to perform power calculations and identify optimal compositing approaches, ensuring that the campaign has sufficient statistical power to detect anticipated changes in. In general, reliable inferences about the short-term effect of management interventions on soil C require larger sample sizes and less compositing than is commonly used. We demonstrate that Student’s *t*-test has highly inflated false-positive rates in scenarios that may be common in the field and suggest caution when using Student’s t-tests and its relatives for verifying changes, especially when sample sizes are small. Nonparametric statistical methods can control false positives for any sample size, without assumptions about SOC distributions; providing more reliable, trustworthy results. The power of nonparametric tests can be increased using transparent, verifiable assumptions (e.g. geochemical constraints on the maximum SOC). Careful planning and continued collaboration between soil scientists and statisticians will help improve accuracy and precision of SOC measurements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Jacob Spertus’ work was supported by the United States National Science Foundation Graduate Research Fellowship Program (Grant number: DGE 1752814). We are grateful to Garrett Long, who was an invaluable part of our rangeland sampling crew; Greg Richardson (Director of Monitoring and Research at Paicines Ranch), who coordinated and helped sample at Paicines Ranch; Ky Cooper (Facilities and Welding at Paicines Ranch), who graciously lent us equipment and helped auger hundreds of samples; and Ky’s dog Pepper, who kept us company in the field.

Data availability and code

All data and code used in this paper are available online at: https://github.com/spertus/soil-carbon-statistics. A web application to facilitate investigators with pre-sample planning, including *a priori* power analyses, determining optimal compositing schemes, and budget planning, is available at: https://scf.berkeley.edu/shiny/bosf/soil-carbon-statistics/. Our R package to facilitate wider use of nonparametric tests can be installed from the R console by running devtools::install_github(“spertus/nptests”).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2022.116323.

References

Anderson, T.W., 1969. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. US Dept of the Navy. https://doi.org/10.21236/AD0696676.

Bai, X., Huang, Y., Ren, W., Coyne, M., Jacinthe, P.-A., Tao, B., Hui, D., Yang, J., Matocha, C., 2019. Responses of soil carbon sequestration to climate-smart agriculture practices: A meta-analysis. Glob. Chang. Biol. 25, 2591–2606. https://doi.org/10.1111/gcb.14658.

Baker, J.M., Ochsner, T.E., Venterea, R.T., Griffis, T.J., 2007. Tillage and soil carbon sequestration—What do we really know? Agric. Ecosyst. Environ 118, 1–5. https://doi.org/10.1016/j.agee.2006.05.014.

Beem-Miller, J.P., Kong, A.Y.Y., Ogle, S., Wolfe, D., 2016. Sampling for soil carbon stock assessment in rocky agricultural soils. Soil Sci. Soc. Am. J. 80, 1411–1423. https://doi.org/10.2136/sssaj2015.11.0405.

Bispo, A., Andersen, L., Angers, D.A., Bernoux, M., Brossard, M., Cécillon, L., Comans, R.N.J., Harmsen, J., Jonassen, K., Lamé, F., Lhuillery, C., Maly, S., Martin, E., Mcelnea, A.E., Sakai, H., Watabe, Y., Eglin, T.K., 2017. Accounting for Carbon Stocks in Soils and Measuring GHGs Emission Fluxes from Soils: Do We Have the Necessary Standards? Front. Environ. Sci. 5 https://doi.org/10.3389/fenvs.2017.00041.

Brus, D.J., de Gruijter, J.J., 2011. Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data. Geoderma 164, 172–180. https://doi.org/10.1016/j.geoderma.2011.06.001.

Carey, C.J., Weverka, J., DiGaudio, R., Gardali, T., Porzig, E.L., 2020. Exploring variability in rangeland soil organic carbon stocks across California (USA) using a voluntary monitoring network. Geoderma Regional 22, e00304.

Chappell, A., Sanderman, J., Thomas, M., Read, A., Leslie, C., 2012. The dynamics of soil redistribution and the implications for soil organic carbon accounting in agricultural south-eastern Australia. Glob Change Biol 18, 2081–2088. https://doi.org/10.1111/j.1365-2486.2012.02682.x.

Chatterjee, A., Lal, R., Wielopolski, L., Martin, M.Z., Ebinger, M.H., 2009. Evaluation of different soil carbon determination methods. CRC Crit. Rev. Plant Sci. 28, 164–178. https://doi.org/10.1080/07352680902776556.

Cochran, W., 1977. Sampling Techniques. John Wiley & Sons.

Conant, R.T., Cerri, C.E.P., Osborne, B.B., Paustian, K., 2017. Grassland management impacts on soil carbon stocks: a new synthesis. Ecol. Appl. 27, 662–668. https://doi.org/10.1002/eap.1473.

Davis, M., Alves, B., Karlen, D., Kline, K., Galdos, M., Abulebdeh, D., 2017. Review of soil organic carbon measurement protocols: A US and Brazil comparison and recommendation. Sustainability 10, 53. https://doi.org/10.3390/su10010053.

de Gruijter, J.J., McBratney, A.B., Minasny, B., Wheeler, I., Malone, B.P., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130. https://doi.org/10.1016/j.geoderma.2015.11.010.

Devine, S.M., O’Geen, A.T., Liu, H., Jin, Y., Dahlke, H.E., Larsen, R.E., Dahlgren, R.A., 2020. Terrain attributes and forage productivity predict catchment-scale soil organic carbon stocks. Geoderma 368, 114286. https://doi.org/10.1016/j.geoderma.2020.114286.

Ellert, B.H., Janzen, H.H., Entz, T., 2002. Assessment of a method to measure temporal change in soil carbon storage. Soil Sci. Soc. Am. J. 66, 1687–1695. https://doi.org/10.2136/sssaj2002.1687.

Franzluebbers, A., 2005. Soil organic carbon sequestration and agricultural greenhouse gas emissions in the southeastern USA. Soil Tillage Res. 83, 120–147. https://doi.org/10.1016/j.still.2005.02.012.

Goidts, E., Van Wesemael, B., Crucifix, M., 2009. Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. Eur. J. Soil Sci. 60, 723–739. https://doi.org/10.1111/j.1365-2389.2009.01157.x.

Gosnell, H., Robinson-Maness, N., Charnley, S., 2011. Profiting from the sale of carbon offsets: A case study of the trigg ranch. Rangelands 33, 25–29. https://doi.org/10.2111/1551-501X-33.5.25.

Homann, P.S., Sollins, P., Fiorella, M., Thorson, T., Kern, J.S., 1998. Regional soil organic carbon storage estimates for western oregon by multiple approaches. Soil Sci. Soc. Am. J. 62, 789–796. https://doi.org/10.2136/sssaj1998.03615995006200030036x.

Jackson Hammond, A.A., Motew, M., Brummitt, C.D., DuBuisson, M.L., Pinjuv, G., Harburg, D.V., Campbell, E.E., Kumar, A.A., 2021. Implementing the soil enrichment protocol at scale: opportunities for an agricultural carbon market. Front. Clim. 3 https://doi.org/10.3389/fclim.2021.686440.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D.D., Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. Sci. Total Environ. 468–469, 376–383. https://doi.org/10.1016/j.scitotenv.2013.08.026.

Jones, D.L., Rousk, J., Edwards-Jones, G., DeLuca, T.H., Murphy, D.V., 2012. Biochar-mediated changes in soil quality and plant growth in a three year field trial. Soil Biol. Biochem. 45, 113–124. https://doi.org/10.1016/j.soilbio.2011.10.012.

Kravchenko, A.N., Robertson, G.P., 2011. Whole-Profile Soil Carbon Stocks: The Danger of Assuming Too Much from Analyses of Too Little. Soil Sci. Soc. Am. J. 75 (1), 235–240.

Kuzyakov, Y., Blagodatskaya, E., 2015. Microbial hotspots and hot moments in soil: Concept & review. Soil Biol. Biochem. 83, 184–199. https://doi.org/10.1016/j.soilbio.2015.01.025.

Learned-Miller, E., Thomas, P., 2019. A New Confidence Interval for the Mean of a Bounded Random Variable. University of Massachusetts.

Lehmann, J., Kinyangi, J., Solomon, D., 2007. Organic matter stabilization in soil microaggregates: implications from spatial heterogeneity of organic carbon contents and carbon forms. Biogeochemistry 85, 45–57. https://doi.org/10.1007/s10533-007-9105-3.

Lehmann, E.L., Romano, J.P., 2010. Testing Statistical Hypotheses (Springer Texts in Statistics), 3rd ed. Springer.

Majumder, S., Neogi, S., Dutta, T., Powel, M.A., Banik, P., 2019. The impact of biochar on soil carbon sequestration: Meta-analytical approach to evaluating environmental and economic advantages. J. Environ. Manage. 250, 109466 https://doi.org/10.1016/j.jenvman.2019.109466.

Marin-Spiotta, E., Chaopricha, N.T., Plante, A.F., Diefendorf, A.F., Mueller, C.W., Grandy, A.S., Mason, J.A., 2014. Long-term stabilization of deep soil carbon by fire and burial during early Holocene climate change. Nature Geosci. 7, 428–432. https://doi.org/10.1038/ngeo2169.

Miller, B.A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K., Wehrhan, M., Sommer, M., 2016. Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. Soil Tillage Res. 156, 194–208. https://doi.org/10.1016/j.still.2015.07.004.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.

Mishra, U., Riley, W.J., 2015. Scaling impacts on environmental controls and spatial heterogeneity of soil organic carbon stocks. Biogeosciences 12, 3993–4004. https://doi.org/10.5194/bg-12-3993-2015.

Natali, C., Bianchini, G., Carlino, P., 2020. Thermal stability of soil carbon pools: Inferences on soil nature and evolution. Thermochim. Acta 683, 178478. https://doi.org/10.1016/j.tca.2019.178478.

Necpalova, M., Anex, R.P., Kravchenko, A.N., Abendroth, L.J., Del Grosso, S.J., Dick, W. A., Helmers, M.J., Herzmann, D., Lauer, J.G., Nafziger, E.D., Sawyer, J.E., Scharf, P. C., Strock, J.S., Villamil, M.B., 2014. What does it take to detect a change in soil carbon stock? A regional comparison of minimum detectable difference and experiment duration in the north central United States. J. Soil Water Conserv. 69, 517–531. https://doi.org/10.2489/jswc.69.6.517.

O' Rourke, S.M., Holden, N.M., 2011. Optical sensing and chemometric analysis of soil organic carbon - a cost effective alternative to conventional laboratory methods? Soil Use Manage. 27, 143–155. https://doi.org/10.1111/j.1475-2743.2011.00337.x.

Oldfield, E.E., Eagle, A.J., Rudek, R.L., Sanderman, J., Gordon, D.R., 2021. Agricultural soil carbon credits: Making sense of protocols for carbon sequestration and net greenhouse gas removals. Environmental Defense Fund, New York, New York.

Pesarin, F., Salmaso, L., 2010. In: Permutation tests for complex data. John Wiley & Sons Ltd, Chichester, UK. https://doi.org/10.1002/9780470689516.

Poffenbarger, H.J., Olk, D.C., Cambardella, C., Kersey, J., Liebman, M., Mallarino, A., Six, J., Castellano, M.J., 2020. Whole-profile soil organic matter content, composition, and stability under cropping systems that differ in belowground inputs. Agric. Ecosyst. Environ 291, 106810. https://doi.org/10.1016/j.agee.2019.106810.

Post, W.M., Izaurralde, R.C., Mann, L.K., Bliss, N., 2001. Monitoring and verifying changes of organic carbon in soil. In: Rosenberg, N.J., Izaurralde, R.C. (Eds.), Storing CArbon in AgriculturAl Soils: A Multi-Purpose EnvironmentAl StrAtegy. Springer, Netherlands, Dordrecht, pp. 73–99. https://doi.org/10.1007/978-94-017-3089-1_4.

Robertson, G.P., Klingensmith, K.M., Klug, M.J., Paul, E.A., Crum, J.R., Ellis, B.G., 1997. Soil resources, microbial activity, and primary production across an agricultural ecosystem. Ecol. Appl. 7, 158–170. https://doi.org/10.1890/1051-0761(1997)007[0158:SRMAAP]2.0.CO;2.

Romano, J.P., Wolf, M., 2000. Finite sample nonparametric inference and large sample efficiency. Ann. Statist. 28 https://doi.org/10.1214/aos/1015951997.

Ryals, R., Kaiser, M., Torn, M.S., Berhe, A.A., Silver, W.L., 2014. Impacts of organic matter amendments on carbon and nitrogen dynamics in grassland soils. Soil Biol. Biochem. 68, 52–61. https://doi.org/10.1016/j.soilbio.2013.09.011.

Saby, N.P.A., Bellamy, P.H., Morvan, X., Arrouays, D., Jones, R.J.A., Verheijen, F.G.A., Kibblewhite, M.G., Verdoodt, A.N.N., Üveges, J.B., Freudenschuß, A., Simota, C., 2008. Will European soil-monitoring networks be able to detect changes in topsoil organic carbon content? Glob. Chang. Biol. 14, 2432–2442. https://doi.org/10.1111/j.1365-2486.2008.01658.x.

Sanderman, J., Baldock, J.A., 2010. Accounting for soil carbon sequestration in national inventories: a soil scientist's perspective. Environ. Res. Lett. 5 (3) https://doi.org/10.1088/1748-9326/5/3/034003, 034003.

Silver, W.L., Ryals, R., Eviner, V., 2010. Soil carbon pools in California's annual grassland ecosystems. Rangeland Ecol. Manage. 63, 128–136. https://doi.org/10.2111/REM-D-09-00106.1.

Slessarev, E., Zelikova, J., Hamman, J., Cullenward, D., Freeman, J., 2021. Depth matters for soil carbon accounting. CarbonPlan.

Spertus, J.V., 2021. Optimal sampling and assay for estimating soil organic carbon. OJSS 11, 93–121. https://doi.org/10.4236/ojss.2021.112006.

Stark, P.B., 2009. Risk-Limiting Postelection Audits: Conservative P-Values from Common Probability Inequalities. IEEE Trans. Inform. Forensic Secur. 4, 1005–1014. https://doi.org/10.1109/TIFS.2009.2034190.

Stark, P.B., 2023. ALPHA: Audit that Leverages Previously Hand-Audited Ballots. Ann. Appl. Stat. in press.

Syswerda, S.P., Corbin, A.T., Mokma, D.L., Kravchenko, A.N., Robertson, G.P., 2011. Agricultural management and soil carbon storage in surface vs. deep layers. Soil Sci. Soc. Am. J. 75 (1), 92–101.

Tautges, N.E., Chiartas, J.L., Gaudin, A.C.M., O'Geen, A.T., Herrera, I., Scow, K.M., 2019. Deep soil inventories reveal that impacts of cover crops and compost on soil carbon sequestration differ in surface and subsurface soils. Glob. Chang. Biol. 25, 3753–3766. https://doi.org/10.1111/gcb.14762.

Walter, K., Don, A., Tiemeyer, B., Freibauer, A., 2016. Determining soil bulk density for carbon stock calculations: A systematic method comparison. Soil Sci. Soc. Am. J. 80, 579–591. https://doi.org/10.2136/sssaj2015.11.0407.

Waudby-Smith, I., Ramdas, A., 2020. Estimating means of bounded random variables by betting. arXiv:2010.09686.

Wendell, J.P., Schmee, J., 1996. Exact Inference for Proportions from a Stratified Finite Population. J. Am. Stat. Assoc. 91, 825–830. https://doi.org/10.1080/01621459.1996.10476950.

Yan, X., Cai, Z., Wang, S., Smith, P., 2011. Direct measurement of soil organic carbon content change in the croplands of China. Glob. Chang. Biol. 17, 1487–1496. https://doi.org/10.1111/j.1365-2486.2010.02286.x.

Zhang, Y., Lavallee, J.M., Robertson, A.D., Even, R., Ogle, S.M., Paustian, K., Cotrufo, M. F., 2021. Simulating measurable ecosystem carbon and nitrogen dynamics with the mechanistically-defined MEMS 2.0 model. doi:10.5194/bg-2020-493.