A GlobalSoilMap prototype for the Conterminous United States: Characterizing Uncertainty

Norman B. Bliss
Volunteer, U.S. Geological Survey (USGS), Earth Resources Observation and Science (EROS) Center,
Sioux Falls, South Dakota, USA, E-mail bliss@usgs.gov https://orcid.org/0000-0003-2409-5211

Abstract
The legacy soil survey geographic databases for the United States have been processed to match the standards of the GlobalSoilMap. The source data are primarily from the detailed Soil Survey Geographic (SSURGO) dataset, and areas that are not yet mapped by SSURGO are filled from the General Soil Map (STATSGO2). These legacy databases were developed by the U.S. National Cooperative Soil Survey under the leadership of the U.S. Department of Agriculture, Natural Resources Conservation Service. The source data structure includes map units, which are delineated spatially on a map, and components, which have attribute information representing a percentage of the map unit but are not mapped. Although the 'representative value' for each quantitative attribute was used to calculate the values used for the GlobalSoilMap results, the source databases also include 'low' and 'high' values that can be used to represent the uncertainty. Although the standards for coding 'low' and 'high' values were not well defined, by assuming that they represent a prediction interval at the 90% confidence level, it is possible to estimate the uncertainty for each component. Each component may have a distinct range of 'low' and 'high' values, but the goal is to have an uncertainty measure representing the map unit. I developed a method to create 1000 'pseudo-observations' for the map unit, distributed according to the prediction interval for each component, from which it is possible to determine the 90% or 95% prediction interval for the map unit. This method makes it possible to merge component records with differing prediction intervals into a single map unit prediction interval. Not all data records are populated with the 'low' and 'high' values. For example, in the attribute representing the percentage of sand, only 115,274 of the 309,916 map units (37%) had both 'low' and 'high' values. The size of the prediction interval at the 90% confidence interval varies considerably for these records. Three methods of estimating the prediction interval at the 90% confidence level are discussed in this poster, and illustrated with soil organic carbon data as an example for the entire process. The pseudo-observation method is also compared with other factors influencing uncertainty: the original scale of mapping and the area represented by the map unit. Providing uncertainty estimates will fulfill a requirement of the GlobalSoilMap specifications and enable these data to be distributed, leading to improved decision making by farmers and resource managers.

Introduction
A map of the soil organic carbon in the Conterminous United States is shown in Figure 1. Figure 2 shows that the distribution is non-linear, such that the more arid half of the land area has about 20% of the SOC, whereas the wetter half of the land area has about 80% of the SOC.

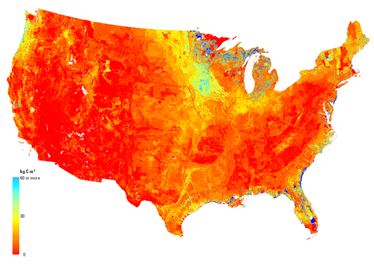


Figure 1. A map of the soil organic carbon in the conterminous United States, using SSURGO data where available and filling with STATSGO2 data in other locations (generally large holdings of government-owned land in the Western United States). Units are kilograms of carbon per square meter for the total profile, and the plot uses standard deviation scaling so extreme values are more evident.

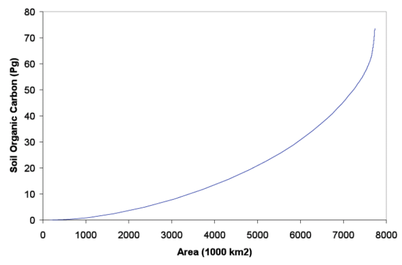


Figure 2. The non-linear distribution of soil organic carbon in the conterminous United States. Wetlands are represented at the right-edge of the curve with high carbon densities.

The SSURGO and STATSGO2 data structure
The data structure for SSURGO and STATSGO2 is shown in Figure 3. An example of the Organic Matter (OM) and bulk density data used for calculating SOC is shown for one map unit in Figure 4.

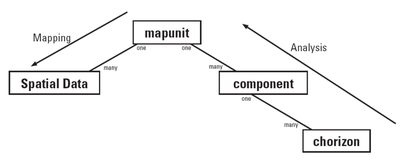


Figure 3. The hierarchical structure of attribute data in SSURGO and STATSGO2. Each map unit can have one or more components representing the distinct soils present but not spatially represented on the map, with the proportion of land area represented as a component percentage. Each component can have zero or more horizons, representing changes in soil properties with depth in the profile.

mapunit	component		chorizon							
mukey	cokey	comppct_r	chkey	hzdept_r	hzdepb_r	om_l	om_r	om_h	dbthirdbar_r	
10001	15001	70	17001	0	15	4.00	5.00	6.00	1.10	
10001	15001	70	17002	15	30	1.00	1.50	2.00	1.35	
10001	15001	70	17003	30	152	0.05	0.06	0.07	1.45	
10001	15002	30	17004	0	25	2.00	3.00	3.50	1.30	
10001	15002	30	17005	25	75	0.50	0.70	0.90	1.40	
10001	15002	30	17006	75	75	NoData	NoData	NoData	NoData	bedrock

Figure 4. A hypothetical data record showing information used for the SOC calculation for one map unit. Organic matter percentage (om_r) is multiplied by .0058, the horizon thickness (cm) and by the bulk density to convert units to grams of carbon per square centimeter of soil. For the GlobalSoilMap calculation the mass of soil is also calculated and the SOC is represented as g SOC per kilogram of soil fines (< 2 mm particle size).

The analysis works from the chorizon table up through the component level, to calculate a measure for each variable at the mapunit level. This is done for each GlobalSoilMap variable by each depth zone for the low (_l), representative (_r), and high (_h) values. Values are weighted by the mass of soil fines and the component percentage (comppct_r) when accumulating to the tables higher in the hierarchy.

Methods
Three methods of representing uncertainty are compared in this poster:
1) The limits of the prediction interval are estimated using a low (_l) and high (_h) value calculated as a weighted average of the low and high values of each of the components for which low and high estimates were available. The mass of soil fines is used as the weighting factor.
2) The limits of the prediction interval are estimated by taking the "low of the low" (_ll) and the "high of the high" (_hh) among all components in the map unit. This method was found to be reasonable based on a sample of map units by Helmick et al. (2014). The method selects the minimum from among the low values of all components (_ll) and the maximum from among the high values all components (_hh).
3) The method of "pseudo-observations" proposed here.

The pseudo-observation method
The term "pseudo-observation" is used because we do not have thousands of real observations for each map unit. A normal distribution is fit to the low (_l) and high (_h) values of each variable, using the assumption that these represent the prediction interval limits at a 90% confidence level. Figure 5 shows how the area under the curve can be divided into zones with identifiable proportions of the population of observations.

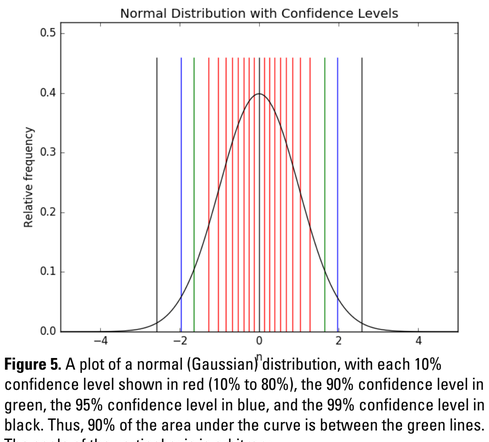


Figure 5. A plot of a normal (Gaussian)[1] distribution, with each 10% confidence level shown in red (10% to 80%), the 90% confidence level in green, the 95% confidence level in blue, and the 99% confidence level in black. Thus, 90% of the area under the curve is between the green lines. The scale of the vertical axis is arbitrary.

Steps in the pseudo-observation method:
1) Assume an input data prediction interval confidence level (here 90%).
2) Fit a normal curve using the low (_l) and high (_h) values for a component.
3) Select 1000 pseudo-observations to closely approximate the normal curve.
4) Randomly select from the component's pseudo-observations in proportion to the component percentage (comppct_r).
5) Pool the selections from all the components into a single list of 1000 pseudo-observations, and sort the values.
6) Pick the values at locations 50 and 950 from the sorted list to represent the lower and upper limits of the prediction interval at the 90% confidence level (5% of the values are in each tail of the distribution)

Figure 6 shows three alternative assumptions for the confidence level of the input data, although only the 90% confidence level was used in this study.

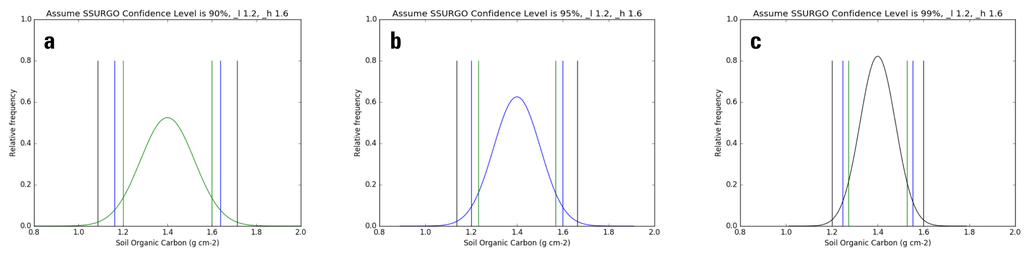
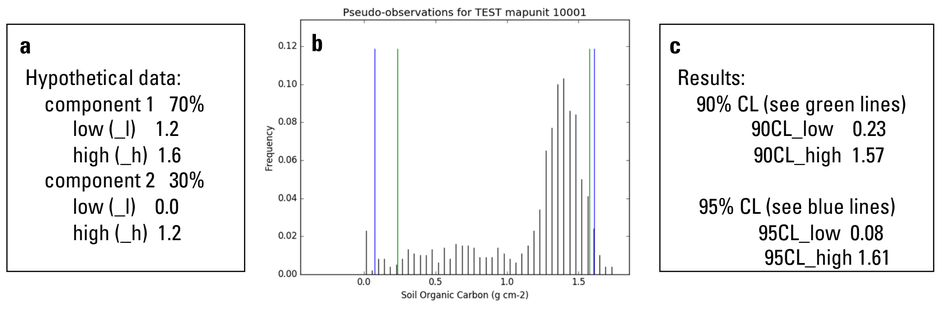


Figure 6. (a) A 90% confidence level is assumed for the input SSURGO or STATSGO2 data. The green curve represents the fit of a normal (Gaussian) curve to a hypothetical data record with a prediction interval "low" value of 1.2 and a "high" value of 1.6. (b) The distribution (blue curve) from the same points, assuming that the source data low and high prediction interval limits represent a 95% confidence level. (c) Similarly, assuming a 99% input data confidencde level. The scale of the vertical axis is arbitrary. The 90% assumption is used for the remainder of this poster (as in (a)).

Methods (continued)
Steps in the pseudo-observation method:
1) Assume an input data prediction interval confidence level (here 90%).
2) Fit a normal curve using the low (_l) and high (_h) values for a component.
3) Select 1000 pseudo-observations to closely approximate the normal curve.
4) Randomly select from the component's pseudo-observations in proportion to the component percentage (comppct_r).
5) Pool the selections from all the components into a single list of 1000 pseudo-observations, and sort the values.
6) Pick the values at locations 50 and 950 from the sorted list to represent the lower and upper limits of the prediction interval at the 90% confidence level (5% of the values are in each tail of the distribution).

The method results in two new map unit level variables for the attribute (for each depth zone): I label these "_90CL_low" to represent the lower bound of the prediction interval at the 90% confidence level, and "_90CL_high" to represent the upper bound. For example, the map unit level variable for the lower bound of soil organic carbon (soc) in the 0 to 5 cm depth zone is labeled "mu_soc_90CL_low_mr_g_gF_000_005" with units of grams of carbon (g) per gram of soil fines (gF). Figure 7 shows selection of the prediction interval for a hypothetical map unit.



Two more measures that are not reportable as a prediction interval but which influence the uncertainty are the map scale (the "projectscale" variable) and the total area of the map unit (km2).

The following questions have guided the evaluation presented here:
1) How does the pseudo-observation method (_90CL_low) compare the average method (_l) and the low-of-the-low (_ll) method (and the same questions for the high values)?
2) Is there a relationship between the size of the prediction interval (_90CL_high minus _90CL_low) and the map scale (projectscale)?
3) Is there a relationship between the size of the prediction interval (_90CL_high minus _90CL_low) and the map area (km2)?

Results
The soil organic carbon (soc) variable was selected for evaluation here. The plots shown below show that the pseudo-observation method may provide a wider range of values that the other two methods. It is even possible to get negative values, especially if the "low" value coded for a variable was zero, because the method creates a "tail" for the distribution. These negative values should be reset to zero.

Comparing (_90CL_low) with (_l) and (_ll)
When plotting the nearly 300,000 map units for the conterminous United States, there are many possible patterns for the results among the three methods. The three ways of estimating the lower bound of the prediction interval are compared in Figure 8.

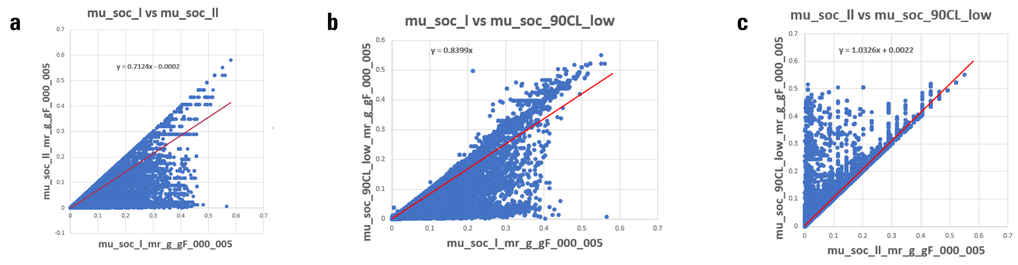


Figure 8. The lower bounds are compared between (a) "_l" and "_ll", (b) "_l" and "_90CL_low", and (c) "_ll" and "_90CL_low". The pattern in (c) seems to have the closest relationship, so the methods will produce similar results for many map units (the red line is close to the 1:1 line).

Comparing (_90CL_high minus _90CL_low) with (_hh minus _ll)
The width of the prediction interval (the upper bound minus the lower bound) is compared for two of the methods in Figure 9.

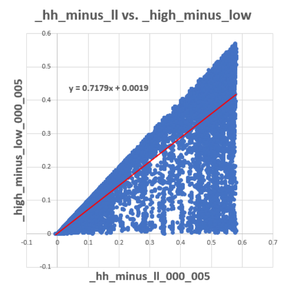


Figure 9. The differences between the higher bound and the lower bound are compared for "_hh minus_ll" versus "_90CL_high minus_90CL_low". The results are similar with some tendency for the pseudo-observation method to produce narrower limits (the red line is below the 1:1 line). There are also many exceptions where the methods produce very different results.

Comparing (_90CL_high minus _90CL_low) with (projectscale) and (km2)
In Figure 10, the width of the prediction interval using the pseudo-observation method is compared with other potential influences on overall uncertainty, the map scale (projectscale) and the land area represented by the map unit in square kilometers (km2).

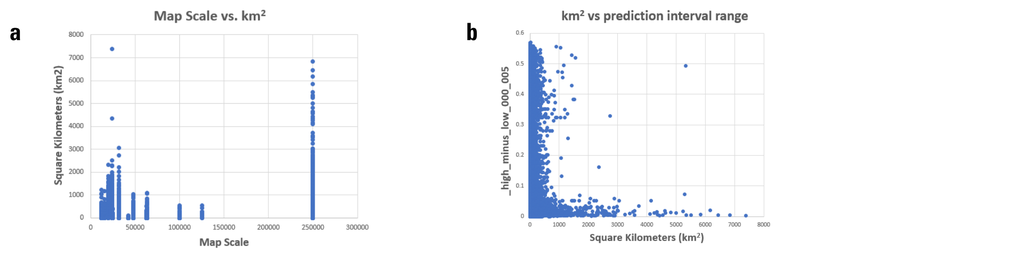


Figure 10. The differences between the higher bound and the lower bound for the pseudo-observation method (Y-axis) are shown according to (a) the map scale (projectscale) and (b) the land area of the map unit (km2). Many of the cases of very large areas also have a low difference between the high and the low, but that could reflect the inherently low carbon values in the Western United States.

Discussion
The pseudo-observation method is likely more robust for representing complex map units with greatly contrasting soil properties. For example, if there are two contrasting components (e.g., #1 with 95% of the land area and a deep organic soil and #2 with 5% of the area in a very sandy (low carbon) soil, then the pseudo-observation method may give more appropriate prediction interval.

Conclusions
1) With insufficient real samples on a nationwide basis, it is not possible to rigorously test the various methods of representing uncertainty.
2) The new method using pseudo-observations has been demonstrated for SOC, and can be applied to the other GlobalSoilMap variables.
3) Future research can investigate how these measures and additional measures such as map scale and map unit area could be synthesized to create an enhanced method for characterizing uncertainty.

References
Helmick, J.L., T.W. Nauman, and J.A. Thompson. 2014. Developing and assessing prediction intervals for soil property maps derived from legacy databases. In D. Arrouays, N. McKenzie, J. Hempel, A.C. Richer de Forges, and A. McBratney (Eds.), GlobalSoilMap: Basis of the Global Soil Information System. CRC Press, New York. p. 359-366.

USDA-NRCS, 2015, The Soil Survey Geographic (SSURGO) database for FY2016, accessed November 20, 2015 (https://sdmdataaccess.nrcs.usda.gov/ or https://gdg.sc.egov.usda.gov/)

USDA-NRCS, 2015, The General Soil Map of the United States (STATSGO2) database, accessed December 30, 2009 (https://sdmdataaccess.nrcs.usda.gov/ or https://gdg.sc.egov.usda.gov/)

Acknowledgments
Thanks to the United States Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS) for providing the soil survey datasets, and to the United States Geological Survey (USGS) Earth Resources Observation an Science (EROS) Center for providing facilities and computer support.