DOI: 10.1007/s00267-003-9119-0

A Revised Methodology for Estimation of Forest Soil Carbon from Spatial Soils and Forest Inventory Data Sets

BEYHAN Y. AMICHEV
Department of Forestry
Virginia Polytechnic Institute and State University
321 Cheatham Hall (0324)
Blacksburg, Virginia 24061, USA

JOHN M. GALBRAITH
Crop & Soil Environmental Sciences Department
Virginia Polytechnic Institute and State University
239 Smyth Hall (0404)
Blacksburg, Virginia 24061, USA

ABSTRACT / Soil organic carbon (SOC) represents the largest constituent of the global C pool and is used by researchers in C cycling, global climate change, and soil quality studies. Spatial, pedon, and soil interpretation record databases are widely used to estimate regional SOC. This study compared published SOC estimates with estimates of mass SOC to 2 m in Maine and Minnesota using STATSGO data tables edited and filled by automated software scripts. Valid STATSGO soil property data were used to produce replacement values for invalid or missing data after grouping by soil order, MLRA, layer number, and texture. Area-weighted mass SOC was calculated using log-transformed data. Between 30% and 54% of the large rock fragment data were invalid, and between 18% and 48% of the missing OM values were replaced. The log-transformed area-weighted mass SOC to 2 m was 7.88 kg/m^2 (SD = 9.24 kg C/m^2 CV = 117.2%) for Maine and 17.38 kg/m^2 (SD = 15.30 kg C/m^2 CV = 88.1%) for Minnesota. These values were lower than earlier estimates because of the log-transformation and because our error checking increased the volume of rock fragments. The FIA database was merged with STATSGO to produce mass SOC by forest-type group. The elm–ash–cottonwood (7.22 kg C/m^2) and the spruce–fir (17.73 kg C/m^2) forest-type groups had the highest SOC (to 1 m depth) in Maine and Minnesota, respectively. The methods and scripts used in this study can be easily adjusted, and as they are improved, they in turn can improve the quality of data in STATSGO tables.

Soil organic C inventory and analysis are required for soil quality assessments (Sikora and Stott 1996) and C cycling predictions (Ellert and others 2002) and are used for state and regional planning by politicians, regulators, and agency employees. Models of global climate change need accurate and complete soil organic C (SOC) inventories because the SOC pool represents the largest component of the global C pool (Jobbagy and Jackson 2000) and acts as a regulator of atmospheric CO_2 levels (Amundson 2001). Previous studies have attempted to extrapolate sources of SOC data available in pedon databases over large areas using small-scale digital soil maps (Franzmeier and others 1985, Huntington and others 1988, Davidson and Lefebvre 1993, Kern 1994, Bliss and others 1995, Homann and others 1998, Johnson and Kern 2003, Galbraith and others 2003). These researchers and modelers encountered consistent problems because of the incomplete nature of the soil databases.

Pedon databases seldom contain a complete inventory of the soil series used as map unit components, or fail to include organic C (OC) or organic matter (OM), bulk density (BD), and rock fragment content (RFC) values for all horizons (Davidson and Lefebvre 1993, Johnson and Kern 2003). United States Department of Agriculture-Natural Resources Conservation Service (USDA-NRCS) Soil Interpretation Record (SIR) databases that accompany their soil surveys include all map unit components but are often missing OM from mineral horizons below the surface 18 cm (Bliss and others 1995), a critical limitation to producing accurate SOC estimates (Lacelle and others 2001). For example, Johnson and Kern (2003, p. 55) studied the USDA-NRCS pedon database and reported that 56% of the mass SOC (excluding surface litter layers) occurred between 0.3 and 1.5 m for mineral soil orders other than Gelisols and Oxisols.

The State Soil Geographic database (STATSGO) is a widely used small-scale (1:250,000) SIR database that provides a digital map and 15 different tables for each

KEY WORDS: STATSGO; FIA; Soil organic carbon; Soil survey; Soil carbon maps; Forest type group

Published online December 4, 2003.

state in the United States (http://www.ftw.nrcs.usda.gov/stat_data.html). A complete description is given in Homann and others (1998, p. 791) and in the STATSGO user's guide (National Soil Survey Center 1994). STATSGO was first issued in 1991, revised in 1993, and last revised in December 1994. STATSGO data is used for soil characterization, global climate change modeling (Bliss and others 1995), soil organic carbon storage estimation (Homann and others 1998; Davidson and Lefebvre 1993) and regional mapping projects (Lacalle and others 2001). Modelers must calculate representative values for soil properties in STATSGO tables, unlike the databases currently used by USDA-NRCS that contain representative values assigned by soil scientists familiar with the soil resources.

STATSGO tables contain empty (null) cells and zero values for various reasons. The cells may have been left unfilled because the job was never completed due to staffing cutbacks or because a value was not applicable, such as the OM content for a layer of bedrock. However, blank cells may have been converted to null or zero and null may be converted to zero inadvertently during institutional database conversions from the original USDA-NRCS State Soil Survey Database. Invalid null and zero values are problematic because they cause calculation errors that result in loss of data from that layer and lower values for properties where mass is calculated on a pedon basis. Recently, a few error-checking routines have been developed to check the data in the National Soil Information System database (NASIS) Ver. 5.x (http://nasis.nrcs.usda.gov/) that is used in the detailed soil survey databases (SSURGO). However, the NASIS database is different than that used for STATSGO, and the update of STATSGO data to the NASIS 5.x format is incomplete.

Zero values in STATSGO tables that are invalid can be detected through expert knowledge and should be replaced by estimated or calculated values. For example, Davidson and Lefebvre (1993) used results and data from technical bulletins to replace 0.0 OC concentration values with 0.2 values for lower layers in their STATSGO tables, based on the results of published research. Textures that have rock fragment modifiers but zero values in all rock fragment content fields are an example of data inconsistency. Conversely, soil textures without a rock fragment modifier could possibly contain zero rock fragments. Zero would never be a reasonable value for minimum or maximum bulk density in a soil horizon and therefore should be treated as invalid data. While error-checking models have been devised by USDA-NRCS for detailed soil survey database products, none have been used on STATSGO. This study will use automated scripts to determine invalid null and zero values in STATSGO and fill them with averages from valid existing STATSGO data without incorporating ancillary data from other sources.

A variety of methods and data sources have been used to fill null cells and replace inaccurate zero values in STATSGO databases. Davidson and Lefebvre (1993) and Galbraith and others (2003) used expert judgment to fill missing data based on ancillary data and data from similar soil series. Davidson and Lefebvre (1993) also assigned minimum values for OM in subsoil layers to replace zero values based on studies that had shown that 0.1 to 0.3% OM actually occurred. Kern (1994) assigned OM concentration where it was invalid by taking half the OM value from the horizon above and adjusted bulk density data based on regressions from USDA lab data. Homann and others (1995) replaced missing data with values from adjoining, genetically similar horizons in the same pedon, and by calculating replacement values from accessory sampling data. Lacelle and others (2001, p. 489) filled missing bulk density data with replacement values from adjacent layers with similar clay and OM and used a neural-net relationship to calculate bulk densities in six categories of soils. Homann and others (1998) used ratios of volumetric SOC in the upper 20 cm to determine the relationship between soil map unit components, then applied that ratio to modify existing subsoil layer SOC for use in the subsoil layers of similar soils that had incomplete records. These methods prevented unreasonably low pedon SOC values while extrapolating incomplete SOC data to calculate regional SOC stocks.

Soil properties are influenced or controlled by the five soil-forming factors (Jenny 1941). Major land resource areas (MRLA) (Soil Conservation Service 1981) and soil classification have been used as the basis of national soil carbon maps (Kern 1994, Homann and others 1998) because they isolate major differences in soil forming factors and incorporate many of the soil, climate, vegetation, and geographic variations that influence OC sequestration in soils.

Few researchers have build regression models in order to describe the relationship between SOC content and site characteristics (Burke and others 1989, Homann and others 1995), including the influence of soil texture (Borchers and Perry 1992). Burke and others (1989) used 300 cultivated and 500 rangeland soil pedons to analyze the relationships between mean annual temperature, annual precipitation, soil texture (silt and clay content), and SOC content in the surface 20 cm for the Central Plains of the United States. Although neither silt nor clay by itself were found to be significant predictors for SOC, Burke and others (1989) found that the combined effects of each with

annual precipitation were significant. Results show that predicted SOC content depends on the combination of clay and silt, so that clay and loam soils are similar, "but sandy soils are predicted to have considerably lower organic C than the fine or medium textured soils" (Burke and others 1989).

Similarly, in a regression analysis study of 134 forest pedon data in western Oregon, Homann and others (1995) reported significant effects of mass clay content and the cross-products of slope X silt and actual evapotranspiration X clay on SOC content for the surface 0–20 cm. Borchers and Perry (1992) analyzed the carbon and nitrogen content in fractionated soils from poorly vegetated clear-cuts and adjacent forested areas in southwest Oregon and reported that C and N concentrations increase with decreasing soil particle size. In addition, relative to sand fraction, the fine-clay fractions were reported to be 10 times richer in C and N. Increased clay content slows organic matter decomposition by absorption and aggregation and results in effective increases of soil organic matter (Burke and others 1989).

Unlike pedon databases, STATSGO contains records for each component of each map unit and reports a range of both low and high estimated values for each property. Most investigators find it satisfactory to assume a symmetrical (normal) distribution of STATSGO data and use the simple average between the low and high values as a representative value (Davidson and Lefebvre 1993, Kern 1994, Bliss and others 1995). However, many soil property distributions are skewed rather than symmetrical (Homann and others 1998, Grigal and others 1991, Brejda and others 2000), as shown in Figure 1a. If STATSGO low and high values are not normally distributed, then data transformation must be used to compute a representative value. Homann and others (1998) used coefficient-transformed data assuming skewed distribution and untransformed data assuming normal distribution. However, they chose the simple average of untransformed data because their coefficient-transformed representative C values (Homann and others 1998, p. 792) reduced regional SOC values from 13.8 to 12.9 kg C/m^2 in the upper 1 m. The higher SOC value was in closer agreement to the values calculated from arithmetic means of pedon data sets alone and area-weighted for four map units, but slightly higher than estimates from coarser-scale FAO soils map and ecosystem type maps (Olson and others 1985). The authors stated that the accuracy and uncertainty of the regional mass SOC values cannot be objectively assessed (Homann and others 1998), leaving an uncertainty as to whether or not the transformation improved prediction accuracy. Recently, Brejda and others (2000) proved that $\log_e$ (lognormal) transformed estimates for most soil properties approximated a normal distribution more closely than the distribution of the nontransformed data (Figure 1b). In addition, due to the fact that most soil properties are assigned only positive values and just a few of them appear as outliers, the $\log_e$ transformation reduces variability two- to threefold for most soil attributes (Brejda and others 2000). We hypothesize that using the mean of a lognormal distribution of values should provide a more representative estimate of soil properties than the simple mean of minimum and maximum values, by reducing the variation between extreme values.

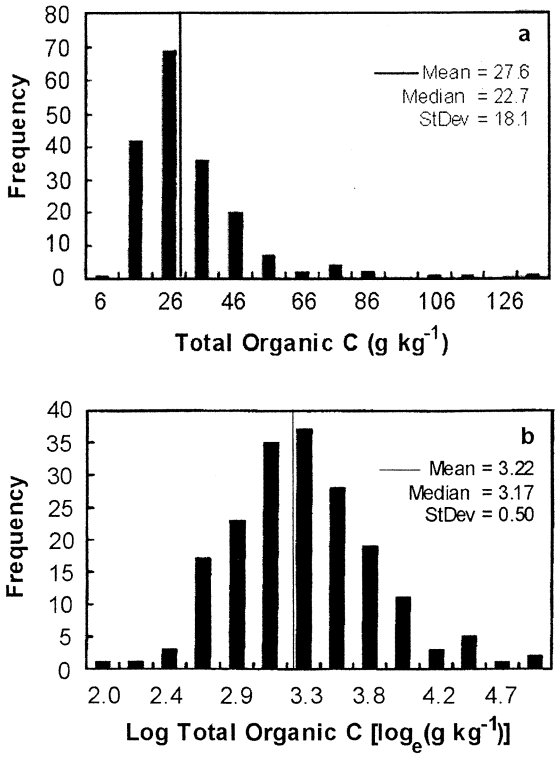

Figure 1. Frequency distribution of total organic C concentrations in MLRA 105, North Mississippi Valley Loess Hills: (a) nontransformed data and (b) log-transformed data (Brejda and others 2000). Vertical lines show means.

Mass SOC data from soils databases are more useful for C sequestration and climate change modeling when combined with biomass or net primary productivity inventories. A preliminary step to determining SOC resources by forest type is to use geographic information systems (GIS) software to integrate national spatial databases such as STATSGO and the US Forest Service Forest Inventory and Analysis (FIA) database (http://fia.fs.fed.us/). Johnson and Kern (2003) used National Soil Characterization Database (NSCD) lab data com-

Table 1. STATSGO soil property variables used in calculation of mass soil OC

Variable code	Variable name
OMH	organic matter high (maximum)
OML	organic matter low (minimum)
BDH	bulk density high (maximum)
BDL	bulk density low (minimum)
INCH3H	percent by weight of rock fragments with size > 25 cm high (maximum)
INCH3L	percent by weight of rock fragments with size > 25 cm low (minimum)
INCH10H	percent by weight of rock fragments with size > 7.5 cm high (maximum)
INCH10L	percent by weight of rock fragments with size > 7.5 cm low (minimum)
NO10H	percent by weight of rock fragments with size < 7.5 cm and which pass through a No. 10 sieve (2 mm screen) high (maximum)
NO10L	percent by weight of rock fragments with size < 7.5 cm and which pass through a No. 10 sieve (2 mm screen) low (minimum)

bined with rock fragment content and soil depth properties from an unfilled STATSGO data set to calculate the mass SOC to 1.5 m from under the forest-type groups classified on Advanced Very High Resolution Radiometer (AVHRR) satellite imagery from across the contiguous United States, but did not report results by individual state.

The objectives of this study are to: (1) identify invalid zero values in STATSGO databases for Maine and Minnesota and then replace nulls and invalid zeros using lookup tables and an automated script for Microsoft Access 2002 (Microsoft Corporation, Redmond, Washington, USA), (2) calculate mass SOC using averages from untransformed and lognormal-transformed database values (called normal and lognormal distribution approaches), (3) present the magnitude of variation in SOC estimates caused by the transformation procedures, and (4) summarize mass SOC by Forest Inventory and Analysis (FIA) forest-type group.

Methods and Materials

Nineteen selected variables (fields) were integrated into one new table using map unit identifier (*MUID*), map unit component (*SEQNUM*), layer number (*LAYERNUM*), and classification code (*CLASCODE*) fields to link data from separate tables. The soil properties used to calculate mass SOC are listed in Table 1. The tables from Maine and Minnesota were sorted and filtered separately to estimate the frequency and distribution of null and zero values. The data was kept separate by state in order to help each state evaluate the quality and target the editing needs of its STATSGO data set.

The texture for any layer was represented by codes listed in the separate *TEXTURE 1, TEXTURE 2,* and *TEXTURE 3* fields. The texture entries were single codes unless the layer contained ≥15% volume of rock fragments. In that case, there was an adjective (rock fragment modifier code) that preceded the texture code and was separated by a dash, such as *STV-FSL* (*STV* = very stony; *FSL* = fine sandy loam). These separate codes are referred to as *TEXTUREx_LEFT* (rock fragment modifier code) and *TEXTUREx_RIGHT* (the fine earth portion).

The approach of this study was to create automated scripts to create texture-based lookup tables in order to identify valid STATSGO values for soil properties used to calculate mass SOC, and to identify invalid entries of zero. The valid entries of these input properties (Table 1) were grouped by layer number, MLRA, and soil order (linked by the *CLASCODE* variable) from Soil Taxonomy (Soil Survey Staff 1999) and used to create lookup tables. The lookup tables were used to fix (replace) nulls and invalid zeros in the same groups that were used to identify the invalid entries and create the lookup table averages.

Assumptions for Modifying Organic Matter Data

Layer number, parent material, soil order, and texture of the fine earth (< 2 mm) were assumed to affect or reflect OM content, but RFC was not. Only *TEXTUREx_RIGHT* codes were used for OM computations and record matching. For example, textures with codes *STV-FSL* and *FSL* were grouped in the same set of records. The following assumptions were considered applicable for determining validity of *OMH* and *OML* records: (a) *OMH* and *OML* should be zero for the following textures: *WB* (weathered bedrock), *UWB* (unweathered bedrock), *CEM* (cemented), and *IND* (indurated); (b) zero value for *OML* is acceptable in mineral or inorganic but not for organic or organic-modified textures; (c) an average value of zero is acceptable for *OMH* for textures that are *ICE* (ice or frozen soil) layers or mostly rock fragments such as *FRAG* (fragmental

material), *G* (gravel), and *CIND* (cinders); and (d) when *TEXTURE 1_RIGHT* codes were *VAR* (variable), *SR* (stratified), and *UNK* (unknown), the texture code in *TEXTURE 2_RIGHT* and *TEXTURE 3_RIGHT* were used as a proxy.

STATSGO data from different states was kept in separate data sets. The following assumptions were considered applicable for grouping valid *OMH* and *OML* records in the Maine and Minnesota data sets: (a) data were separated by MLRA; (b) within each MLRA group, data were separated into four specific soil order groups: Histosols, Spodosols, Andisols, and all others; and (c) data was kept separate by layer in each MLRA and soil order group before averages were calculated at four levels of specificity, ranging from very specific to very general. Invalid records in the STATSGO database were marked for replacement. The phase I lookup tables were created with averages for every possible *TEXTUREx_RIGHT* value. Replacement of invalid data in *TEXTURE 1_RIGHT* was based on an exact match from the phase I lookup table. In case *TEXTURE 1_RIGHT* contained UKN, SR, or VAR, or where there were fewer than three valid values of existing data for the texture in *TEXTURE 1_RIGHT,* the average of the averages for the textures in *TEXTURE 2_RIGHT* and *TEXTURE 3_RIGHT* was used as a replacement. Replacement takes place by joining the original STATSGO *Layer* table with the phase I lookup table by their common fields (*ORDER, MLRA, LAYERNUM,* and *TEXTURE 1*) to transfer replacement values to all records with *OMH;OML* fields that are nulls or zeros. If there were too few valid values to compute a replacement average, no replacement was made and the existing values were grouped into more general categories in phases II–IV. There were 71 unique texture values that were grouped into 19, 9, and 4 groups in phases II–IV. For example, the peat texture in phase I was placed into the fibric group in phase II, the low decomposition OM group in phase III, and the organic group in phase IV. To prevent indecision, the only four groups in the phase IV lookup table were organic, organic modified, mineral, and no carbon. The no-carbon group contained members where the *TEXTUREx_RIGHT* value was *CEMENTED, CINDERS, FRAGMENTAL_MATERIAL, GRAVEL, ICE_OR_FROZEN_SOIL, INDURATED, UNWEATHERED BEDROCK,* or *WEATHERED BEDROCK*. If there were too few valid values to compute a replacement average after phase IV, the zero value was left and the null values were converted to zero and the property table was considered fixed. Altogether, 55 lookup tables were created, 11 for each of the following five properties: OM, BD, NO10, INCH3, and INCH10.

Assumptions for Modifying Bulk Density Data

Layer number, parent material, soil order, texture of the fine earth, rock fragment size, and RFC were assumed to affect the BD of the soil layer. Stones (*ST, STV,* and *STX* texture codes), flags (*FL, FLV,* and *FLX*), and boulders (*BY, BYV,* and *BYX*) were so large that they were assumed not to affect the BD of the fine-earth, but gravel (*G, GRC, GRF, GRV, GRX*), chert (*CR, CRC, CRV, CRX*), cinders (*CIND*), pumice (*PUM, APUM, HPUM, MPUM*), shale (*SH, SHV, SHX*), and channers (*CN, CNV, CNX*) were. Both *TEXTUREx_LEFT* and *TEXTUREx_RIGHT* codes were used for BD computations and record matching. The following assumptions were considered applicable for determining validity of *BDH* and *BDL* records: (a) *BDL* and *BDH* of zero was acceptable for textures *WB, UWB, IND,* and *CEM*. The assumptions for grouping valid *BDH* and *BDL* records and for procedures for replacing invalid *BDH* and *BDL* records were the same described for *OMH* and *OML* above, with the following exceptions: (a) after fixing procedures, a value of 0.00 was used instead of null to prevent calculation errors, although 0.00 was not a reasonable value for bulk density.

Assumptions for Modifying Rock Fragment Data

Layer number, parent material, and texture were assumed to affect the RFC of the soil layer, but soil order was not. It was assumed that soil layers with stones would also contain smaller size rock fragments. The same concept was used when cobbles were present. However, it was not assumed that layers with gravel necessarily contained cobbles or that layers with cobbles always contained stones. Both *TEXTUREx_LEFT* and *TEXTUREx_RIGHT* codes were used for RFC computations and record matching. The following assumptions were considered applicable for determining validity of RFC records: (a) zero values were acceptable for *INCH3L* and *INCH3H* (cobbles) and *INCH10L* and *INCH10H* (stones) if no rock fragment modifier was present; (b) zero values were not acceptable for *INCH3H* (cobbles) and *INCH10H* (stones) if any *TEXTUREx_LEFT* code indicated RFC volume to be $\geq$15%; and (c) zero values were not acceptable for *NO10L* or *NO10H* because those variables represented the percent weight of rock fragments with size less than 7.5 cm plus those that passed through a 2-mm screen, which means that a zero value would not be possible unless there was no fine earth in the layer at all, as in solid bedrock.

The following assumptions were considered applicable for grouping valid RFC records: (a) data were separated by MLRA, and (b) data were kept separate by

layer in each MLRA group before averages were calculated by grouping in phases I–IV. The procedure for replacing invalid RFC records were similar to those described for *OMH* and *OML* above, except that replacement was based on the presence of a code in any of *TEXTURE 1_LEFT, TEXTURE 2_LEFT,* or *TEXTURE 3_LEFT* rather than based on *TEXTURE 1_LEFT,* alone. Replacement was based on the first *TEXTUREx_LEFT* code found in the layer, checked in order from TEXTURE 1 to 3.

Formulae for Calculating SOC from STATSGO Tables

The amount (kilograms of carbon per square meter) of layer organic carbon (LOC) was calculated for each unique layer (*LAYERNUM* variable) in the *Layer* table. Layers composed of bedrock or indurated materials were assigned 0.00 kg C/m^2. Layer organic carbon was calculated as 0.00 kg C/m^2 only when *OML* and *OMH* were 0.00 (Bliss and others 1995, p. 288). The LOC normal method assumed a normal distribution of nontransformed property values and used formulas described by Davidson and Lefebvre (1993) and Bliss and others (1995). The LOC lognormal method assumed a normal distribution of logarithmically transformed property values. LOC-lognormal was calculated the same as LOC-normal but with the following exceptions: (1) the antilog of [(natural logarithm of *OMH* + natural logarithm of *OML*)* 0.5] was used in place of the simple average of the *OMH* and *OML* values, and (2) the antilog of [(natural logarithm of *BDH* + natural logarithm of *BDL*) * 0.5] was used in place of the simple average of the *BDH* and *BDL* values. RFC data were not transformed because of the large number of valid zero and low values for both low and high variables. Mass SOC was calculated using formulas described by Bliss and others (1995) for the upper 2 m of each map unit component (*SEQNUM*) of each map unit (*MUID*) and stored as map unit component organic carbon. Map unit components (*SEQNUM*) of water, rock outcrop, and other miscellaneous land types were assigned 0.00 kg C/m^2. Mass SOC was summed by map unit (MUID_OC) and area-weighted SOC was calculated for both normal and lognormal data sets using formulas in Galbraith and others (2003). The area-weighted standard deviation was calculated as the standard deviation of mass SOC values of all components within a mapping unit rather than the simple standard deviation, which was computed as the variation between SOC values of all mapping units in each state without considering the inherent variation passed from the components.

Combining FIA and STATSGO Data

The FIA database includes three data tables, called *COUNTY, PLOT,* and *TREE,* that are hierarchically related to one other. The most general table is the *COUNTY* table that contains plot-related county and regional unit information. The *PLOT* table provides extensive information on land ownership, current and previous forest type and forest-type group, number of acres that each plot represents on the ground, and a unique plot number. The *TREE* table is the most detailed table and provides tree growth data (Hansen and others 1992). The SOC stores calculated in this study for Maine and Minnesota were reported by FIA forest-type group because much of the area is currently or formerly forested.

The STATSGO spatial layers were reprojected from their Albers Conical Equal Area projection based on the North American Datum of 1927 (NAD27) into unprojected layers with decimal degree units of the NAD27 datum. The latter is currently used by FIA field crews in most US regions to locate inventory plots on the ground with global positioning system (GPS) devices. The SOC data from STATSGO map units were spatially related to FIA forest-type group using the ArcGIS 8.x Geoprocessing Tools (Environmental Systems Research Institute, Inc., Redlands, California, USA). Forest-type group (*ForTypGr*) and area expansion factors (*Expacr*) were the variables extracted from the FIA *Plot* table. Soil organic carbon estimates were assigned to each forest inventory plot by performing spatial overlay analysis of normal and lognormal SOC maps with the FIA plot layer. Then, similar to the approach of Xu and Prisley (2000), SOC values by forest-type group were computed using the following equation:

$$\text{For } TypGr\ SOC = \left(\sum_{F=1}^{j} (MUID\_OC * Expacr)\right) \times \left(\sum_{F=1}^{j} (Expacr)\right)^{-1}$$

where *ForTypGr SOC* is the soil organic carbon by forest-type group (kg m^{-2}), *MUID_OC* is the mass SOC kg m^{-2}, *Expacr* is the expansion factor to relate the area represented by each FIA plot, and *F* is the number of FIA plot records with same forest-type group (F = 1,2,3,…, j).

Results and Discussion

STATSGO *Layer* Table Optimization Results

All null values were replaced by an average or a value of zero. In the Maine STATSGO database, 25% and

Table 2. Inventory of STATSGO *Layer* tables before and after fixing procedures

	Count			% of all records			
Variable	Nulls[a]	Zeros before fixing	Zeros after fixing	Nulls	Zeros before fixing	Zeros after fixing	Records with ≥1 variable fixed
Maine, 3649 records							
INCH10 L	797	2280	1101	22	62	30	54
INCH10 H	797	1045	240	22	29	7	44
INCH3 L	262	2389	824	7	65	23	50
INCH3 H	262	902	65	7	25	2	30
NO10 L	468	1	0	13	< 1	–	13
NO10 H	468	0	0	13	–	–	13
BD L	0	252	246[b]	–	7	7	< 1
BD H	0	252	246[b]	–	7	7	< 1
OM L	0	1975	243	–	54	7	47
OM H	0	906	267	–	25	7	18
Minnesota, 12318 records							
INCH10 L	2651	9665	9663	22	78	78	22
INCH10 H	2651	8515	1372	22	69	11	80
INCH3 L	236	11394	4598	2	92	37	57
INCH3 H	236	6754	729	2	55	6	51
NO10 L	895	0	0	7	–	–	7
NO10 H	895	0	0	7	–	–	7
BD L	0	196	192[b]	–	2	2	< 1
BD H	0	197	192[b]	–	2	2	< 1
OM L	0	6102	193	–	50	2	48
OM H	0	3383	252	–	27	2	25

[a]All null values were considered invalid and changed to zero or larger number.
[b]0.00 values were allowed to prevent computation errors.

54% of the records contained zero values for *OMH* and *OML* before fixing, but only 7% contained zeros after fixing (Table 2). Most of the invalid values were filled during phase I, but some changes were made in all four phases. Bulk density values were almost all valid, as only 7% of the values were zero and < 1% were replaced with nonzero averages. The soils in Maine were dominantly glacial till over hard bedrock, but there were many stony- and cobbly-modified textures that had zero values in *INCH3H* and *INCH10H* fields. The zero values were reduced to 30% and 7% for *INCH10L* and *INCH10H* and reduced to 23% and 2% for *INCH3L* and *INCH3H* after fixing. The 13% nulls in the *NO10* variables were replaced with nonzero averages, since it seemed illogical that any soil horizon would contain zero material that passes through a No. 10 (2 mm) sieve. The results for the Minnesota STATSGO database were similar to Maine for all except the rock fragment variables. The soils in Minnesota had fewer rock fragments (except in northern parts of the state) than the soils in Maine. The zero values were reduced to 78 and 11% for *INCH10L* and *INCH10H* and reduced to 37 and 6% for *INCH3L* and *INCH3H* after fixing (Table 2).

Since a soil may occur up to 21 times in a single map unit and may occur in multiple map units, the averages computed for each lookup table may be heavily influenced by repetitive values from a small number of soil series that occur with high frequency. This would weaken the advantage of using the procedures in this study but could also be easily remedied by filling the database for those frequently occurring soils. These results can be used by the states involved to target improvements in their STATSGO data sets.

The soil order variable conveyed soil morphology inferences that grouped soils with unique properties and materials, such as Histosols, Spodosols, and Andisols. These soil orders were the ones most likely to have significant accumulations of OC below the surface layer and are the three orders with the highest SOC (Kern 1994, Johnson and Kern 2003). However, Histosols and Andisols are uncommon in many states, which may create difficulty in acquiring large enough data sets to produce meaningful average values. Spodosols typically have uniform (often sandy) textures throughout but have a dynamic change in OM content with depth (Hoosbeek and Bryant 1995). Deriving and replacing property averages by layer matching within a texture

Table 3. Descriptive statistics of normal and lognormal area-weighted mass SOC to 1 and 2 m

		Normal (kg C/m^2)			Lognormal (kg C/m^2)		
	Land area km^{2a}	Mean (SD)	*N*	CV (%)	Mean (SD)	*N*	CV (%)
		to 1 m			to 1 m		
Maine	81,457	9.34 (6.65)	69	71.2	6.37 (5.49)	69	86.2
Minnesota	211,902	16.46 (12.45)	321	75.6	13.71 (10.68)	321	77.9
		to 2 m			to 2 m		
Maine	81,457	11.56 (11.46)	69	99.1	7.88 (9.24)	69	117.2
Minnesota	211,902	21.29 (18.61)	321	87.4	17.38 (15.30)	321	88.1

[a]Area determined using Albers equal area projection.

class may be in error if there is inconsistency in the layer number of the E horizon and the spodic horizon. In that case, a layer with very high OM (spodic horizon) may be used to calculate an average for a layer with lower OM (E horizon). Those problems may be remedied by preinspection of the Spodosol data and filling of the subsurface layer OM, BD, and RFC values for series with unusual layer numbering or by using other variables such as pH, BD, or CEC to match spodic layers in different series before creating lookup table averages. If population sizes are large enough, Spodosol data may be grouped by drainage class and regressions may be developed for depth to the spodic horizon as related to the depth of the water table. A similar problem may occur in suborders with buried A horizons (Fluvents) and high subsoil OM (Humults), although those soils are not dominant in any region. The Mollisol (uncommon in Maine and most of Minnesota) soil order is the forth highest in mass SOC and is known to contain high amounts of OC in the surface layer and that value is likely to be valid in STATSGO. However, some Mollisols (especially the Albolls and Aquolls) may also contain accumulations of SOC in layers below layer 1 and therefore could also be separated into a unique soil order group in future uses of this procedure. The rules for the grouping of Mollisols would not apply in all regions, but could easily be incorporated into the procedure for land resource regions (LRRs) that have significant extent of Mollisols, such as in the Midwestern United States.

Davidson and Lefebvre (1993, pp. 116–117) showed that considerably more information was provided by grouping by suborder rather than order, but the number of soils in many of their suborder groups was very low. Kern (1994) reported that grouping pedon data by the great group level of soil taxonomy provided better estimates of SOC than grouping by soil order. He was able to generate meaningful averages by using thousands of pedons sampled from across the nation. That level of detail would not have been possible in this study because the low number of map unit components representing different great groups would not have provided enough data to generate meaningful averages in many cases.

Davidson and Lefebvre (1993) reported a positive relationship at $\alpha = 0.01$ between the area-weighted drainage class and the area-weighted SOC content in the soils of Maine ($R^2 = 0.54$). Their map units contained organic and mineral soils together, so it is unclear if the relationship would be significant for mineral soil series only. The rules for the extra level of grouping would not apply in all regions, but could easily be incorporated into the procedure for states that have significant extent of poorly and very poorly drained soils.

This study has revealed several possible ways to avoid small numbers of valid OM data in future studies. The first solution is to use the USDA-NRCS pedon database used by Kern (1994) and Johnson and Kern (2003) to fill in missing OM values for the map unit components with dominant composition in each MLRA before filling the rest of the database. The matching could be accomplished by series or higher taxa if there were no data for the dominant series. A second solution to increase the size of the data set is to combine data from all adjacent states that are contained in the same MLRA, rather than keeping the data separate by state. A third solution to increase the size of the data set is to aggregate OM data from adjacent MLRAs within the larger LRRs (Soil Conservation Service, 1981), since the soils would still share many soil forming factors and physiographic similarities.

Mass SOC Estimates

Table 3 shows the area-weighted mass SOC for Maine and Minnesota calculated by the normal and lognormal methods. Figure 2 shows the distribution of mass SOC by STATSGO map unit across each state. The area-weighted mass SOC was higher in Minnesota than in Maine because Minnesota contains more Mollisols

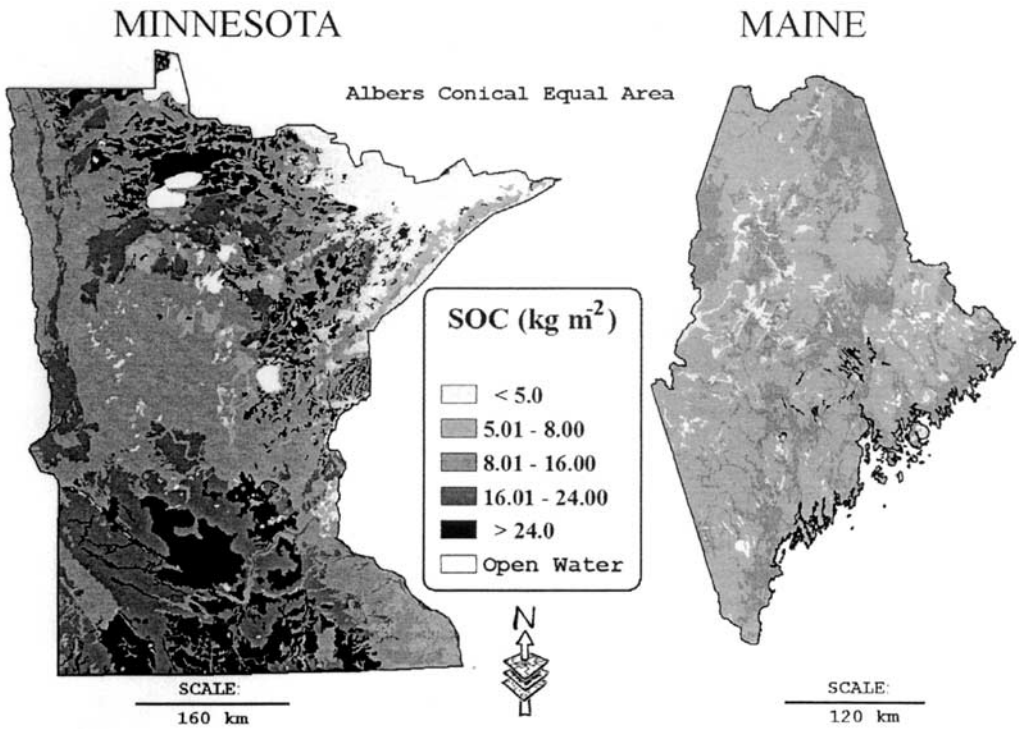

Figure 2. Total SOC maps based on STATSGO map units for the states of Maine and Minnesota computed by the lognormal averaging methods. Open water is included with the < 5 kg/m^2 delineations, and makes up 3.1% of the total area of Maine and 3.0% of Minnesota.

and has deeper soils with fewer rock fragments than in Maine. Both states contain high amounts of Spodosols and Histosols. In Maine, the mass SOC calculated using the normal distribution approach was about 1.5 times higher than the lognormal distribution approach. The same relationship occurred in Minnesota, although the normal mass SOC values were only about 1.2 times higher than the lognormal mass SOC values. Homann and others (1988) reported a similar relationship but a smaller difference between the two methods. There was greater reduction in mass SOC value due to log-transformation than reduction in variability (SD). There was a 1.23- to 1.27-fold times increase in mass SOC between 1 and 2 m in Maine and Minnesota, which is similar to the 1.13-fold times increase between SOC estimates to 1 and 1.5 m reported by Johnson and Kern (2003, p. 55) for mineral soils. In this study, 19% and 21% of the total mass SOC to 2 m was found between the 1- and 2-m depths. The standard deviation (SD) and coefficient of variation (CV) were higher for 2 m than for 1 m SOC, because differences in mass SOC between shallow and deep soils become larger as the deeper layers are analyzed. The SD was higher in Minnesota than in Maine, possibly because of the wider range in climatic and physiographic factors (Soil Conservation Service 1981) as shown in Figure 2.

The lognormal results were compared to other published area-weighted mass SOC estimates in Table 4. Results of other studies using the normal method of SOC calculation were divided by the 1.5 (Maine) and 1.2 (Minnesota) adjustment factors derived from Table 3 to allow comparison to our lognormal SOC values. Franzmeier and others (1985), using a pedon database, reported mass SOC to 1 m ranging from 7.1 to 75 kg C/m^2 within Minnesota, but did not report an area-weighted average by state. However, a careful visual estimate of the area percentage of each soil association in Minnesota (Franzmeier and others 1985, p. 703) times its average SOC resulted in a weighted average estimate of 15.5 kg C/m^2 to 1 m. The lognormal-adjusted value of 12.9 is very similar to our value of 13.7 kg C/m^2. Bliss and others (1995) used STATSGO data and reported a value of 8.4 kg C/m^2 to a variable depth up to 1.65 m for Maine but that was based on only 40% of the land area that had OM and BD data but apparently uncorrected RFC values. Their lognormal-adjusted mass SOC in Maine was lower than ours because they did not fill in missing data in or below layers that

Table 4. Comparison of results between log-transformed SOC data from this study and lognormal-adjusted data from other studies[a]

		Normal (kg C/m^2)		Lognormal[b] (kg C/m^2)	
Data source and date	Depth (m)	Maine	Minnesota	Maine	Minnesota
Lognormal 2003	0–1	—	—	6.37	13.71
Lognormal 2003	0–2	—	—	7.88	17.38
Franzmeier and others (1985)	0–1	—	15.5	—	12.92
Davidson and Lefebvre (1993)	0–1.65	15.5	—	10.33	—
Bliss and others (1995)	0–1.5[c]	8.4[d]	21.3[e]	5.60	17.75
Bliss and others (2003)	0–1.5[f]	14.06	22.95	9.37	19.13
Kern (1994)	0–1	16.82	25.25	11.21	21.04
Johnson and Kern (2003)	0–1	13.67	21.57	9.11	17.98
Johnson and Kern (2003)	0–1.5	16.44	28.14	10.96	23.45

[a]Original (normal) data are shown for comparison.
[b]Estimated by dividing normal SOC from other studies by 1.5 in Maine and by 1.2 in Minnesota.
[c]SOC was not calculated in or below layers where OML, OMH, BDL, and BDH were null or zero.
[d]Based on 40% of the land area.
[e]Based on 95% of the land area.
[f]Approximate depth.

had zero values for both OM and BD. Their lognormal-adjusted value for Minnesota was 17.8 kg C m^{-2} based on 95% of land area and that was very similar to our lognormal value to 2 m. Bliss and others later revised their data with a filled database and 100% of the land area, and their revised lognormal-adjusted SOC values were higher than our lognormal data (N.B. Bliss, personal communication of unpublished data). They apparently did not correct the rock fragment data in STATSGO. Davidson and Lefebvre (1993) reported an average mass SOC of 15.5 kg/m^2 from 0 to about 1.65 m for Maine, using a normal approach on the STATSGO database. Their lognormal-adjusted mass SOC was about 1.6 times higher than our estimate, apparently because they used uncorrected RFC values from STATSGO. Our STATSGO data set for Maine included 30 to 54% nulls and invalid zero values for large rock fragments (Table 2). STATSGO was last revised in December 1994 and they may also have been using a preliminary data set. The area-weighted averages for Maine and Minnesota using the normal method were 16.8 and 25.3 kg C/m^2 to 1 m (J. S. Kern unpublished data 1994). Kern's averages from a pedon database were higher than those in this study because he did not yet correct for rock fragment volume. Johnson and Kern (2003) also calculated higher SOC values than ours (J. S. Kern personal communication of unpublished data) because they apparently did not correct the RFC in STATSGO and because they deleted pedon data from soils with indications of agricultural land use and thus lower OM values

The STATSGO datasets used in this study were insufficient in size to compute separate SOC estimates for mineral and organic soils as did Johnson and Kern (2003). Therefore, we evaluated area-weighted mass SOC by forest-type groups and compared results to those of Johnson and Kern (2003, p. 66) for mineral soils only (Figure 3). Our methods and sources differed from those of Johnson and Kern (2003) because we used lognormal calculation methods and FIA data to estimate extent and location of forest-type groups only in Maine and Minnesota, and they used normal methods of SOC calculation and Advanced Very High Resolution Radiometer (AVHRR) and Landsat Thematic Mapper remote-sensing data across the entire United States. Variation in differences could be explained by different spatial patterns of forest-type groups between FIA versus remote sensing data. Those differences could be quantified using GIS software in future studies. The FIA data are presumed to be more accurate because the plots were visited and verified in the field during data collection but the remote sensing data was apparently not field verified. The elm–ash–cottonwood (7.22 kg C/m^2) and the spruce–fir (17.73 kg C/m^2) forest-type groups had the highest SOC (to 1 m depth) in Maine and Minnesota, respectively (Figure 3). The range of SOC values in Maine was from 5.90 (maple–beech–birch) to 7.22 kg C/m^2 (elm–ash–cottonwood) (Figure 3).

Soil organic carbon estimates by forest-type group are significant piece of information that contributes to a better understanding of the carbon cycle and particularly the complex interaction between trees and soils

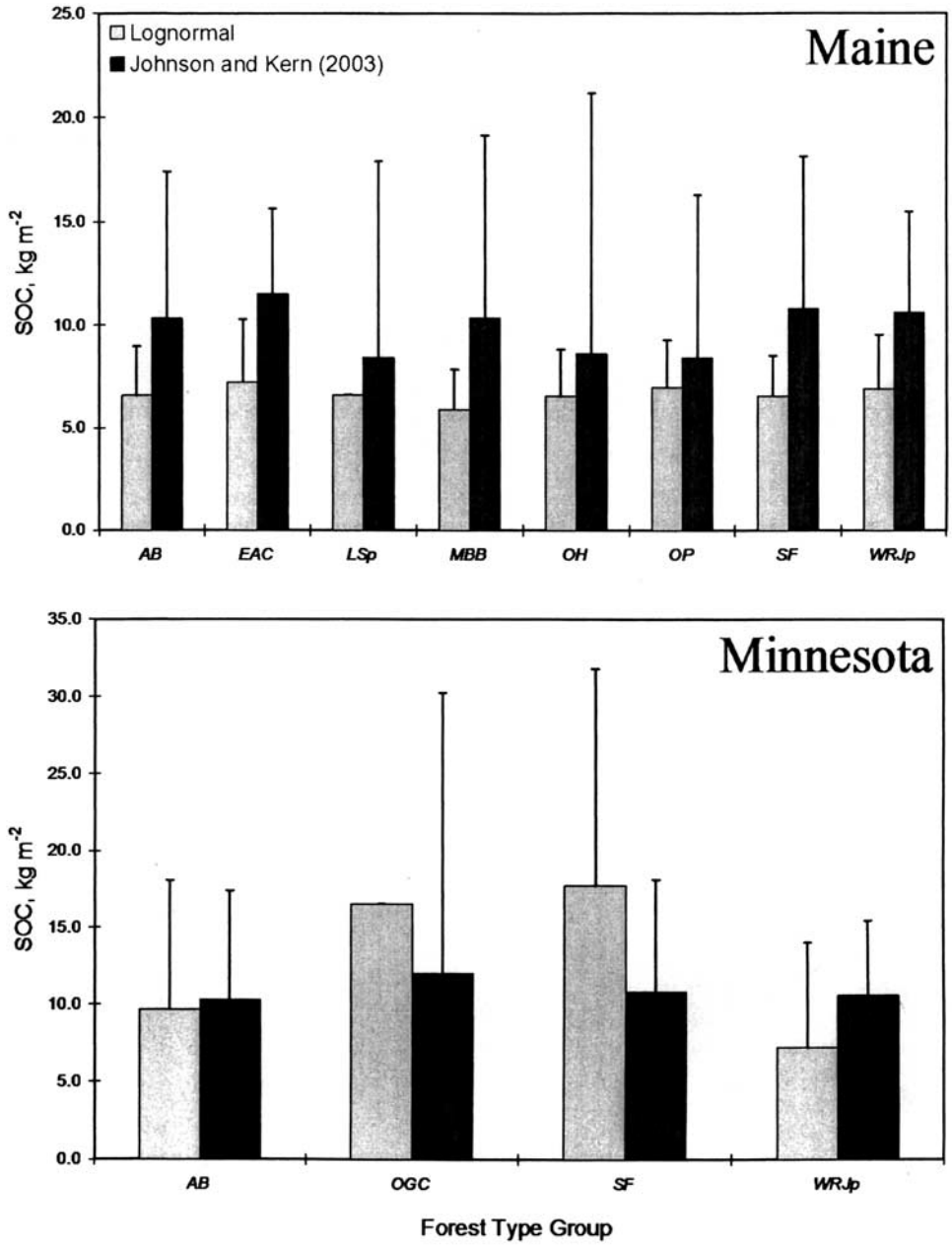

Figure 3. Lognormal SOC and data from Johnson and Kern (2003; p. 66) for mineral soils to 1 m depth with error bars representing one standard deviation from the mean.

within forest ecosystems. The kind of trees influence soil properties and affect the distribution and quantity of SOC in the soil (Johnson and Kern 2003). Over the long term, with the means of remote sensing data (satellite imagery, aerial photography) land use changes detection would be extended to mass SOC flux estimates in the forests. In addition, SOC by forest type group may be used in the USDA-FS FORCARB forest carbon budgeting model (Heath and others 2002).

Conclusions

The averaging and replacement methods used in this study are flexible tools for dealing with null values and identifying invalid zero values in STATSGO databases. This study was conducted as a preliminary effort to automate and critique procedures and rules for SOC estimation from Maine and Minnesota STATSGO data, but the automated, systematic procedures can be used

to edit STATSGO data from multiple states. These methods will have use for correcting current and future versions of STATSGO (scheduled for update in 2004) because STATSGO has not been updated since 1994 and because not all of the data from county soil surveys that will be used as input data for the update has been uniformly and rigorously edited and filled. The Microsoft Access scripts will be made available upon request and can be customized for use in specific regions from menu choices without the use of programming language.

STATSGO was produced on a state-by-state basis. The USDA-NRCS has reorganized and now deals with soil survey by MLRA rather than by state boundaries. The procedures in this study may be helpful to producers of MLRA revision through aggregation of STATSGO (MARTHA'S) databases (http://soils.usda.gov/soil_survey/geography/mlra/marthas.htm) in the future as they identify and remove some STATSGO spatial and map unit composition disagreements along state borders and reduce abrupt changes in SOC along state lines.

This study revealed several ways that improvements could be made in our procedures and in STATSGO data. The averages derived from the grouping procedures in this study would be improved if pedon data were used to replace missing information and verify existing property values of the dominant soils in each MLRA or LRR. This addition would require a reliable method for relating STATSGO database layers to soil series genetic horizons. To increase the data pool size, STATSGO data from different states would have to be aggregated before calculation of average values by MLRA or LRR. Then, if sufficient numbers of valid data occur to produce meaningful averages, data could be grouped for all soil orders or for lower taxonomic levels than soil order. If there are not enough data, then data for Andisols, Spodosols and Mollisols could be grouped by drainage class categories. For Spodosols, all drainage classes should probably be separated, while for the other orders it may be sufficient to group poorly and very poorly drained soils from the better-drained soils.

Addition of surface litter horizons has been recommended by Homann and others (1998), Johnson and Kern (2003), and Galbraith and others (2003). Johnson and Kern (2003) also recommend screening of STATSGO components for land use, so that OM values specific to land use or vegetation group can be developed. STATSGO was easily joined with FIA data to produced SOC averages by forest-type group, making the FIA database more complete and leading the way to producing total ecosystem C estimates in the forests that include C from standing biomass, surface litter, dead roots, and soil organic matter.

Regional or national carbon inventories for other vegetation types can also be completed by integrating digital land cover data with mass SOC from STATSGO and vegetation biomass and root production data from other sources. Finally, soil sampling and litter layer collection could be incorporated with future data collection at USDA-NRCS National Resource Inventory and FIA plots and the SOC data used to conduct some validation studies of regional and national SOC inventories such as the one in this study.

Acknowledgements

The authors wish to thank the USDA-FS for funding this project and Norman Bliss of the USGS and Sharon Waltman and Cathy Seybold of USDA-NRCS for STATSGO technical advice. Forest soil expertise was provided by Linda Health of USDA-FS and Stephen Prisley of Virginia Tech. Elizabeth LaPoint, National Forest Inventory and Analysis Geospatial Service Center, provided the FIA plot data and overlay analysis of the STATSGO and FIA datasets. Jeff Kern of Dynamac Corporation, US EPA National Health and Environmental Effects Laboratory, provided unpublished data and methodology.

References

Amundson, R 2001. The carbon budget in soils. Annual review Earth planet. *Science* 29:535–562.

Bliss, N. B., S. W. Waltman, and G. W. Peterson. 1995. Preparing a soil carbon inventory for the United States using geographic information systems. Pages 275–295 *in* R. Lal, J. Kimble, E. Levine, and B. Stewart. Eds, Soils and global change. CRC Press Inc., Boca Raton, Florida.

Borchers, J. G., and D. A. Perry. 1992. The influence of soil texture and aggregation on carbon and nitrogen dynamics in southwest Oregon forests and clearcuts. *Canadian Journal of Forest Research* 22:298–305.

Brejda, J. J., T. B. Moorman, J. L. Smith, D. L. Karlen, D. L. Allan, and T. H. Dao. 2000. Distribution and variability of surface soil properties at a regional scale. *Soil Science Society of America Journal* 64:974–982.

Burke, I. C., C. M. Yonker, W. J. Parton, C. V. Cole, K. Flach, and D. S. Schimel. 1989. Texture, climate, and cultivation effects on soil organic matter content in US grassland soils. *Soil Science Society of America Journal* 53:800–805.

Davidson, E. A., and P. A. Lefebvre. 1993. Estimating regional carbon stocks and spatially covarying edaphic factors using soil maps at three scales. *Biogeochemistry* 22:107–131.

Ellert, B. H., H. H. Janzen, and T Entz. 2002. Assessment of a method to measure temporal change in soil carbon storage. *Soil Science Society of America Journal* 66:1687–1695.

Franzmeier, D. P., G. D. Lemme, and R. J. Miles. 1985. Organic carbon in soils of the North Central United States. *Soil Science Society of America Journal* 49:702–708.

Galbraith, J. M., P. J. A. Kleinman, and R. B. Bryant. 2003: In press. Sources of uncertainty affecting soil organic carbon estimates in Northern New York. *Soil Science Society of America Journal* 67(4).

Grigal, D. F., R. E. McRoberts, and L. F. Ohmann. 1991. Spatial variation in chemical properties of forest floor and surface mineral soil in the north central United States. *Soil Science* 151:282–290.

Hansen, M. H., T. Frieswyk, J. F. Glover, and J. F. Kelly. 1992. The eastwide forest inventory database: users manual. General Technical Report NC-151. USDA Forest Service. .

Heath, L. S., R. A. Birdsey, and D. W. Williams. 2002. Methodology for estimating soil carbon for the forest carbon budget model of the United States, 2001. *Environmental Pollution* 116:373–380.

Homann, P. S., P. Sollins, H. N. Chappell, and A. G. Stangenberger. 1995. Soil organic carbon content of mountainous, forested region: relation to site characteristics. *Soil Science Society of America Journal* 59:1468–1475.

Homann, P. S., P. Sollins, M. Fiorella, T. Thorson, and J. S. Kern. 1998. Regional soil organic carbon storage estimates for western Oregon by multiple approaches. *Soil Science Society of America Journal* 62:789–796.

Hoosbeek, M. R., and R. B. Bryant. 1995. Modeling the dynamics of organic carbon in a Typic Haplorthod. Chapter 34. *in* R. Lal, J. Kimble, E. Levine, and B. A. Stewart. Eds, Advances in soil science: soils and global change. CRC Press, Boca Raton, Florida.

Huntington, T. G., D. F. Ryan, and S. P. Hamburg. 1988. Estimating soil nitrogen and carbon pools in a northern hardwood forest ecosystem. *Soil Science Society of America Journal* 52:1162–1167.

Jenny, H. 1941. Factors of soil formation: a system of quantitative pedology. McGraw-Hill, New York.

Jobbagy, E. G., and R. B. Jackson. 2000. The vertical distribution of organic carbon and its relation to climate and vegetation. *Ecological Applications* 10:423–436.

Johnson, M. G., and J. S. Kern. 2003. Quantifying the organic carbon held in forested soils of the United States and Puerto Rico. Pages 47–72 *in* J. M. Kimble, L. S. Heath, R. A. Birdsey, and R. Lal. Eds, The potential of US forest soils to sequester carbon and mitigate the greenhouse effect. CRC Press, Boca Raton, Florida.

Kern, J. S 1994. Spatial patterns of soil organic carbon in the contiguous United States. *Soil Science Society of America Journal* 58:439–455.

Lacelle, B., S. Waltman, N. Bliss, and F. Orozco-Chavez. 2001. Methods Used to Create the North American Soil Organic Digital Database. Pages 485–494 *in* R. Lal, J. Kimble, R. Follett, and B. Stewart. Eds, Assessment methods for soil carbon. Lewis Publishers, Boca Raton, Florida.

National Soil Survey Center. 1994. State Soil Geographic (STATSGO) Data base, Data use information. Miscellaneous Publication Number 1492. USDA-NRCS, Lincoln, NE. Available at *http://www.ftw.nrcs.usda.gov/stat_data.html* (verified 14 January 2003). Users Guide available at http://www.ftw.nrcs.usda.gov/pdf/statsgo_db.pdf.

Olson, J. S., J. A. Watts, and L. J. Allison. 1985. Major world ecosystem complexes ranked in live vegetation: a database. NDP-017. Carbon Dioxide Information Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Sikora, L. J., and D. E. Stott. 1996. Soil organic carbon and nitrogen. Pages 157–168 *in* J.W. Doran, and A.J. Jones. Eds, Methods for assessing soil quality. Soil Science Society of America Special Publication 49. Soil Science Society of America, Madison, Wisconsin.

Soil Conservation Service 1981 (verified 16 May 2003). Land resources regions and major land resource areas of the US. USDA-SCS. Agriculture Handbook No. 296. US Government Printing Office, Washington, DC. Available on-line at http://soils.usda.gov/soil_survey/geography/mlra/main.htm. .

Soil Survey Staff. 1999. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys, 2nd ed. USDA-SCS. Agriculture Handbook No. 436. US Government Printing Office, Washington, DC.

Xu, Y. J., and S. P. Prisley. 2000 (verified 16 January 2003). Linking STATSGO and FIA data for spatial analyses of land carbon densities. Proceedings of the Third USDA Forest Service Southern Forestry GIS Conference, 10–12 October 2000. Athens, Georgia. Available on-line at http://www.soforgis.net/2000/cdrom/posters.html. .