Environmental Factors Relating to Landscape Variation in Soil Carbon Fractions in Florida.

D. Brent Myers[1], Sabine Grunwald[2], Aja M. Stoppe[3], Nick B. Comerford[3], and Willie G. Harris[2]

[1,2]USDA-ARS-Cropping Systems and Water Quality Unit, Columbia, MO
[2]Department of Soil and Water Science, University of Florida, Gainesville, FL
[3]North Florida Research and Education Center, Department of Soil and Water Science, University of Florida, Quincy, FL

Supported by National Research Initiative Competitive Grant no. 2007-35107-18368 from the USDA National Institute of Food and Agriculture

Introduction

Carbon (C) is a spatially and chemically dynamic component in the soil landscape. The dynamic processes controlling the spatial variability of C fractions may be explored by analysis of environmental information correlated to soil C. Some examples of potential environmental correlates are soil map-unit, vegetative indices, rainfall, and elevation. Environmental correlation models may be developed to quantify total soil C and chemical C fractions in a soil-landscape and may help to understand the processes and dynamics of the C cycle. One type of environmental correlation model is the soil factorial model. Soil factorial models relate variability of soils into key conceptual genetic gradients (e.g. SCORPAN: soil, climate, organisms, relief, parent material, age, time). Digital soil mapping requires the production of a statistical, geostatistical, or mathematical model to predict soil properties, and may be based on the SCORPAN model. There are significant challenges to implement environmental correlation models based on large datasets of predictors due to the large computational infrastructure needed to manage the data and models.

Objective

Identify parsimonious sets of environmental variables for large extent application of models of soil C fractions.

Materials and Methods

Experimental Design

A statewide dataset was collected at 1014 sites from the top 20 cm of Florida soils. A stratified random spatial sampling design (n=1014) was developed to efficiently cover the range of soil-landscape and environmental factors controlling soil C distribution and the range of C fractions. Points were located and mapped by a differentially corrected GPS (fig. 1). Four 20 x 5.8 cm soil cores were collected from each of the sample locations. Oven-dry bulk density was measured . All C measurements refer to this fine-earth fraction of the soil on a kg m^{-2} basis for a 20 cm surface soil profile.

Soil Carbon Measurements

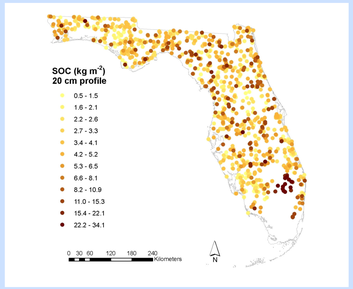

Figure 1. A random stratified sampling design (strata: land cover—soil suborder) was implemented for the measurement of soil C fractions including soil organic carbon (SOC).

Total soil C (TC) and inorganic C (IC) were measured by gas analyzer (Shimadzu TOC -V/SSM-5000). Soil organic C was calculated by subtracting IC from TC. Recalcitrant C (RC) was measured as the residual C after digestion of 4g soil with 20 ml 6M HCL for 16 hours under reflux conditions. Hot water extractable C (HC) was measured as C dissolved after 16 hours of incubation at 80°C.

Environmental Data Layers

Environmental covariates (n=185) with statewide coverage were collected to serve as potential predictors of soil C fractions. These variables represent a variety of data types including floating point, integer, nominal, ordinal, and binary. Some variables represent multi-temporal sequences such as monthly rainfall or monthly vegetation indices. Some variables were collected at multiple scales such as digital elevation models (30, 90, and 1000 meters). Some variables such as LANDSAT and MODIS products are raster images while others such as SSURGO soil maps and geologic provinces are mapped as polygons. The environmental variables were organized by SCORPAN factors and converted to a common 30 meter grid resolution. See tables 1 and 2 for a more detailed summary of environmental variables.

Data Mining Approach

Few traditional modeling approaches are able to handle the heterogeneous nature of the dataset collected here. However, regression tree and ensemble regression tree methods offer opportunities to model complex, high-order interactions between SOC and large sets of environmental data. These data mining methods are non-parametric, can handle different data types, and large sets of predictor variables. We use the random forest method (Brieman, 2001) for model production and comparison due to the minimal parameter adjustment necessary and low sensitivity to parameter settings. A training/testing framework was used for model fitting and assessment. The 1014 sites were randomly divided into calibration and validation datasets using a 70/30 split. Random forest models were fit to the calibration datasets while model fit statistics were calculated from the test dataset.

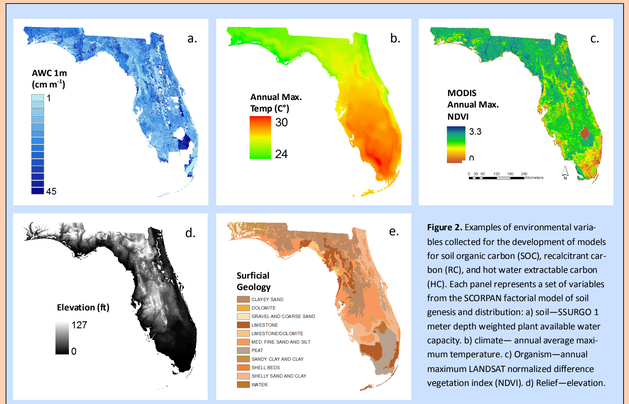

Figure 2. Examples of environmental variables collected for the development of models for soil organic carbon (SOC), recalcitrant carbon (RC), and hot water extractable carbon (HC). Each panel represents a set of variables from the SCORPAN factorial model of soil genesis and distribution: a) soil—SSURGO 1 meter depth weighted plant available water capacity. b) climate— annual average maximum temperature. c) Organism—annual maximum LANDSAT normalized difference vegetation index (NDVI). d) Relief—elevation.

Table 1. Brief listing of some environmental variables used to model SOC, RC, and HC.

SCORPAN Factor	n	Example Variables	Important properties
Soil	28	SSURGO attributes: taxonomy, drainage class, texture, om, albedo	polygon, mixed continuous and categorical
Climate	38	monthy average precip, mintemp, maxtemp, annual averages	multiple raster resolutions (1-4 km)
Organisms	99	land-use, land-cover, and ecoregion classifications	mixed polygon and raster datasets, multiscale
		LANDSAT and MODIS remote sensing products (vegetative indices, phenology, spectral bands)	multitemporal and multiresolution rasters
Relief	15	Digital Elevation Models (DEM) National Elevation Dataset, USGS elevation, Hydro 1K (elevation, slope, topographic wetness index)	multiple resolutions (30, 90, 1000 meter)
Parent Material	5	Physiographic province, surficial geology	large polygons

Results and Discussion

Full Model Performance and Variable Ranking

Random forest models of SOC and RC performed moderately well, and substantially better than HC models (fig. 3). The performance of these models is sufficient to produce digital soil maps of C fractions, however the large set of raster images needed to produce a map are difficult to manage computationally, particularly in regard to sophisticated data mining models such as random forest models.

Variable importance measures (fig. 3 d-f) highlight some of the processes at work in Florida's landscape to control C distribution. Soil taxonomy, soil properties, land-cover/land-use (LCLU), and hydrologic properties are key variables indicating that an interaction between soil, plant communities, and soil profile moisture are controlling C fraction distribution in the landscape (see table 2 for a listing of names and descriptions for some important variables). Geologic variables were also prominent indicating different equilibrium states of C fraction content for different geomorphology, age, and types of surficial materials.

Incremental Variable Performance

Figure 4 shows the incremental improvement of root mean squared error (RMSE) of random forest models as predictor number (n) increases from 5 to 185. For each soil C fraction there is a point at which additional variables do not provide model improvement. For SOC this occurs somewhere around 25 variables, while for RC and HC about 100 variables are needed to maximize RMSE.

As predictor count increases, so too does the likelihood that redundant information is present in the dataset. This is undesirable to different degrees depending on the regression model chosen. Some, like random forests are resistant to problems caused by multicolinearity. However, a larger problem exists for the purpose of digital soil mapping when redundant variables are included in a model. Predictor variables translate into raster datasets that must be collected, managed, processed, and to which models must be applied in order to generate prediction maps. This can be computationally prohibitive. The best and smallest set of predictors is ideal for the purpose of applying sophisticated data mining models on a pixel by pixel basis.

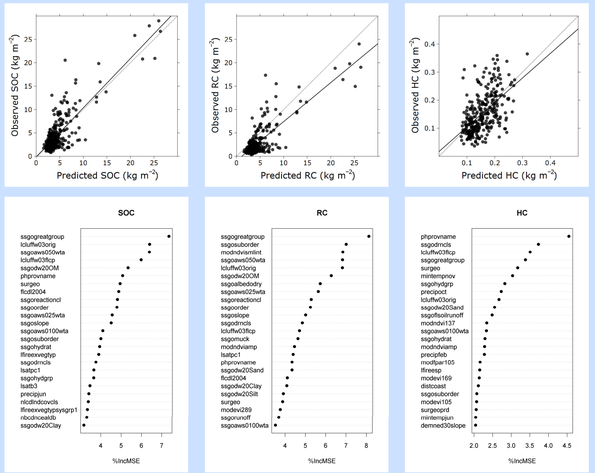

Figure 3. **a-c.** Validation plots visually show the performance of random forest models of soil organic C (SOC), recalcitrant C (RC), and hot water extractable C (HC). **d-f.** Variable importance plots show the relative ranking of contributions from the top 25 variables form each training model (see table 2 for a listing of names and descriptions for some important variables). Percent increase in mean squared error (%IncMSE) indicates the improvement a given variable provides to a random forest model compared to its random permutation.

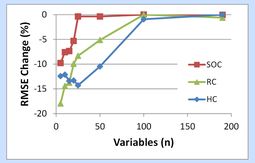

Figure 4. Reduction in root mean squared error (RMSE) of an independent validation dataset due to models fit with reduced numbers of variables as compared to a full 185 variable random forest model.

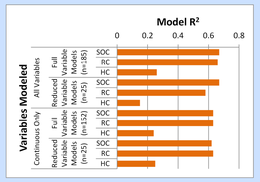

Figure 5. Random forest regression results for variable subsets: all 185, top 25, all 152 continuous, top 25 continuous.

Continuous Variable Performance

In addition to the computational challenges of large predictor datasets, there is significant challenge when working with categorical variables due to small or empty membership in some classes. This results in the problem that classes within certain variables may not be present in the training versus the raster and testing datasets. This requires grouping, reclassification, or imputation of missing variables to successfully create predictions for these locations. Continuous variables do not exhibit these problems.

We compared the effectiveness of continuous predictors only against the full model (categorical and continuous). Further, we compared models of the top 25 variables from each of these groups to their full model counterpart. Figure 5 shows that there is minimal difference between the fit of the full variable model and the continuous variable model. The top 25 variables from both of these sets perform almost identically to their full model results.

Factor Group Performance

Compared to other factors, soil and organism factors provided the most predictive power for any of the carbon fractions (Figure 6). The success of soil variables is largely due to hydrologic attributes highlighting the process linkage between soil moisture and C processes. The importance of organism variables indicates the importance of biomass productivity and plant communities on C fraction levels. Relief repeatedly scored worse than other factor groups. This is contradictory to what actually happens in Florida's relatively flat landscape. Small variations in topography can lead to large differences in hydrology and readily observable differences in C accumulation. The poor performance of relief variables is more likely due to the quality, scale, and relative error of the elevation data available for Florida. The low gradients across most of Florida result in poor interpolation of hypsographic lines and subsequently noisy derivatives and interpolation artifacts. More accurate and finer resolution elevation data might be more useful than these results indicate.

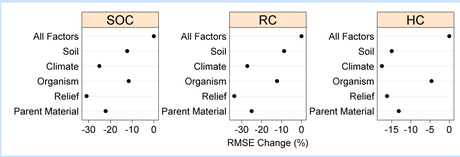

Figure 6. Validation results from SCORPAN factor group random forest models of soil organic C (SOC), recalcitrant C (RC), and hot water extractable C (HC) fractions. Percent change in root mean squared error (RMSE) indicates reduction in RMSE of the single factor group model as compared to the all factors model.

Conclusions

This research addresses model development and implementation issues that provide challenges to estimate soil C fractions across a large subtropical region composed of diverse soil-hydrology and land uses.

- Prediction models for TC and RC showed higher accuracy when compared to HC potentially due to the spatio-temporal dynamics of labile C, which makes it challenging to achieve good predictions.
- Overparametrization of digital soil prediction models was addressed with a novel data reduction method focused to minimize prediction errors AND delineate minimal environmental predictor datasets.
- A relatively small set of 25 continuous variables is nearly as successful as a 185 variable categorical-continuous dataset to model SOC.
- Key environmental predictors were identified which related to SOC, RC, and HC, respectively. Soil-hydrologic and taxonomic as well as vegetation/biomass, land use, and phenology properties showed the highest predictive capabilities to infer on soil C forms.
- Interestingly, climatic properties which represent the long-term forcings of global climate change in FL did not show a close linkage to soil carbon.

Table 2. Important variables for the prediction of soil C fractions.

SCORPAN	Variable	Description
O	flcdl2004	NASS cropland datalayer classification
O	lcluffw03orig	Florida Fish and Wildlife LC/LU
O	lcluffw03flcp	Florida Fish and Wildlife LC/LU
O	lfireexvegtyp	LANDFIRE existing veg. type
O	lsatb4	LANDSAT band 4
O	nbcdbawht	national biomass carbon dataset basal area weighted height
C	precipoct	monthly average precipitation
S	ssgoaws050wta	SSURGO depth weighted available water supply to 50 cm
S	ssgodrncls	SSURGO drainage class
S	ssgodw20OM	SSURGO depth weighted organic matter
S	ssgodw20Sand	SSURGO depth weighted sand
S	ssgodw20Silt	SSURGO depth weighted silt
S	ssgoflsoilrunoff	SSURGO Florida soil runoff potential
S	ssgogreatgroup	SSURGO taxonomic great group
S	ssgohydgrp	SSURGO hydrologic group
S	ssgohydrat	SSURGO hydric rating
S	ssgomuck	SSURGO series name 'muck' indicator
S	ssgoorder	SSURGO taxonomic order
S	ssgoreactioncl	SSURGO taxonomic reaction class
S	ssgoslope	SSURGO slope
S	ssgosuborder	SSURGO taxonomic suborder
P	surgeo	surficial geology