Data mining reveals relationships between soil C and environmental factors at Tier 2 sites

C. Wade Ross[1], Sabine Grunwald[1], Jason Vogel[2], Allan Bacon[1], Eric J. Jokela[3], Rosvel Bracho-Garrilo[3],
Daniel Markewitz[4], Madison Akers[4], Joshua Cucinella[3], Andy Laviner[5], Tom Fox[5], Tim Martin[3], Carlos Gonzalez[6]
[1] University of Florida, Department of Soil and Water Science, [2]Texas A&M – Ecosystem Science and Management, [3]University of Florida, School of Forest Resources and Conservation,
[4]University of Georgia, Warnell School of Forestry and Natural Resources, [5]Virginia Polytechnic Institute and State University, [6]Oregon State University

Rationale

Soils of the US Southeast are *estimated* to store between 8.9 and 51.2 Pg carbon (C), accounting for more than 1/3 of the total soil C storage for the conterminous US when using the median value. The large range of estimates is due to 1) high spatial variance of soil C, 2) difficulties associated with collecting and analyzing enough samples to capture this variance across large regions, and 3) estimating soil C from highly aggregated legacy datasets with limited sample support. Additionally, many studies do not attempt to quantify C below the top soil (0-20 cm). To improve soil C estimates to 1m depth in forested ecosystems of the US Southeast, we used machine learning (Random Forests) with measured soil properties from the Tier 2 network in conjunction with a large suite of environmental data. This *data mining* approach revealed relationships between soil C and environmental covariates.

Materials & Methods

The concentration of soil C was analyzed across the Tier 2 network (N ~ 322) at four depth profiles: 0-10, 10-20, 20-50, and 50-100 cm. To predict soil C to 1m, we aggregated these depth-based measurements to 1m with *smoothing splines*.
We acquired nearly ***7 terabytes of data*** from a multitude of agencies to represent the 8 factors in the STEP-AWBH modeling framework. After spatially extracting this information to Tier 2 sites, we applied machine learning (Random Forests) algorithms to this large suite of environmental covariates (N ~ 300) to predict soil C to 1m.

Table 1: Factors representing the **STEP-AWBH** modeling framework used with Random Forests

Factors	Property	Factors	Property
Soil	Soil order	**Ecology**	Monthly LAI
	Soil suborder		Annual min, max and mean NDVI
	Soil great group		NDVI green-up & brown-down
	Soil texture		NDVI greenup and browndown rate
	Cation exchange capacity		NDVI Season length
	Bulk density		NDVI amplitude and base NDVI level
	Organic matter content		Max peak NDVI
	Site index		Canopy coverage and Imperviousness
	Phosphorus		Basal area weighted canopy height
	Nitrogen		Aboveground live dry biomass
	Potassium		Gross and net primary production
	Calcium		Woody biomass
	Zinc		Root biomass
	Manganese		Basal area
	Copper		Leaf area index
	pH		Stem volume over bark
	Base saturation		Coarse root production
	Boron		Foliage biomass production
	Magnesium		Woody biomass production
	gamma absorbed dose		Total net primary production (NPP)
	gammaPotassium		Above ground NPP
	gammaThorium		Site index
	gammaUranium	**Parent Material**	Age
	Iron		Predominant lithology
Topography	Slope		Secondary lithology
	Aspect	**Atmospheric**	Precipitation
	Curvature		Maximum temperature
	Profile curvature		Minimum temperature
	Flow direction		Maximum relative humidity
	Flow accumulation		Minimum relative humidity
	Latitude / longitude		Solar radiation
	Topographic wetness index	**Water**	Available water capacity
Ecology	Major Land Resource Area		Drainage class
	Biophysical settings		Hydric rating
	Disturbance		Saturated hydraulic conductivity
	Land cover		Soil moisture (seasonal & annual)
	Environmental site potential	**Biota**	Foliage
	Monthly MODIS FPAR		Vegetation type
	Vegetation height		Vegetation type system group 1
	Vegetation cover		Vegetation type system group 2
	Forest canopy properties		Vegetation type order
	Landsat ETM þ bands		Vegetation type class
	Landsat ETM þ tasseled cap indices		Vegetation type subclass
	Landsat ETM þ principal components	**Human**	Land use
	Monthly NDVI		Trees per hectare
	Monthly EVI		Stand / Plot age

Workflow

The conceptual model below (Figure 1) illustrates the modeling workflow.
1) Aggregate measured soil C to 1m using a spline function
2) Compile data
3) Extract data to Tier 2 sites
4) Perform data munging / cleaning / aggregating
5) Create a single 'master file' containing all relevant data
6) Split data into training (70%) and validation (30%) sets
7) Random Forest generates an ensemble of *decision trees* for soil C predictions
8) Finally, model performance is tested using the validation dataset

Figure 1: Conceptual model of workflow.

Results

- **Top predictors** - Major Land Resource Areas, geology, biomass, climate data, and soil properties
- **Model performance** - adjusted R^2 of 0.93 and 0.52 for training and validation sets using all data
- **Model performance** - adjusted R^2 of 0.90 and 0.36 (training & validation) when using only ancillary data

The best models were achieved using a comprehensive set of STEP-AWBH predictors, which consists of measured soil properties in conjunction with a wide variety of publicly available data sources from various foreign and domestic agencies, which resulted in an adjusted R^2 of 0.93 and 0.52 for training and validation sets, respectively. The top five predictors using this approach include eco-regions, soil great group, concentration of soil nitrogen, major land resource areas, and age of parent material (geology). Of these predictors, four represent categorical data, while the remaining represents continuous data.

Random Forests also performed well when using only ancillary data (not explicitly measured at Tier 2) as predictors, and resulted in an adjusted R^2 of 0.90 and 0.36 for training and validation datasets, respectively. The top five predictors using this approach include eco-regions, soil great group, major land resource areas, parent material age, and soil suborder. The top five predictors all represent categorical data.

Refer to Figure 2 and 3 below for the full list of important variables.

Important variable using measured Tier 2 soil data

Predictors	Data source
Eco-region level 4	Eco-regions
Soil great group	gSSURGO
Nitrogen concentration	Tier 2
Major land resource area	MLRA
Parent material age	US geology
Organic matter (1m)	Tier 2
Cation exchange capacity	Tier 2
Magnesium ppm	Tier 2
Soil suborder	gSSURGO
Calcium ppm	Tier 2
Secondary lithology	US geology
Bulk density	Tier 2
Predominant lithology	US geology
Soil order	gSSURGO
Soil moisture (April)	SMOS
Manganese ppm	Tier 2
Trees per hectare	Tier 2
Minimum solar radiation	METDATA
gammaPotassium	Gamma Ray
GammaAbsrorbedDose	Gamma Ray
Clay content	gSSURGO
Drainage class	gSSURGO
Potassium ppm	Tier 2
Land use / Land cover	NASS
Temperature (March-May)	METDATA
Disturbance	LANDFIRE
Phenology production	AVHRR
Foliage	Tier 2
Basal area	Tier 2
Precipitation (June-August)	METDATA

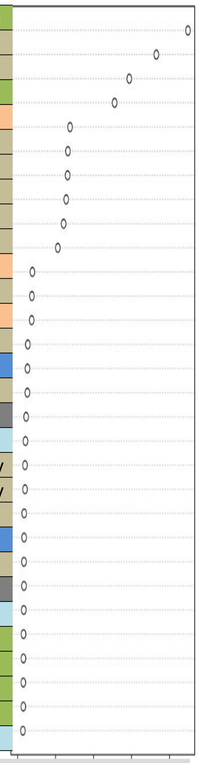

Figure 2. Variables of importance when predicting soil C to 1m depth using measured soil data collected from the Tier 2 network in addition to ancillary data.

Important variables without measured Tier 2 soil data

Predictors	Data source
Eco-region level 4	Eco-regions
Soil great group	gSSURGO
Major land resource area	MLRA
Age of parent material	US geology
Soil suborder	gSSURGO
Predominant lithology	US geology
Secondary lithology	US geology
Topographic curvature	SRTM
Land use / land cover	NASS
Trees per hectare	Tier 2
Sand content	gSSURGO
Stem volume over bark	Tier 2
Leaf are index	Tier 2
Organic matter	gSSURGO
Drainage class	gSSURGO
Minimum solar radiation	METDATA
Soil moisture (April)	SMOS
Soil moisture (July)	SMOS
Clay content	gSSURGO
Disturbance	LANDFIRE
Foliage production	Tier 2
Site index	Tier 2
gammaPotassium	Gamma Ray
Phenology	AVHRR
Soil moisture (Jan.-March)	SMOS
Biomass	NBCD
Precipitation (June-August)	METDATA
Max. temp. (Dec. Feb.)	METDATA
Flow accumulation	SRTM
Basal area	Tier 2

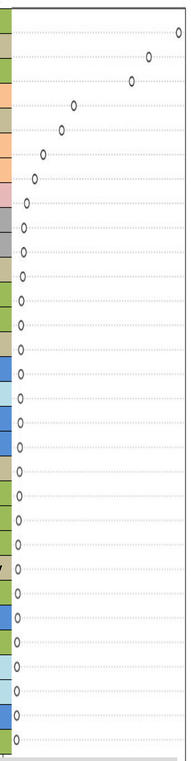

Figure 3. Variables of importance when predicting soil C to 1m depth using only ancillary data (soil data not explicitly measured at the Tier 2 network).

Concluding Remarks

Although climate is an important factor in determining soil C at global and regional scales, other factors proved more important for predicting soil C at the regional scale. This is explained by 1) scaling considerations and 2) ecosystem properties. In regards to scaling, climate varies much more at global scales compared to smaller, regional scales, therefore reducing it's prediction power in this model. Additionally, many of the top predictors, such as MLRAs and eco-regions, are ecological factors, and are not only functions of their biotic components, but also abiotic components such as soil, water, **and climate**. Therefore, it would be wrong to assume that climate is not an important factor for soil C at regional scales. However, ecosystem properties, which include climate information, proved to be most relevant for predicting soil C.

PINEMAP is funded by the USDA –NIFA Award # 2011-68002-30185