Which covariates are needed for soil carbon models in Florida?

X. Xiong & S. Grunwald
Department of Soil and Water Science, University of Florida, Gainesville, Florida, USA

D.B. Myers
USDA-ARS-Cropping Systems and Water Quality Unit, Columbia, Missouri, USA

J. Kim & W.G. Harris
Department of Soil and Water Science, University of Florida, Gainesville, Florida, USA

N.B. Comerford
North Florida Research and Education Center, University of Florida, Quincy, Florida, USA

ABSTRACT: In soil-landscape modeling, selection of optimal covariates has always been a critical question facing modelers. However, this topic has not been fully investigated. In this study a total of 1,192 soil samples in the topsoil (0-20 cm) across Florida were taken between 2008 and 2010 and a comprehensive pool of 212 environmental covariates covering all STEP - AWBH variables representing - S: soils, T: topography, E: ecology, P: parent materials, A: atmosphere/climate, W: water, B: biota, and H: human - were compiled. Data mining and machine learning techniques (Regression Tree, Bagged Regression Tree, Boosted Regression Tree, Random Forest and Support Vector Machine) were used to develop models to predict total soil carbon (TC) stocks. Results showed that soil-water properties, biota, human and parent material were the dominating factors controlling TC variation in Florida. A simplified Random Forest model with approximate 50 predictors performed comparable to the exhaustive model with all 212 predictors.

1 INTRODUCTION

Multi-variate soil-landscape modeling has been widely used to investigate soil carbon patterns and processes as well as upscale site-specific observations (McBratney et al., 2003; Grunwald, 2009). Soil carbon variation is complex to model because it is governed by various soil forming, environmental and anthropogenic factors that operate at distinct scales (Vasques et al., 2012). However, limited research has been conducted on systematically studying which covariates should be used in soil carbon modeling at a specific scale and region.

With the advent of the information age, the innovative delivery of remote sensing and proximal sensing products has been increasingly providing the soil science community with more accessible dataset characterizing environmental soil-landscapes and even internal soil information directly (Grunwald, 2009). Meanwhile, the rapid development and introduction of data mining and machine learning techniques equip pedometricians with powerful tools that can deal with huge volume of data. Tree-based modeling techniques and Support Vector Machine are among the most widely used in ecology and soil science (De'ath & Fabricius, 2000; Prasad et al., 2006; Vasques et al., 2008).

Therefore, the objective of this study was to identify the critical soil, environmental and anthropogenic covariates that can make the best prediction of soil carbon at regional scale (Florida, USA, about 150,000 km^2) from a huge pool of covariates, as well as reveal the dominating properties/processes controlling soil carbon in Florida.

2 MATERIALS AND METHODS

2.1 *Study area*

The study area is the State of Florida, located in the southeastern United States, with latitudes from 24°27′ N to 31° N and longitudes from 80°02′ W to 87°38′ W. The whole of Florida covers approximately 150,000 km^2.

The climate of North and Central Florida is humid subtropical. South Florida has a tropical climate according to the Koppen Classification Map. Dominant soil orders of Florida are: Spodosols (29%), Entisols (20%), Ultisols (17%), Alfisols (12%) and Histosols (10%). Overall, soils in Florida are sandy in texture. Land use/land cover consists mainly of Open Water (18%), Pinelands (16%), High Impact Urban (7%), Improved Pasture (7%), and Freshwater Marsh and Wet Prairie (5%). The topography consists of gentle slopes varying from 0 to 5% in almost the whole state. Elevation ranges from sea level up to approximately 114 m at the Panhandle.

2.2 *Soil data and environmental covariates*

A total of 1,192 soil samples in the topsoil (0-20 cm) across Florida were taken between 2008 and 2010

based on random design stratified by the combination of soil suborder and land cover / land use (Figure 1). Total carbon (TC) was analyzed by combustion catalytic oxidation (Shimadzu SSM-5000a). The laboratory TC measurements in mass unit (mg kg^{-1}) were converted to stock units (kg m^{-2}) using the measured bulk density and soil depth (20 cm). The whole dataset was randomly divided into calibration dataset (70%) and validation dataset (30%) for model development and assessment, respectively.

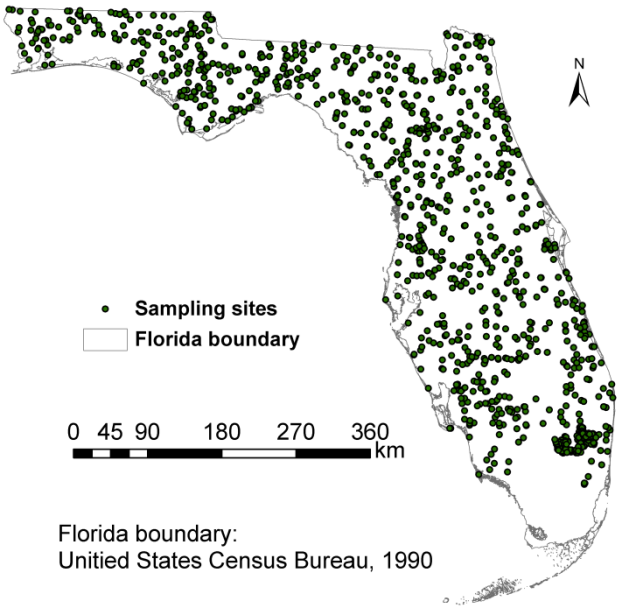

Figure 1. Soil sampling sites in Florida, USA

An unprecedented amount of environmental covariates (212 variables) covering all STEP - AWBH covariates representing - S: soils, T: topography, E: ecology, P: parent materials, A: atmosphere/climate, W: water, B: biota, and H: human - were compiled from various data sources with ArcGIS (Grunwald et al., 2011) (compare Table 1).

2.3 *Modeling methods*

Various data mining techniques were used to identify and model the relationships between TC stocks and environmental covariates, i.e., Regression Trees (RT), Bagged Regression Trees (BaRT), Boosted Regression Trees (BoRT), Random Forest (RF), and Support Vector Machine (SVM). Models were built in R 2.13.1 (Package rpart, ipred, gbm, randomForest and e1071). The models were assessed by comparing the model predicted values with the independent validation samples using the coefficient of determination (R^2), root mean squared deviation of the independent validation dataset (RMSD),

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}} \quad (1)$$

and residual prediction deviation (RPD) (Williams, 1987),

$$RPD = \frac{SD}{RMSD\sqrt{n/(n-1)}} \quad (2)$$

where, $\hat{y}_i$ are the model predicted values, y are the observed values, n is the number of predicted or observed values in the validation dataset with $i = 1, 2, \ldots, n$, and SD is the standard deviation of the validation dataset.

Table 1. Compiled environmental covariates.

STEP – AWBH factors	Data sources	Number of variables
Soil / water	USDA – NRCS's soil survey / state geographic database (Soil Data Mart)*	31
Topography	USGS's national elevation dataset*	4
Topography	USGS's digital elevation model	8
Ecology	FDEP's ecoregions*	1
Parent material	USGS's physiography	2
Parent material	USGS's geology	4
Climate	NREL's spectral solar radiation data base*	14
Climate / water	PRISM climate dataset*	39
Biota	LANDFIRE's national vegetation products*	14
Biota	National land cover data	3
Biota	National biomass & carbon dataset	3
Biota	Landsat ETM*	18
Biota	MODIS for NACP*	64
Biota / human	National cropland data layer	1
Biota / human	FFWCC's land cover / land use	2
Geography / topography / water	Other	4

* USDA – NRCS: U.S. Department of Agriculture – Natural Resources Conservation Service, USGS: U.S. Geological Survey, FDEP: Florida Department of Environmental Protection, NREL: National Renewable Energy Laboratory, PRISM: Parameter-elevation Regressions on Independent Slopes Model, LANDFIRE: Landscape fire and resource management tools project, ETM: Enhanced Thematic Mapper, NACP: North American Carbon Program, FFWCC: Florida Fish and Wildlife Conservation Commission.

2.4 *Model simplifying method*

Since the exhaustive model that uses all 212 predictors is too complex to be applied to make spatial predictions across the whole study area, a simplified RF model that uses only the most important predic-

tors in the exhaustive RF model was developed. The "most important predictors" were identified by iteratively eliminating the least important predictor in the RF model until the model reached a threshold at which its performance degraded significantly.

3 RESULTS AND DISCUSSION

3.1 *Descriptive statistics of TC in Florida*

The TC stock values in the top 20 cm soil largely ranged from 0.45 kg m^{-2} to 34.15 kg m^{-2} with a mean of 5.62 kg m^{-2} and a median of 3.81 kg m^{-2}. The standard deviation of TC was 4.66 kg m^{-2} and the data were strongly positively skewed (1.99). The distribution of TC could not be well transformed to a Gaussian shape by log transformation due to high TC values, which restricted the use of parametric modeling techniques.

3.2 *Models*

All the five data mining techniques yielded fairly good model performance with that more than 60% of TC variation in the validation dataset could be explained by the models (Table 2), despite the large variation of TC in Florida as shown above. Some variation of model performance could be found among the five models. The RF model ranked the best in terms of R^2, RMSD, and RPD, while the BaRT model had fairly close performance compared with the RF model. However, another ensemble tree method – BoRT ranked last which was even worse than the single tree method – RT. SVM generally performed not as well as ensemble tree-based modeling techniques.

Table 2. Total carbon model validation.

Model	R^2	RMSD	RPD
		kg m^{-2}	
Regression Tree	0.62	2.96	1.62
Bagged Regression Tree	0.72	2.56	1.87
Boosted Regression Tree	0.67	2.80	1.71
Random Forest	0.73	2.48	1.93
Support Vector Machine	0.63	3.07	1.56

However, it is difficult to draw a firm conclusion on which modeling technique is superior to another simply based on the error metrics. The major reason is that a "suitable" predictive digital model should also be easy to be interpreted and populated to the mapped area. In this sense, the RT model is preferable over ensemble tree models which are more complex and computationally intensive when making a soil map.

3.3 *Factors controlling soil carbon variation*

Table 3 shows the variable importance of the most important predictors in the RF model that was identified to perform best.

Table 3. Top 20 important predictors in the exhaustive Random Forest model.

STEP – AWBH factors	Variables	Relative importance *
		%
Soil	Soil organic matter (20cm)	12.2
Soil / water	Soil available water storage (50cm)	8.7
Soil	Soil great group	7.8
Biota / human	Land cover / land use	5.4
Soil / water	Soil available water storage (25cm)	4.0
Soil	Soil suborder	3.9
Biota	Existing vegetation cover	2.6
Soil	Soil silt content (20cm)	2.6
Biota	Existing vegetation type	2.5
Biota	Land cover class	2.4
Soil	Soil sand content (20cm)	2.2
Biota / human	Cropland data layer	2.2
Parent material	Surficial geology	2.1
Soil	Soil clay content (20cm)	2.0
Biota / human	Land cover / land use (reclassified)	1.8
Soil	Soil reaction class	1.5
Biota	Environmental site potential	1.5
Biota	Existing vegetation type group	1.2
Soil	Soil order	1.1
Biota	Biophysical settings	0.9

* Relative importance was calculated as the percentage of the absolute importance index which is defined as the total decrease in node impurities measured by residual sum of squares from splitting on the variable, averaged over all trees.

In general, soil properties, soil-water, soil taxonomic variables, biota and human variables dictated the model, while parent material also accounted for some TC variation. It indicates that the existing NRCS soil database contains much useful information about TC variation. In particular, it is not surprising that historical soil organic matter played a dominant role in the model because a major part of TC in Florida soils is in the form of organic carbon (Vasques et al., 2012). Among all of the STEP - AWBH factors, the soil-water content was a key factor that could explain much of the TC variation as indicated by two soil available water storage variables ranked high as predictors in the model (Table 3). In addition, soil taxonomic variables, e.g., soil

suborder and great group, which contains mixed information of soil moisture condition, texture, and more accounted for much TC variation. It is worth mentioning that land cover / land use was also one of the key players in the model along with other biota variables, e.g., existing vegetation indices. Considering anthropogenic-induced land cover / land use changes in the future, we could also expect change in soil carbon. Loss of highly organic wetland soil and deforestation as a result of urbanization would possibly cause the loss of soil carbon. On the other hand, restoration of wetlands and forests might help accumulate carbon in soils (Lal, 2003; Chmura et al., 2003). Furthermore, soil texture was an important factor affecting soil carbon patterns. It is also interesting to observe that surficial geology was selected by the model indicating that parent materials could explain some of the TC variation as well. In contrast to climatic variables, such as temperature and precipitation, that did not demonstrate predictive power to infer on soil TC.

3.4 *Model simplification*

As discussed above, besides good performance, a predictive model should also provide ease to be populated and applied to the mapping area. Applying a complex model with 212 predictors to a large area like Florida is a computationally prohibitive task. Therefore, it is desirable to reduce the number of predictors and simplify the model once knowing the key predictors in the model.

In Figure 2, it is interesting to observe that with approximate 50 most important variables the RF model could achieve equivalent performance as the exhaustive model which used all 212 predictors in terms of all three error metrics. This is a straightforward method for model simplification since it greatly reduced the number of predictors without sacrificing model performance. However, it does not necessarily indicate the remaining 162 predictors had no predictive power because their predictive power might have been suppressed by more powerful predictors. More useful model simplification strategies may lie in how to take advantage of these "less important" variables.

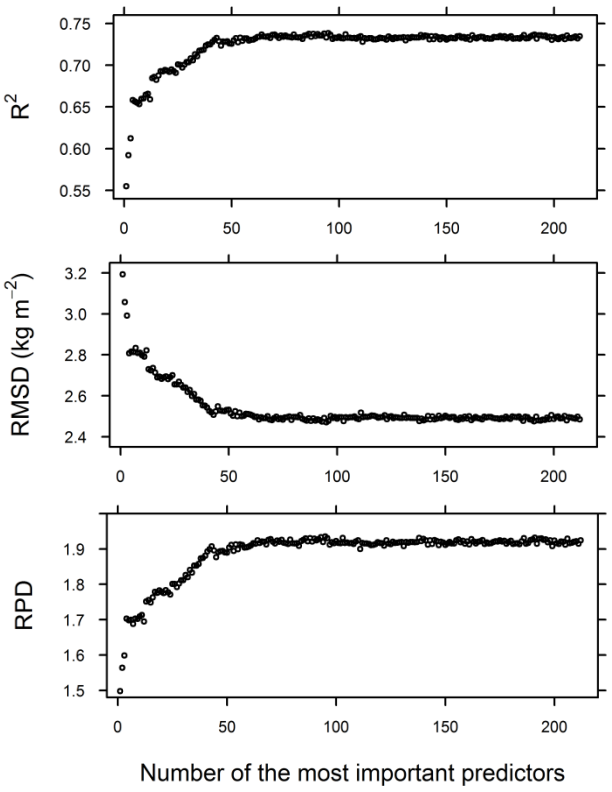

Figure 2. Change of Random Forest model performance as the model eliminates the least important predictor one at each time (view from left of the horizontal axis to the right)

4 CONCLUSIONS

This study showed the large variation of TC stock in Florida, USA. Our models captured the controlling factors and forcings of TC variation in Florida. The most important factors entailed soil-water, biota, human, and parent material. It indicated that the current readily available national soil database contains much critical information that can be used to model soil carbon. The importance of land cover / land use variables indicated that human activities, such as the change of land use, might have great influence on soil carbon stock. The exhaustive models that used a comprehensive environmental covariate dataset were successfully developed. All the five data mining and machine learning techniques produced good TC models, while the RF model performed the best, followed closely by the BaRT model, then the BoRT model, the SVM model, and the RT model. In order for the models to be applied to map soil carbon in Florida, a straightforward model simplification method showed that it was possible to use much fewer predictors (~50) in the model without jeopardizing model performance. However, it may be noted that in this study the starting point was an exhaustive set of potential environmental covariates (n=212) that was reduced to a parsimonious model with 50 variables to predict soil TC stocks. In essence the attribute space of a complex model was reduced to develop a more simplified model that still predicts

TC at the same level as the complex model. Internal pruning techniques aiming to reduce the complexity of a given set of predictor variables in an RT or ensemble tree model are common. But pruning is constraint to fixed attribute space used to build a model. Furthermore, the user specific bias of preselected environmental covariates to build a soil prediction model is inherent in many digital soil mapping studies. This has implications that were addressed in the presented study, which provides future investigators guidance to assemble those covariates that control soil TC in Florida and possibly similar subtropical and tropical soil-landscapes, such as the southern coastal plain of the U.S.

ACKNOWLEDGEMENTS

This study was funded by USDA-CSREES-NRI grant award 2007-35107-18368 "Rapid Assessment and Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape" (National Institute of Food and Agriculture (NIFA) – Agriculture and Food Research Initiative (AFRI)). The authors would like to thank Aja Stoppe, Christopher Wade Ross, Samiah Moustafa, Lisa Stanley, Adriana Comerford, Xiaoling Dong, and Anne Quidez for their hard work in field soil sampling and lab analyses.

REFERENCES

Chmura, G.L., Anisfeld, S.C., Cahoon, D.R., & Lynch, J.C., 2003. Global carbon sequestration in tidal, saline wetland soils. *Global Biogeochemical Cycles* 17, 1111–1122.

De'ath, G., & Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81, 3178–3192.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma* 152, 195–207.

Grunwald, S., Thompson, J.A., & Boettinger, J.L., 2011. Digital soil mapping and modeling at continental scales: finding solutions for global issues. *Soil Science Society of America Journal* 75, 1201-1213..

Lal, R., 2003. Global potential of soil carbon sequestration to mitigate the greenhouse effect. *Critical Reviews in Plant Sciences* 22, 151–184.

McBratney, A.B., Mendonça Santos, M.L., & Minasny, B., 2003. On digital soil mapping. *Geoderma* 117, 3–52.

Prasad, A.M., Iverson, L.R., & Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9, 181–199.

Vasques, G.M., Grunwald, S., & Myers, D.B., 2012. Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. *Landscape Ecology* 27, 355–367.

Vasques, G.M., Grunwald, S., & Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* 146, 14–25.

Williams, P.C. 1987. Variables affecting near-infrared refl ectance spectroscopic analysis. In P. Williams and K. Norris (eds), *Near-infrared technology in the agricultural and food industries*: 143–167. St. Paul, Minnesota: American Association of Cereal Chemists.