Geoderma 146 (2008) 102–113

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis

R. Grimm [a,*], T. Behrens [b], M. Märker [a], H. Elsenbeer [a,c]

[a] Institute of Geoecology, University of Potsdam, Karl-Liebknecht-Straβe 24–25, 14476 Potsdam, Germany
[b] Institute of Geography, Eberhard-Karls-University of Tübingen, Rümelinstraβe 19–23, 72070 Tübingen, Germany
[c] Smithsonian Tropical Research Institute, Balboa, Panama

ARTICLE INFO

Article history:
Received 27 October 2007
Received in revised form 20 April 2008
Accepted 15 May 2008
Available online 1 July 2008

Keywords:
Soil organic carbon
Digital soil mapping
Random Forest
Spatial pattern
Tropical soils
Barro Colorado Island

ABSTRACT

Spatial estimates of tropical soil organic carbon (SOC) concentrations and stocks are crucial to understanding the role of tropical SOC in the global carbon cycle. They also allow for spatial variation of SOC in environmental process models. SOC is spatially highly variable. In traditional approaches, SOC concentrations and stocks have been derived from estimates for single or very few profiles and spatially linked to existing units of soil or vegetation maps. However, many existing soil profile data are incomplete and untested as to whether they are representative or unbiased. Also single means for soil or vegetation map units cannot characterize SOC spatial variability within these units. We here use the digital soil mapping approach to predict the spatial distribution of SOC. This relies on a soil inference model based on spatially referenced environmental layers of topographic attributes, soil units, parent material, and forest history. We sampled soils at 165 sites, stratified according to topography and lithology, on Barro Colorado Island (BCI), Panama, at depths of 0–10 cm, 10–20 cm, 20–30 cm, and 30–50 cm, and analyzed them for SOC by dry combustion. We applied Random Forest (RF) analysis as a modeling tool to the SOC data for each depth interval in order to compare vertical and lateral distribution patterns. RF has several advantages compared to other modeling approaches, for instance, the fact that it is neither sensitive to overfitting nor to noise features. The RF-based digital SOC mapping approach provided SOC estimates of high spatial resolution and estimates of error and predictor importance. The environmental variables that explained most of the variation in the topsoil (0–10 cm) were topographic attributes. In the subsoil (10–50 cm), SOC distribution was best explained by soil texture classes as derived from soil mapping units. The estimates for SOC stocks in the upper 30 cm ranged between 38 and 116 Mg ha^{-1}, with lowest stocks on midslope and highest on toeslope positions. This digital soil mapping approach can be applied to similar landscapes to refine the spatial resolution of SOC estimates.

1. Introduction

Soils store about three times more organic carbon than is held in the plant biomass of terrestrial ecosystems and about twice as much than is current in the atmosphere (Batjes and Sombroek, 1997). Tropical soils contain about 26% of the soil organic carbon (SOC) stored in the soils of the world (Batjes, 1996). Global environmental conditions such as climate, biochemical cycles and vegetation are related to SOC. In order to understand the role of tropical SOC in the global carbon cycle as well as to incorporate variations of SOC into environmental process modeling, accurate estimates of the amount of SOC with high spatial resolution are necessary.

One common way of deriving the spatial distribution of soil is the analysis of the factors controlling soil formation. Jenny (1941) described soil as a function of climate, organisms, topographic relief, parent material, and time. Within digital soil mapping (also called soil-landscape modeling (Gessler et al., 1995) or predictive soil mapping (Scull et al., 2003) a Jenny-like approach is used but rather for quantifying than for explaining spatial soil class/property distribution. Digital soil mapping is characterized by formulating empirical spatial or non-spatial soil inference systems between soil observations and spatially referenced environmental "scorpan" factors (*s*oils and/or soil properties, *c*limate and/or climate properties, *o*rganisms like flora and fauna and human activities, *r*elief settings, *p*arent material, *a*ge, and spatial coordinate *n*) (McBratney et al., 2003).

A broad range of statistical methods have been applied towards digital soil mapping of SOC or organic matter (OM). Most commonly, multiple- and linear regression have been used for spatial quantifications of SOC/OM (Moore et al., 1993; Arrouays et al., 1995; Chaplot et al., 2001; Florinsky et al., 2002; Powers and Schlesinger, 2002; Thompson and Kolka, 2005; Thompson et al., 2006). This modeling technique has several advantages, such as simplicity in application and ease of interpretation (Hastie et al., 2001). Fewer studies used generalized linear models (McKenzie and Ryan, 1999), tree models (Kulmatiski et al., 2004; Henderson et al., 2005) or artificial neural networks (Minasny et al.,

* Corresponding author. Tel.: +49 179 4607517.
E-mail address: rosigrimm@gmx.de (R. Grimm).

0016-7061/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2008.05.008

2006) for relating SOC and OM storage to environmental predictors. The latter techniques have the potential for discovering non-linear relationships and might therefore prove more powerful for digital SOC mapping. From the field of machine learning, ensemble approaches like bagging (Breiman, 1996), boosting (Freund and Schapire, 1996) or Random Forest (Breiman, 2001) could be applied for SOC prediction in order to enhance prediction accuracy. These approaches, however, have not yet been reported in SOC prediction literature. Within geostatistics soil forming factors in terms of ancillary environmental predictors can be used to estimate the spatial distribution of SOC/OM through kriging with external drift or co-kriging (Simbahan et al., 2006). Bhatti et al. (1991), Hengl et al. (2004) as well as Simbahan et al. (2006) applied regression kriging to predict SOC or OM.

Random Forest (RF), a new method of data mining, has several advantages compared to most of the modeling techniques mentioned above, such as (Breiman, 2001; Liaw and Wiener, 2002): Ability of modeling high dimensional non-linear relationships, handling of categorical and continuous predictors, resistance to overfitting, relative robustness with respect to noise features, implemented unbiased measure of error rate, implemented measures of variable importance, and only few user defined parameters.

We hypothesized that terrain-driven hydrological flow patterns and mass-movement are the dominating processes responsible for SOC redistributions. We furthermore assumed that SOC storage on Barro Colorado Island (subsequently referred to as BCI) is related to the spatial distribution of soil properties such as soil texture, color, and mineralogy, as well as geology and forest history.

The aim of this study was to propose a RF-based digital SOC mapping framework from which knowledge of soil processes can be derived. Using this framework we were able to estimate the SOC concentrations and stocks on BCI in the spatial domain more realistically than by simply relating mean SOC values to soil mapping units as has been the traditional approach.

2. Methods

2.1. Study site

Barro Colorado Island (BCI) (Fig. 1) (9°9′N, 79°51′W) was formed by the flooding of Lake Gatun in the Panama Canal basin in 1914. The 1500 ha former hilltop rises 137 m above the lake level. The climate is tropical with a mean annual temperature of 27 °C. The annual precipitation averages 2600 mm with 90% of the rainfall occurring in the wet season between May and December (Dietrich et al., 1982). BCI is entirely covered by semi-deciduous lowland moist tropical forest. Parts of the island were cleared for agricultural purpose before and during the creation of the Panama Canal. The southwest of the island is old growth which has not been disturbed for at least 200 years (Leigh, 1999), whereas the northeast is younger regrowth with 100 or more years in age (Foster and Brokaw, 1996).

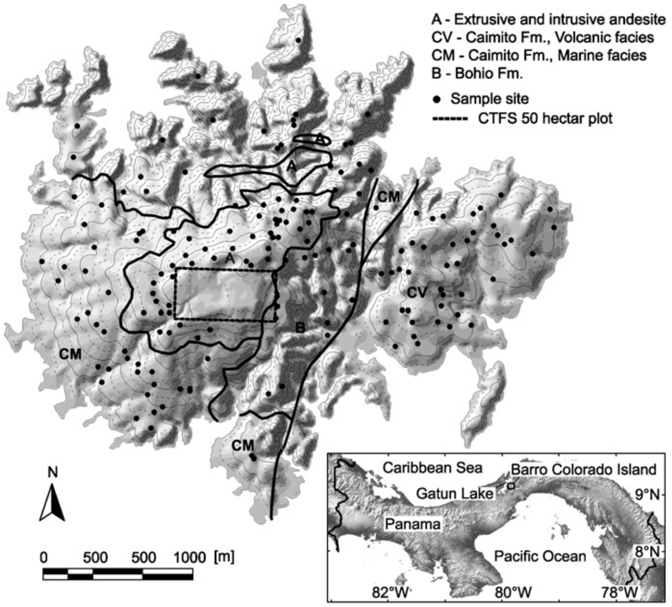

Fig. 1. Hillshade of Barro Colorado Island derived from the digital elevation model superimposed on the locations of sampling sites, contour lines, and the geological map (Johnsson and Stallard, 1989). Inset: Barro Colorado Island's location in Panama.

The geology (Fig. 1) consists of two main formations: the Bohio dating back to the early Oligocene (Ministerio de Comercio e Industrias, 1976) and the younger Caimito formation from the late Oligocene (Woodring, 1958). Both formations are sedimentary and each consists of two facies: volcanic and marine. In addition, there are extrusive and intrusive igneous rocks from the Oligocene and early Miocene age (Johnsson and Stallard, 1989). The main extrusive component is an andesite flow, which caps the island (Johnsson and Stallard, 1989) forming a flat, slightly tilted hilltop. The most obvious structural feature is the sinistral strike-slip fault system that trends NNE–SSW across the centre of the island (Fig. 1).

The dominant soils on BCI are immature Cambisols, and most of the more mature soils are Ferralsols (Fig. 5; Table 1; WRB, 2006). They are clay- and nutrient-rich and contain high amounts of calcium, magnesium, nitrogen, and potassium, but presumably low amounts of phosphorus (Dietrich et al., 1982; Yavitt et al., 1993; Yavitt, 2000; Barthold et al., 2008). Until now, SOC estimates for BCI were limited to parts of the island using only few samples and limited spatial coherence (Yavitt et al., 1993; Yavitt, 2000; Yavitt and Wright, 2002).

Soil clay mineralogy of the andesite plateau and the Caimito volcanic facies is dominated by kaolinite. Furthermore, the deep red clays on the Bohio formation and the Caimito marine facies are also dominated by kaolinite. The loams of the Bohio formation and the Caimito marine facies, however, also contain substantial amounts of smectite. On the other hand, the pale swelling clays on all geological units are dominated by smectite (Baillie et al., 2006). Johnsson and Stallard (1989) related the presence of smectites in this highly weathered environment to the rapid erosion on steep slopes leading to shallow soils with short residence times of minerals.

2.2. Data

2.2.1. Soil organic carbon

2.2.1.1. Soil sampling. In order to analyze and predict the amount and distribution of SOC most efficiently, we established a design-based, stratified, two-stage sampling plan (McKenzie and Ryan, 1999) with topography and geology as the stratifying variables. A digital quantification of catenary landscape position was calculated from the digital elevation model (DEM) by combining the compound topographic index (CTI) (Beven and Kirkby, 1979) with the projected distance to stream (PD02) (Behrens, 2003). Lower slope positions within a distance of 100 m of Gatun Lake were not considered in this quantification because these represent former mid-slope positions before the flooding of the Panama Canal, and hence do not show classical catenary soil attributes (Baillie et al., 2006). The 50 ha plot of the Centre for Tropical Forest Science (CTFS) in the centre of BCI was also excluded from the sampling scheme (Fig. 1), in order to avoid disturbance.

We calculated four CTI classes and three PD02 classes on an equal area basis. Those classes were combined and the resulting 12 distinct environments were further stratified into four geological units. From those 48 distinct environments we randomly selected three replicate sites. Additionally, 21 sites were chosen randomly from all distinct environments in order to enlarge the sampling size, resulting in a total of 165 sites.

Soil samples were taken during the beginning of the wet season, between June and September 2005. At each site a 50 cm deep soil

Table 1
BCI soil mapping units and corresponding geological units (A: andesite flow, B: Bohio formation, CM: Caimito marine facies, CV: Caimito volcanic facies), mean field description of topsoil (T) and subsoil (S) depth, texture, color, and correlation to WRB soil classification system[a]

Soil mapping unit (abbreviation)	Geological unit	Mean depth [m]	Mean texture[b]	Most frequent Munsell soil color	WRB[a] correlation
Ava (A)	A	T: <0.04	sicL	7.5YR 3/3, 7.5YR 3/4	Hypereutric & Haplic Ferralsol
		S: >2	siC, C	5YR 4/4, 5YR 4/6	
Marron (M)	A	T: <0.05	sicL	7.5YR 3/2, 7.5YR 3/4	Leptic & Eutric Cambisol
		S: <1	siC	7.5YR 4/6, 5YR 4/4	
Lake (L)	A	T: <0.03	siC	7.5YR 3/4	Vertic Luvisol & Acrisol & Vertic Eutric or Alumic Gleysol
		S: <1.5	C	Mottled 10YR 6/3	
Swamp (Sw)	A	T: <5	cL	7.5YR 3/1	Mollic, Eutric & Haplic Gleysol
		S: >1	C, siC	Mottled 2.5Y 4/3, 2.5Y 5/1	
Fairchild (F)	B	T: <0.04	sicL	7.5 YR 3/3	Leptic & Eutric Cambisol
		S: <0.6	siC	2.5YR 4/4, 2.5YR 4/6	
Standley (S)	B	T: <0.15	siC	7.5YR 3/2, 5YR 3/3	Leptic & Eutric Cambisol
		S: <0.5	sicL, siC	7.5YR 4/4, 5YR 4/4	
Gross (G)	B	T: <0.04	sicL	7.5 YR 3/3	Vertic Luvisol & Acrisol & Vertic Eutric or Alumic Gleysol
		S: >2	C, siC	Mottled 5Y 6/2	
Poacher (P)	CM	T: <0.08	sicL	7.5 YR 3/3, 5YR 3/3	Hypereutric & Haplic Ferralsol
		S: >2	sicL, siC	5YR 4/6, 2.5YR 4/6	
Wetmore (W)	CM	T: <0.1	cL, sicL	7.5 YR 3/3	Leptic & Eutric Cambisol
		S: <1	sicL	7.5YR 4/4, 5YR 4/4	
Lutz (Lu)	CM	T: <0.1	sicL, siC	7.5 YR 3/3, 5YR 3/3	Ferric & Hypereutric Ferralsol
		S: <1	siC, C	7.5YR 4/4, 5YR 4/4	
Zetek (Z)	CM	T: <0.05	siC	7.5YR 3/2	Vertic Luvisol & Acrisol & Vertic Eutric or Alumic Gleysol
		S: >2	C, siC	Mottled 2.5Y 7/2, 5Y 7/3	
Harvard (H)	CV	T: <0.03	cL, sicL	7.5 YR 3/3, 5YR 3/3	Hypereutric & Haplic Ferralsol
		S: >1.5	siC, C	5YR 4/6, 2.5YR 4/6	
Hood (Ho)	CV	T: <0.05	cL, sicL	7.5 YR 3/3, 5YR 3/3	Leptic & Eutric Cambisol
		S: <0.5	cL, sicL	7.5YR 3/4, 5YR 4/4	
Barbour (B)	CV	T: <0.05	sicL	7.5YR 3/2	Vertic Luvisol & Acrisol & Vertic Eutric or Alumic Gleysol
		S: >2	C, siC	Mottled 5Y 6/3	

[a] WRB (2006).
[b] For texture classes see FAO (2006).

profile was dug and a soil sample of 250 g taken at the depth intervals of 0–10, 10–20, 20–30 and 30–50 cm.

2.2.1.2. Laboratory analyses. The samples were oven-dried at 60 °C and passed through a 2 mm sieve; recognizable undecomposed OM particles were removed. A sub-sample of about 20 g was finely ground and dried to constant weight at 105 °C. Total carbon was measured by dry combustion using at least three 20 mg sub-samples form each sample until the coefficient of variation of replicate measurements was below 0.05. According to Baillie et al. (2006), carbonates were not to be expected in these soils on account of their low pH; therefore we assumed that total carbon equals organic carbon. SOC stock (also called SOC storage, SOC pool or SOC density), i.e. carbon mass per unit area for a given depth, was calculated according to:

$$\mathrm{SOC_{stock}} = \mathrm{SOC_{conz}} \times \rho \times (1{-}ST) \times \Delta d \times UFC \qquad (1)$$

where $\mathrm{SOC_{stock}}$ is soil carbon stock (kg $\mathrm{ha^{-1}}$), $\mathrm{SOC_{conz}}$ is soil carbon concentration (%), ρ is bulk density of the fine earth (kg $\mathrm{m^{-3}}$), ST (stoniness) is the volumetric percentage (vol.%) of fragments of >2 mm, Δd is the thickness of the layer (m), and UFC is a unit conversion factor (100 $\mathrm{m^2\ ha^{-1}}$). Carbon mass per unit area for a given depth was calculated by summing $\mathrm{SOC_{stock}}$ over all layers.

Bulk density was determined for 24 sites by the compliant cavity method (Soil Survey Staff, 1996). Baillie et al. (2006) determined stoniness based on field estimates of volume percentage of fragments >2 mm.

As the absolute uncertainty of $\mathrm{SOC_{stock}}$ is a function of individual uncertainties (Eq. (1)) — we assumed that these factors are independent — the error propagation rule for multiplications of independent factors was used for its determination (Taylor, 1997). For cumulative $\mathrm{SOC_{stock}}$ total uncertainty for a given depth was calculated by using the error propagation equation for summations of independent summands (Taylor, 1997).

2.2.2. Environmental predictors

2.2.2.1. Topography. Topography has the potential to explain large parts of the variation of SOC. Thus, models accounting for terrain attributes can provide more realistic estimates of SOC pools. Terrain attributes, the most extensively used environmental predictors in digital soil mapping (McBratney et al., 2003), approximate water, solute, and sediment fluxes throughout the landscape. These are driven by gravity, solar insolation and micro-climate, and hence may control spatial patterns of soil properties such as SOC.

Terrain attributes were derived from the 5 m DEM of BCI, which is based on a topographic map in 1:25,000 scale with 10 m contour intervals (Kinner et al., 2002). A total of 13 terrain attributes were calculated and extended to 15 datasets by deriving additional variations (Table 2). Terrain parameters can be grouped into local, regional and combined terrain attributes. Local terrain attributes are based on a moving window technique with the same spatial extent for each cell. Regional terrain attributes are based on contributing area, and combined terrain attributes derived through combinations of local and regional attributes (Behrens, 2003).

2.2.2.2. Soil. The BCI soil map, whose taxa are based on geology, general topographic indicators, soil color and texture (Baillie et al., 2006), was used as soil factor within the digital SOC mapping framework (Fig. 5). Soil color relates to SOC with darker colors generally indicating higher SOC concentrations (Konen et al., 2003; Viscarra-Rossel et al., 2006). Soil texture, particularly soil clay content, is positively correlated to SOC (Arrouays et al., 1995; Powers and Schlesinger, 2002; Kahle et al., 2002). Additionally, a variety of authors (Van Breemen and Feijtel, 1990; Torn et al., 1997; Baldock and Skjemstadt, 2000; Six et al., 2002) propose that SOC stabilization is influenced by clay mineralogy, with 2:1 clays like smectite stabilising SOC to a greater extend than 1:1 clays like kaolinite.

Table 1 shows the local BCI soil units, some general field descriptions and the corresponding WRB (WRB, 2006) soil names. The Swamp soil unit was merged with Lake on the basis of parent material, soil color, and texture since there were no observations in the Swamp unit. All other soil units remained unchanged.

2.2.2.3. Geology. Soils are the weathering product of the parent material. Parent material was derived from the geological map of BCI (Fig. 1) (Woodring, 1958; Johnsson and Stallard, 1989). The andesite flow is a resistant and non-vesicular rock, with phenocrysts consisting primarily of

Table 2
Terrain attributes used for digital soil organic carbon mapping

	Terrain attribute	Abbreviation	Author
Local	Slope	SLT	Tarboton, 1997
	Horizontal curvature	CHOS	Shary et al., 2002
	Mean curvature	CMES	Shary et al., 2002
	Profile curvature	CPES	Shary et al., 2002
	Landform evolution	LEV	Nogami, 1995
Regional	Contributing area	CA	Dietrich and Montgomery, 1998
	Stream power index	SPI	Moore et al., 1991 (calculation based on Dietrich and Montgomery, 1998; Tarboton, 1997)
	Transport capacity	TC	Schmidt and Dikau, 1999 (calculation based on Dietrich and Montgomery, 1998; Tarboton, 1997)
	Relative hillslope position	RHP	Behrens, 2003
	Local elevation (for 0.2 ha and 0.5 ha)	LE02 LE05	Behrens, 2003
	Projected distance to stream (for 0.2 ha and 0.5 ha)	PD02 PD05	Behrens, 2003
Combined	Compound topographic (wetness) index	CTI	Beven and Kirkby, 1979 (calculation based on Dietrich and Montgomery, 1998; Tarboton, 1997)
	LS-factor	LS	Feldwisch, 1995 (calculation based on Dietrich and Montgomery, 1998; Tarboton, 1997)

plagioclase (Johnsson and Stallard, 1989). The main rock type on the Bohio volcanic facies is a conglomerate, which consists of basaltic clasts of all sizes (pebbles, cobbles and boulders) in a matrix of finer basaltic clasts (Woodring, 1958). The marine facies is interlayered with the conglomerate and consists of greywacke sandstone of poorly sorted angular basaltic coarse sand in a finely-grained matrix containing feldspars and some quartz (Woodring, 1958). The latter facies is not separately delineated in Fig. 1. The main constituents of the Caimito volcanic facies are a basaltic agglomerate and different kinds of greywacke, varying only in the degree of sorting of the grains. The Caimito marine facies primarily consist of foraminiferal limestone with abundant pelecypods and a large detrital component in the form of vitric volcaniclastic debris, plagioclase and quartz (Johnsson and Stallard, 1989). Furthermore, Fig. 1 does not display intrusive basaltic to basaltic andesitic dikes which can be found mainly within the Bohio formation and the volcanic facies of the Caimito formation (Johnsson and Stallard, 1989).

2.2.2.4. Forest history. The relationship between land use history and SOC was observed with SOC stocks being relatively lower in secondary compared to primary forests depending on type of past land use and forest age (e.g. Silver et al., 2000; Paul et al., 2002).

Svenning et al. (2004) derived the forest history from an 1927 aerial photograph of BCI by converting it into a grayscale grid. They delineated three forest history classes (old growth, tall secondary forest, low secondary forest) by sequentially grouping different grey values. Bright colors were assigned to be younger or cleared forests whereas dark colors are high forest areas. Due to the fact that no absolute time periods were assigned to each forest history class, forest history can only be an ordinal predictor for the spatial prediction of SOC concentrations and stocks.

2.3. Data pre-processing

Prior to modeling we identified and removed outliers from the SOC dataset, which were taken as values deviating >2× interquartile range away from the upper and lower quartile. This resulted in 3–5 SOC data point exclusions in each depth interval. Fig. 2 shows the SOC raw and pre-processed dataset.

2.4. Random Forest

Random Forest (RF) is an example of a machine learning method. RF consists of an ensemble of randomized classification and regression trees (CART) (Breiman, 2001). We assume familiarity with the basics of CART (Breiman et al., 1984). Numerous trees are generated within the algorithm and finally aggregated to give one single prediction. In regression problems the prediction is the average of the individual tree outputs, whereas in classification the trees vote by majority on the correct classification.

Within the training procedure, the RF algorithm produces multiple CART-like trees, each based on a bootstrap sample (sample with replacement) of the original training data. In addition to this normal bagging function (Breiman, 1996), the best split at each node of the tree is searched only among a randomly selected subset of predictors. All trees are grown to maximum size without pruning.

RF has several advantages over other statistical modeling approaches (Breiman, 2001; Liaw and Wiener, 2002). Its variables can be both continuous and categorical. The RF algorithm is quite robust to noise in predictors and thus does not require a pre-selection of variables (Diaz-Uriate and de Andres, 2006). As only a limited random number of predictors is used to search for the best split at each node, the diversity of the forest is increased (low correlation of individual trees) and the computational load is reduced. Pruning the trees is not necessary; it results in low bias and high variance trees and also saves computation time (Svetnik et al., 2003). As a large number of trees are averaged RF achieves both low bias and low variance (Diaz-Uriate and de Andres, 2006). The algorithm is robust to overfitting since each tree is trained on a unique bootstrap subsample of the data (Arun and Langmead, 2005). RF provides reliable error estimates by using the so called Out-Of-Bag (OOB) data (the proportion which is not used in the bootstrap subset — on average about on third of the data is excluded, while some others will be repeated in the sample), and thus eliminates the need for an independent validating dataset. The latter advantage should be of particular interest to soil science, since the collection of soil samples and laboratory analyses is in many cases time-consuming and expensive.

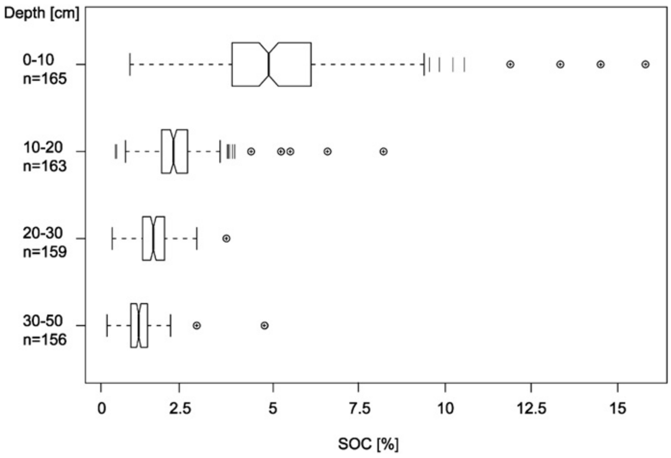

Fig. 2. Boxplots of soil carbon concentrations as a function of depth. The crossbar within the box shows the median, the length of the box reflects the interquartile range, the fences are either marked by the extremes if there are no outliers, or else by the largest and smallest observation that is not an outlier. Bars are outliers >1.5 times from the interquartile range away from the upper/lower quartile, whereas circles with a cross are outliers>2 times the interquartile range away from the upper/lower quartile. The notches represent the 95% confidence interval around the median.

RF depends only on three user defined parameters: the number of trees (n_{tree}) in the forest, the minimum number of data points in each terminal node (nodesize), and the number of features tried at each node (m_{try}). The default of n_{tree} is 500. However, more stable results of estimating variable importance (see below) are achieved with a higher number of n_{tree}(Diaz-Uriate and de Andres, 2006), thus we used n_{tree}= 1000. For nodesize we used the default for regression RF which is 5 instances in each terminal node. Concerning m_{try} the default for regression problems is one third of the total number of predictors. However, as RF prediction performance can be sensitive to m_{try} (Breiman and Cutler, 2004) we used an iterative approach to determine the best m_{try} in terms of smallest OOB mean square error (Eq. (2)). Within each depth interval we applied the RF algorithm with n_{tree} = 1000, nodesize = 5, and m_{try} values of 1, 3, 6, 9, 12, 15 and 18 with 100 replicate models for each m_{try} value (Section 3.2.1).

The main disadvantage of RF and ensemble algorithms in general is their limited interpretability. This algorithm is therefore often called a "black box" approach, since the relationship between predictors and response cannot be examined individually for every tree in the forest. In CART models, in contrast, a predictor variable in a single tree is related to the predictions. It should, however, be mentioned that single CART models are unstable in that minor changes in the training sample can lead to changes in the predictors, which are used for the splits. One should therefore be careful with drawing conclusions from single CART models concerning variable importance (Sutton, 2005). Moreover, as complexity in terms of the number of terminal nodes of single trees rises with increasing number of instances, interpretation of CART may become confusing.

The interpretation in RF is facilitated by two measures of variable importance. The first is the difference between the OOB error (Eq. (2)) of each tree and the same computed after permuting a predictor. The change in OOB error for each randomly permutated predictor gives an indication of the importance of this particular predictor. Random permutation should therefore have little effect on the estimated OOB error if a predictor is irrelevant. The second variable importance measure is the same as used in the CART algorithm and represents the total decrease in node impurity from splitting on the variable as measured for regression by the residual sum of squares and averaged over all trees. As the latter is computed on the training data its conclusion is based on overfitted models (Prasad et al., 2006). Hence, we only report the first variable importance measure.

Model performance is ideally addressed by using a large independent test dataset that was not used in the training procedure. When data is limited, k-fold cross-validation is often used. RF uses an extension of cross-validation, where each OOB sample is predicted by its corresponding bootstrap training tree. By aggregating the OOB predictions of all trees in the forest the mean square error (MSE) can be estimated (Liaw and Wiener, 2002):

$$\mathrm{MSE}^{\mathrm{OOB}} = \frac{\sum_{i=1}^{n}\left\{Y_i - \widehat{Y}_i^{\mathrm{OOB}}\right\}^2}{n} \tag{2}$$

Svetnik et al. (2003) showed that the OOB estimate of prediction accuracy yields results comparable to k-fold cross-validation. However, OOB estimates of error rate are computationally less expensive than standard k-fold cross-validation. As the MSE is scale dependent it cannot be used for comparing SOC model performance in different soil layers. Therefore, we additionally reported the normalized OOB mean square error ($\mathrm{NMSE}^{\mathrm{OOB}}$) which was calculated as:

$$\mathrm{NMSE}^{\mathrm{OOB}} = \frac{\mathrm{MSE}^{\mathrm{OOB}}}{\mathrm{Var}(Y_k)} \tag{3}$$

where Var is the total variance of carbon concentrations Y in the depth interval k.

In soil science in general and pedometrics in particular, RF has not yet to be applied widely as a modeling tool. The only spatial mapping applications so fare have been an investigation of risk mapping of tick-borne disease (Furlanello et al., 2003), the prediction of tree species distributions under future climate scenarios (Prasad et al., 2006), and remote sensing studies (Ham et al., 2005; Pal 2005; Gislason et al., 2006; Lawrence et al., 2006). However, RF has frequently been applied to non-spatial analyses in biology, biometrics, genetics and bioinformatics (Gunther et al., 2003; Svetnik et al., 2003; Bureau et al., 2003; Schwender et al., 2004; Parkhurst et al., 2005).

For all RF computations, we used the "RandomForest" package (Liaw and Wiener, 2002) for the *R* statistical language (R Development Core Team, 2006).

3. Results and discussion

3.1. Soil carbon concentrations and stocks

SOC concentrations decreased with depth as expected, and varied significantly between the observed depth intervals, but insignificantly between soil types (Fig. 2; Table 3). In general, high SOC concentrations translated into high carbon stocks. The pale swelling clays (Vertic Luvisol, Acrisol and Vertic Eutric or Alumic Gleysol) constitute an exception in that they have low SOC concentrations but high carbon stocks. This is due to their comparatively high bulk density and low stoniness. Within every soil class SOC stocks of the upper 10 cm were significantly higher than subsoil SOC stocks. The overall SOC stocks in the upper 30 cm was 72.61 Mg ha^{-1} and to a depth of 50 cm 92.72 Mg ha^{-1}.

Table 3
Soil organic carbon (SOC) concentrations and stocks (±1 SD)

Soil depth	All observations		Hypereutric, Haplic, Ferric Ferralsol[a]		Vertic Luvisol & Acrisol & Vertic Eutric or Alumic Gleysol[a]		Leptic, Eutric, Ferralic Cambisol[a]	
cm	*n*	%SOC	*n*	%SOC	*n*	%SOC	*n*	%SOC
0–10	161	5.00 (1.77)	18	4.38 (1.88)	39	5.35 (1.95)	104	4.97 (1.66)
10–20	158	2.13(0.64)	18	2.22 (0.53)	40	1.91 (0.68)	100	2.21 (0.62)
20–30	158	1.53 (0.46)	18	1.65 (0.46)	40	1.35 (0.41)	100	1.59 (0.46)
30–50	154	1.10 (0.35)	18	1.18 (0.35)	40	0.94 (0.26)	96	1.16 (0.37)
cm	*n*	Mg SOC ha^{-1}	*n*	Mg SOC ha^{-1}	*n*	Mg SOC ha^{-1}	*n*	Mg SOC ha^{-1}
0–10	161	38.05 (15.29)	18	33.27 (15.11)	39	45.04 (17.96)	104	35.98 (14.51)
10–20	158	17.84 (6.18)	18	17.95 (4.91)	40	18.11 (6.95)	100	17.67 (6.02)
20–30	158	13.57 (4.87)	18	13.61 (4.62)	40	13.76 (4.72)	100	13.49 (4.75)
30–50	154	21.02 (8.08)	18	21.80 (7.02)	40	20.71 (7.63)	96	20.87 (7.95)
0–30		69.46 (17.20)		64.83 (16.55)		76.91 (19.83)		67.14 (16.41)
0–50		90.48 (19.00)		86.62 (17.98)		97.62 (21.24)		88.01 (18.24)

[a] WRB (2006).

The standard deviation of SOC stocks was high (Table 3), as was also observed in most of the studies referred to in Table 4. These estimates of uncertainty are coarse at best and ignore the bias of site selection, since boulders and rock outcrops obstruct manual soil pit excavation. Several attempts may therefore be necessary before an observation can actually be recorded. Hence, the soil map may underestimate the spatial extent of rocky areas and stoniness on BCI.

Comparisons with other tropical regions (Table 4) show that SOC stocks on BCI are higher than the estimates of global tropical means (Post et al., 1982; Batjes, 1996; Amthor and Huston, 1998; Jobbagy and Jackson, 2000), as well as of the Brazilian Amazon (Batjes and Dijkshoorn, 1999; Bernoux et al., 2002; Cerri et al., 2003). Compared to Ecuador (Rhoades et al., 2000; de Koning et al., 2003), Mexico (Hughes et al., 1999) and Hawaii (Bashkin and Binkley, 1998), the SOC stock on BCI is significantly lower. Compared to Costa Rica (Powers and Schlesinger, 2002; Powers, 2004; Veldkamp et al., 2003; Powers and Veldkamp, 2005) and Puerto Rico (Brown and Lugo, 1990; Li et al., 2005), SOC stocks on BCI are lower. The ranges of differences, however, were narrower. These differences underline the strong influence of climate and ecosystem properties including soil properties, such as clay content and mineralogy.

3.2. Digital soil organic carbon mapping using Random Forests

3.2.1. Parameter optimization

In order to optimize RF prediction performance in terms of lowest OOB normalized mean square error (NMSE^{OOB}), we used an iterative approach with m_{try} model settings of 1, 3, 6, 9, 12, 15 and 18, each replicated 100 times (Fig. 3). As the total range and the differences between tested m_{try} settings were relatively small, with most changes occurring in the third position after decimal point, these differences were not relevant despite their frequent significance. Those parameter settings of m_{try} within our dataset, which were influencing prediction performance to less than the second decimal place of NMSE^{OOB}, were regarded as having an equal quality of prediction performance. In the topsoil the lowest m_{try} of 1 was the best model parameter setting, whereas in the soil layers 10–20 and 20–30 cm m_{try}=12 performed best with ranges of equal prediction performances of 6 to 18 and 9 to 18, respectively. Setting m_{try} equal to the total number of predictors corresponds to a normal bagging function (Breiman, 1996). Therefore, in the depth interval 10–20 and 20–30 cm the improvement of prediction accuracy by using RF instead of bagging was insignificant. Between 30 and 50 cm tested m_{try} values between 3 and 12 performed best, with the default m_{try} of 6 randomly selected features at each split revealing best prediction accuracies.

In correspondence with other studies (e.g. Svetnik et al., 2003; Diaz-Uriarte and de Andres, 2006), we suggest that the default of m_{try} is often a good choice. Higher numbers of m_{try} than the default value indicate that some noise variables are contained in the total set of environmental predictors. The best NMSE^{OOB} performance in the topsoil could be achieved with the lowest m_{try}m$_{try}$ values. This results in the highest randomness in feature selection at each node. Model performance evaluation by feature selection approaches (e.g. Behrens et al., 2007) is, however, beyond the scope of this study.

3.2.2. Model performance

Table 5 shows the RF prediction performance based on the OOB mean square error (MSE^{OOB}), OOB root mean square error (RMSE^{OOB}), and the OOB normalized mean square error (NMSE^{OOB}). In general model performances were limited. Prediction accuracy was on average lowest in the topsoil with NMSE^{OOB}=0.94 compared to the subsoil ranging between 0.75 and 0.91 in NMSE^{OOB}. These results suggest that particularly in the topsoil the spatial distribution patterns of SOC are highly variable due to small scale variations in input, redistribution, stabilization, as well as in intrinsic random variability of SOC. They are therefore difficult to approximate with state of the art soil-landscape modeling assessments of environmental layers. Furthermore, technical sources of uncertainties, as for instance the accuracy of the DEM and the localization of sampling sites with the global positioning system (GPS), limit the model performance. On the other hand, we did not find any residual spatial structure, and therefore, we could not adopt a regression–kriging strategy (Odeh et al., 1995) in order to improve prediction accuracy.

3.2.3. Variable importance

Variable importance revealed different dominating environmental features between topsoil (0–10 cm) and subsoil (10–50 cm) RF SOC models (Fig. 4). Regarding the topsoil on average erosion processes approximated by regional (e.g. contributing area (CA), relative hillslope position (RHP)), and combined terrain attributes (e.g. combined topographic index (CTI), LS-factor (LS)) were most relevant, followed by the local attributes like slope (SLT) and curvatures (CHOS, CMES, CPRS). The categorical predictors soil, geology, and forest history were of little importance for topsoil SOC prediction, suggesting that neither soil forming processes nor past land use changes influence the topsoil SOC distribution. The topsoil SOC is dependent on the present-day biomass input to soil, which, however, is not covered by forest history because past land use was derived form an aerial photograph of 1927. Since prediction performance is low in the topsoil (Section 3.2.2), variable importance is restricted in terms of interpretation.

Variable importance among predictors showed similar patterns in the subsoil below 10 cm. Similarly to the topsoil, topography had a strong impact on SOC predictions. Regional and combined parameters were more crucial than local terrain attributes. CTI, which is a proxy for soil moisture (Beven and Kirkby, 1979), was highly influential in the soil layer between 10 and 20 cm. The soil map was the most important predictor for the whole depth interval of 10–50 cm, indicating that soil texture and/or color determines the subsoil SOC distribution. As with topsoil, geology and forest history were weak predictors within the RF models. Between 30 and 50 cm the importance value of forest history was even below zero, indicating that random noise would be a better predictor in this soil depth.

Although certain predictors are more important within each RF model, we could not quantitatively determine their functional relationship to SOC. In this respect, spatial visualizations of prediction results were essential to understanding the driving processes behind SOC predictions (Section 3.2.4.).

3.2.4. Spatial prediction

We spatially predicted the SOC concentration in the depth intervals 0–10, 10–20, 20–30 and 30–50 cm using RF, respectively (Fig. 5a–d). In order to compare lateral and vertical SOC distribution patterns we computed the SOC stocks (Fig. 6a–d), since natural pedons include non-soil components such as rocks and pebbles. SOC stocks therefore reflect SOC distributions more realistically. As we did not have a spatial representation of neither bulk density nor stoniness that should be used for the spatial conversion of SOC concentrations (Fig. 5) to SOC stocks (Fig. 6), we used mean values stratified by soil units. This approach therefore cannot account for variability in bulk density and stoniness within single soil units, and hence might mask meaningful variations.

Both the SOC concentration (Fig. 5a–d) and stock (Fig. 6a–d) maps mirror the high importance of topography and soil units for SOC distribution (Section 3.2.3). Clear catenary soil patterns were dominant in each layer, with highest SOC stocks in toeslope and lowest in midslope positions. These patterns were less distinctive within the subsoil, suggesting that the erosive power of surface processes is limited to shallow depths. In contrast to that the impact of soil units was more accentuated in the subsoil.

In each of the observed soil depth intervals, the pale swelling clays of the soil units Zetek, Barbour, Lake (including Swamp), and Gross (Table 2) showed on average higher SOC stocks than the other soil units (Fig. 6a–d). This might relate to subsoil properties such as the high clay content (Table 2) that predominantly consists of expandable smectite minerals, which show a higher SOC stabilization effect than

Table 4
Estimates of tropical soil organic carbon (SOC) stocks

Author	Region	Ecosystem	Soil type	0–5	0–10	0–20	0–30	0–50	0–100	[cm]
						Mg SOC ha^{-1}				
Post et al., 1982	Global	Tropical very dry forest[a]							62	
		Tropical dry forest[a]							102	
		Tropical moist forest[a]							115	
		Tropical wet forest[a]							210	
Batjes, 1996	Global		Ferralsols[f]				57	176	107	
			Cambisols[f]				50	69	96	
			Luvisols[f]				31	43	65	
Amthor and Huston, 1998	Global	Tropical forest							83	
Jabbagy and Jackson, 2000	Global	Tropical deciduous forest[b]							158	
		Tropical evergreen forest[b]							186	
		Tropical grassland/savanna[b]							132	
Bernoux et al., 2002	Brazil	Open Amazon forest[c]	Oxisols[g]				50.8			
		Dense Amazon forest[c]	Oxisols[g]				58.3			
		Amazon savanna[c]	Oxisols[g]				24.6			
		Open Amazon forest[c]	Ultisols[g]				53.9			
		Dense Amazon forest[c]	Ultisols[g]				56.4			
		Amazon savanna[c]	Ultisols[g]				36.2			
Batjes and Dijkshoorn, 1999	Amazon		Cambisols[f]				55.9		95.3	
			Ferralsols[f]				50.5		101.6	
			Luvisols[f]				46.7		88.6	
Cerri et al., 2003	Amazon	Open humid tropical forest with large a number of palms	Ultisols[f]	15.7			34.4			
Rhoades et al., 2000	Ecuador	Old growth lower montane forest	Andic humitropepts[g]				95.6		178.5	
de Koning et al., 2003	Ecuador	Secondary forest (7–30 yr after pasture)	Tropepts, aquents, orthents, fluvents, udalfs, udolls, and psamments[g]					106.1		
Powers and Schlesinger, 2002	Costa Rica	See Powers (2004)	See Powers (2004)		34.1		82.2			
Powers and Veldkamp, 2005	Costa Rica	See Powers (2004)	See Powers (2004)				80.5			
						Mg C ha^{-1}				
Powers, 2004	Costa Rica	Tropical wet forest transitioning to tropical wet–cool transition forest at higher elevations[a]	Tropohumults, dystropepts, dystrandepts[g]		35.2		81.9			
Veldkamp et al., 2003	Costa Rica	Tropical wet forest[a]	Alluvial soils[h]				64			
			Residual soils[h]				96			
Brown and Lugo, 1990	Puerto Rico	Mature forest, wet life zone[a]	Clayey, kaolinitc, isohyperthermic typic tropohumults[g]			0–25 cm: ~100				
	Virgin Islands	Late secondary forest (100 yr), moist life zone[a]	Fine, mixed, isohyperthermic typic argiustolls[g]			0–25 cm: ~80				
	Puerto Rico	Mature forest, dry life zone[a]	Clayey, mixed, isohyperthermic lithic ustorthent[g]			0.25 cm: ~55				
Li et al., 2005	Puerto Rico	Secondary forest (29 yr)	Mixed isothermic tropohumult[g]		34.5	0–25 cm: 56.8				
Bashkin and Binkley, 1998	Hawaii	Wildland forest (never under management)	Typic hydrudands of the Akaka and Kaiwiki series[g]					0–55 cm: 129.8		
Hughes et al., 1999	Mexico	Tall evergreen secondary forest (6 mo to 50 yr)	Well-drained, coarse textured, vitric andosols[g]	12	39	98	138		207	
Brown et al., 1993	Tropical Asia	Tropical forest[d]							148	
Zhong and Zhao, 2001	Tropical and subtropical China	Vegetation categories[e]: coniferous forest, broad-leaf forest, bush and coppice forest, grassland and savannah, meadow and herbaceous swamp, agricultural land				21–94				
Batjes, 2001	Senegal	Forest	Orthic ferrasols[f]				23		47	
			Plinthic ferrasols[f]				35		72	

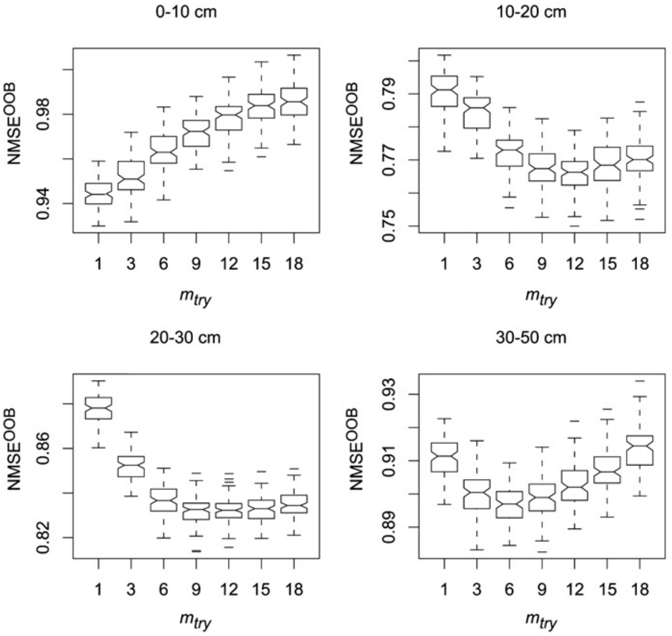

Fig. 3. Iterative determination of best m_{try} values in terms of lowest Out-Of-Bag (the proportion of the dataset which is not used in the bootstrap subset) normalized mean square error (NMSE$^{\text{OOB}}$) for the depth intervals of 0–10, 10–20, 20–30 and 30–50 cm. Each boxplot represents 100 Random Forest runs. See Fig. 2 for details about boxplots.

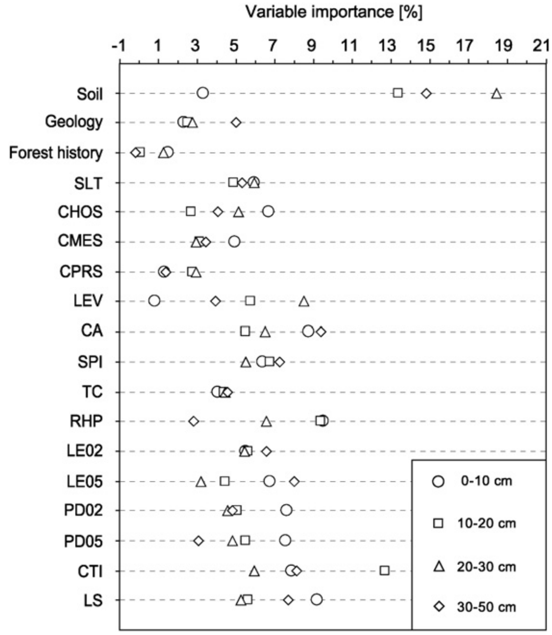

Fig. 4. Variable importance of soil organic carbon predictions averaged over 20 Random Forest runs for each depth interval and normalized to 100% (see Table 2 for terrain parameter abbreviations).

does kaolinite. Furthermore, the pale mottled (heavy) clay lower subsoil with anaerobic soil condition possibly supports shallow rooting trees, and hence leads to an accumulation of SOC near the surface. With increasing depth Ava, Marron, and Harvard contained on average more SOC. The latter soil units are silty clay soils (Table 2) dominated by kaolinite. Except for the silty clay to clay texture of the Lutz soil unit, the remaining soil units, Standley (smectite, kaolinite), Wetmore (smectite, kaolinite), Poacher (kaolinite), and Hood (kaolinite), are somewhat coarser soils with silty clay loam to clay loam textures (Table 2). For the clayey Lutz soil unit, we only had one observation, which in this particular profile is more similar to the Wetmore unit with silty clay loam textured lower subsoil. These results suggest that clay and SOC stock are positively correlated and that clay content is more important than clay mineralogy for stabilizing SOC, as was also observed by Wattel-Koekkoek et al. (2001). Furthermore, the deeply weathered Ava and Harvard are on relatively flat terrain with limited erosion, which supports SOC accumulation. Nonetheless, Marron, situated on the steep sideslopes of the main plateau, contains relatively more SOC. There are two explanations concerning the relatively high SOC stocks of the Marron soil unit: First, Marron is enriched by erosion products originating from the Ava soil unit of the andesite plateau and second, decomposition rates on the sideslopes of the plateau are reduced because of higher soil water contents supplied through subsurface throughflow from the main plateau (Daws et al., 2002).

Considering soil color as integrated in the soil mapping units (Table 2), we could not determine a direct relationship with SOC distribution. This might indicate high contents of hematite, which cover the dark appearance of humic substances.

Geology, an approximation of lithology on BCI, was a relatively weak predictor for SOC prediction in the subsoil (Section 3.2.3). The reason for this could be twofold: Firstly, lithology differs only slightly within the geological formation with mostly andesitic basaltic rock compositions and to a lesser extent foraminiferal limestone. Secondly, geology is incorporated into the spatial extend of soil mapping units, which perhaps more appropriately delineate variations in parent material.

Forest history showed only weak predictive power in the upper soil layers, while below 30 cm it was entirely irrelevant. One could assume that the impact of historical (>100 years ago) land use on the distribution of SOC has faded with forest succession. Brown and Lugo (1990) report that recovery of soil OM takes about 50 years of forest succession.

Finally, we spatially calculated the cumulative SOC stock up to a depth of 30 cm (Fig. 7). In contrast to the traditional approaches where mean SOC stocks were linked to soil map units, we provided a more appropriate estimation of the SOC stock on BCI, accounting for within soil unit variability of SOC stocks. We present spatial SOC estimates up to a depth of 30 cm to facilitate comparison to other studies, as this depth interval has often been used for SOC estimates. The map (Fig. 7) illustrates both, clear catenary SOC patterns as well as the importance of soil texture.

The data ranges of the predicted SOC maps are narrower than those of the pre-processed datasets used for modeling (Section 2.3.; Fig. 2), which is, however, to be expected due to the smoothing effect of the models which tend to predict mean values more often as model accuracy is low. This smoothing effect, however, reduces both the local

Notes to Table 4:

[a] Holdridge life zone classification system.

[b] Biome classification based on Whittaker (1975) and Jackson et al. (1997).

[c] Modified vegetation categories based on the vegetation map of Brazil (IBGE, 1988).

[d] Vegetation map of continental tropical Asia (Food and Agriculture, 1989, K.D. Singh, FAO, pers. comm. 1990); A digital map of the forest areas for insular Asian countries reported by Collins et al., 1991, obtained from the World Conservation Monitoring Centre (WCMC), Cambridge, England.

[e] Modified land use classification based on the Vegetation Map of the People's Republic of China (1:4 M) (Hou, 1982).

[f] FAO World Reference Base for Soil Resources (WRB).

[g] U.S. Soil Taxonomy.

[h] La Selva convention (both groups are Typic Haploperox in U.S. Soil Taxonomy).

Table 5
Model performance from 100 Random Forest runs

		0–10 cm	10–20 cm	20–30 cm	30–50 cm
MSE^{OOB}	Min	2.91	0.30	0.17	0.11
	Med	2.96	0.31	0.17	0.11
	Max	3.00	0.32	0.18	0.11
$RMSE^{OOB}$	Min	1.71	0.55	0.41	0.33
	Med	1.72	0.56	0.41	0.33
	Max	1.73	0.57	0.42	0.33
$NMSE^{OOB}$	Min	0.93	0.75	0.82	0.88
	Med	0.94	0.77	0.83	0.9
	Max	0.96	0.78	0.85	0.91

MSE^{OOB}: Out-Of-Bag mean square error.
$RMSE^{OOB}$: Out-Of-Bag root mean square error.
$NMSE^{OOB}$: Out-Of-Bag normalized mean square error.

variations as well as the effect of random errors, and therefore facilitates the identification of general spatial SOC patterns.

Plant biomass is the main source of OM input to soil. Thus, actual forest composition and structure may be more significant than forest history for making spatial predictions of SOC stocks. However, individual trees possibly drown general forest patterns. Therefore, representations with high spatial resolution such as provided by multi- or hyperspectral remote sensing data are necessary in order to characterize actual forest composition, which might prove powerful for spatial prediction of soil properties such as SOC. Regarding the latter possibility as well as the widespread availability of digital elevation data, digital SOC mapping could be applied to larger areas, helping to refine the resolution of spatial SOC estimates.

4. Conclusion

A large part of the general spatial patterns in SOC variations on Barro Colorado Island in the Panama Canal could be continuously predicted by using the digital soil mapping approach.

In contrast to traditional SOC mapping approaches, where mean SOC concentrations and stocks are spatially linked to soil or vegetation units, the variability of SOC within these units was predicted by integrating empirically derived relationships between SOC and soil forming factors such as topographical (terrain attributes), pedological, lithological, and biological (forest history) attributes into the digital soil mapping framework.

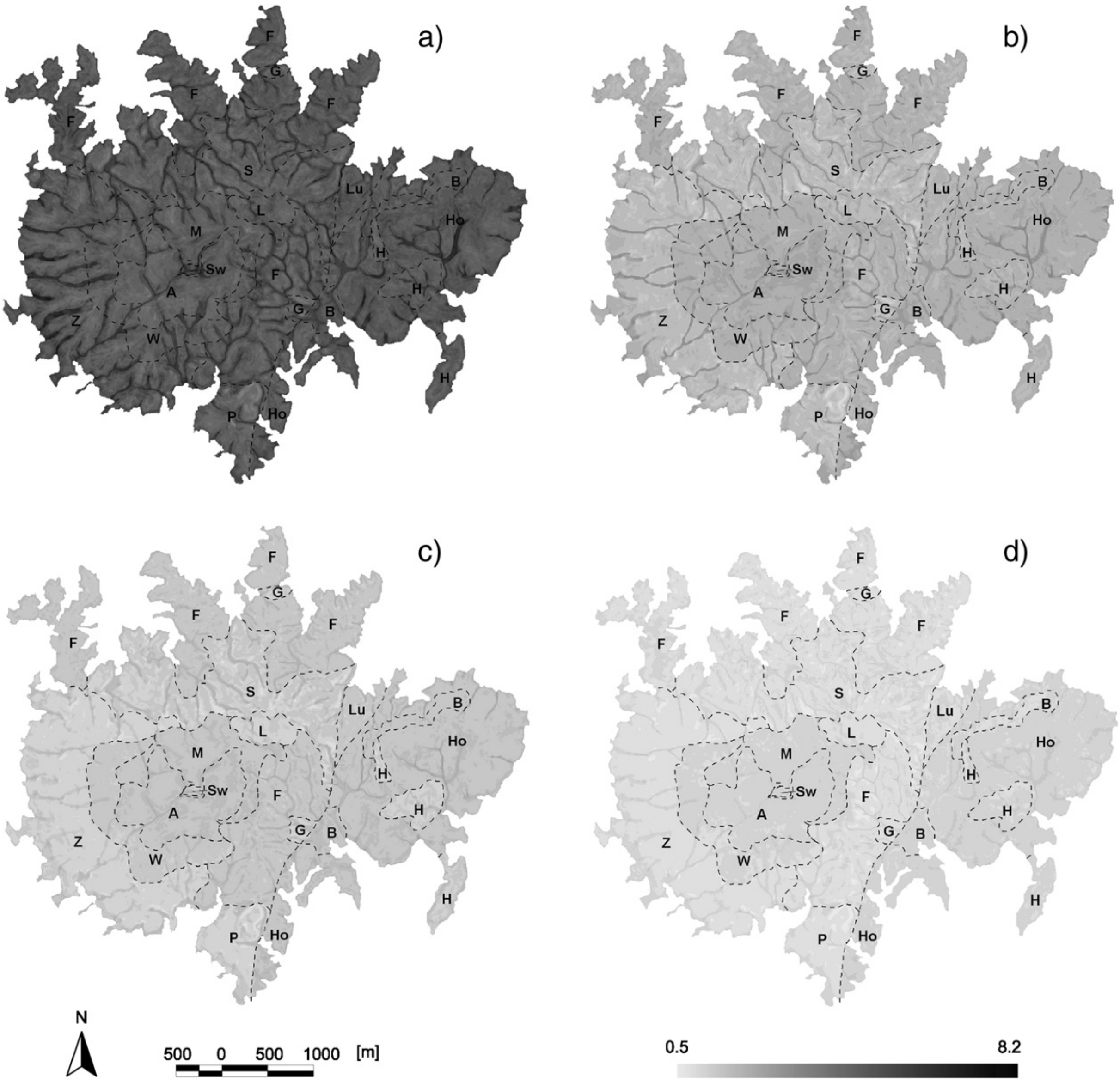

Fig. 5. Soil organic carbon concentrations [%] superimposed on soil units (see Table 1 for soil unit abbreviations) in depth intervals a) 0–10 cm, b) 10–20 cm, c) 20–30 cm, d) 30–50 cm.

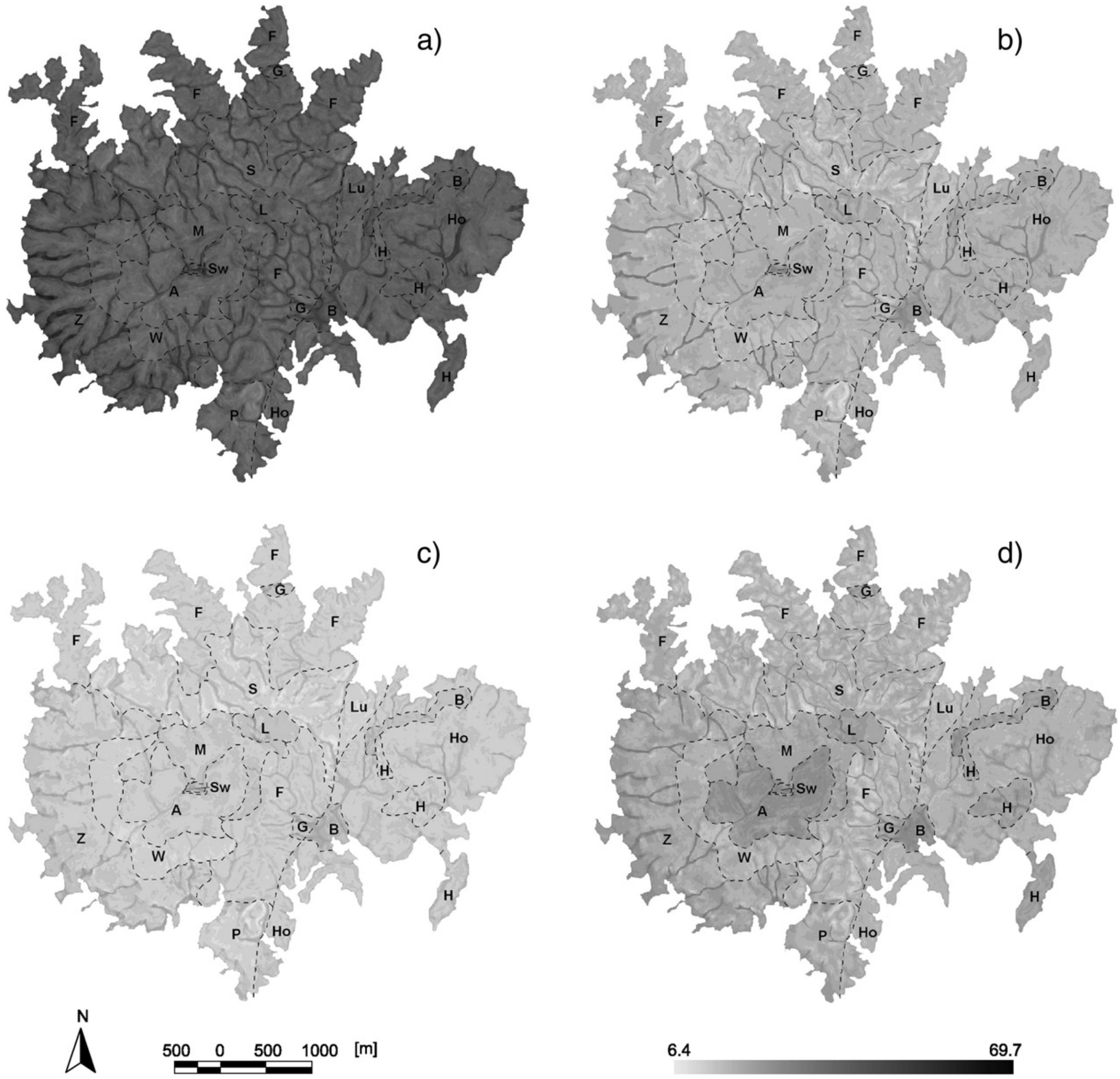

Fig. 6. Soil organic carbon stocks [Mg ha^{-1}] superimposed on soil units (see Table 1 for soil unit abbreviations) in depth intervals a) 0–10 cm, b) 10–20 cm, c) 20–30 cm, d) 30–50 cm.

As a modeling method we applied Random Forest (RF), consisting of an ensemble of CART-like trees, which has proven to be a powerful modeling approach for the spatial prediction of SOC. In order to improve prediction results, the m_{try} parameter settings of the RF algorithm was tested in more detail, revealing that default settings were generally a good choice. Knowledge of soil processes and landscape relationships was drawn from both variable importance measures implemented in RF as well as spatial visualizations of the prediction results. These results indicate that:

- The SOC patterns strongly follow the catena definition of soil properties distribution showing decreasing SOC concentrations and stocks in the sequence of toeslopes > ridges > midslopes.
- In the subsoil, soil units, which represent a generalization of soil and geophysical properties, were most important for SOC concentrations and stocks prediction.
- Neither geology nor forest history were important for SOC concentrations and stocks prediction based on the data available.

In this study we produced a more accurate spatial SOC concentration and stock estimation, which can be used for both understanding the role of tropical soils in the global carbon cycle as well as the incorporation of small scale spatial variations of SOC in future environmental process modeling on BCI.

Acknowledgments

We would like to thank the Smithsonian Tropical Research Institute for the warm welcome and great logistical support and assistance on BCI and at the Tupper campus, Ancon. The constructive comments of Ian Baillie on an earlier draft are gratefully acknowledged. We particularly appreciate the field efforts provided by Frauke Barthold, Julian Schnetzer and Christina Schweipert. We received advice from many scientists in the wider BCI community. Rosina Grimm acknowledges partial financial support by the German Academic Exchange Service (DAAD).

References

Amthor, J.S., Huston, M.I., 1998. Terrestrial ecosystem responses to global change: A research strategy. Oak Ridge National Laboratory. ORNL/TM-1998/27.

Arrouays, D., Vion, I., Kicin, J.L., 1995. Spatial analysis and modeling of topsoil carbon storage in temperate forest humic loamy soils of France. Soil Science 159, 191–198.

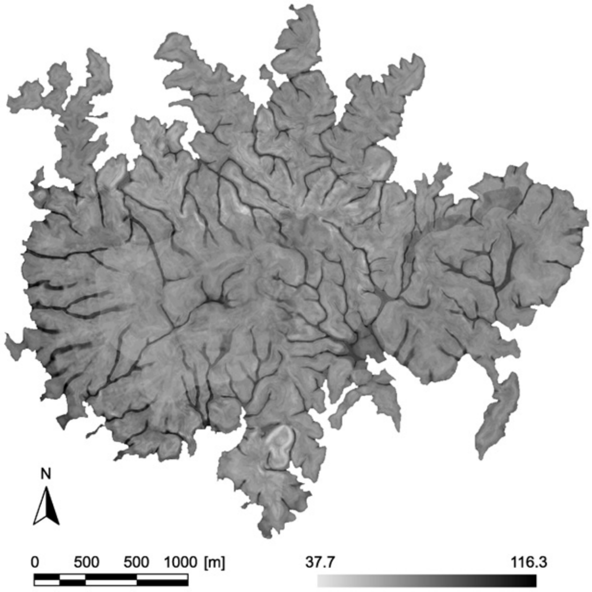

Fig. 7. Soil organic carbon stocks [Mg ha^{-1}] in the upper 30 cm.

Arun, K., Langmead, C.J., 2005. Structure based chemical shift prediction using random forest non-linear regression. Carnegie Mellon University, Pittsburgh. CMU-CS-05-163, on-line: http://reports-archive.adm.cs.cmu.edu/anon/2005/abstracts/05-163.html.

Baillie, I., Elsenbeer, H., Barthold, F., Grimm, R., Stallard, R., 2006. Semi-detailed soil survey of Barro Colorado Island, Panama. on-line: http://biogeodb.stri.si.edu/bioinformatics/bci_soil_map/.

Baldock, J.A., Skjemstadt, J.O., 2000. Role of the soil matrix and minerals in protecting natural organic materials against biological attack. Organic Geochemistry 31, 697–710.

Barthold, F.K., Elsenbeer, H., Stallard, R., 2008. Soil nutrient-landscape relationships in a lowland tropical rainforest in Panama. Forest Ecology and Management 255, 1135–1148.

Bashkin, M.A., Binkley, D., 1998. Changes in soil carbon following afforestation in Hawaii. Ecology 79, 828–833.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science 47, 151–163.

Batjes, N.H., 2001. Options for increasing carbon sequestration in West African soils: An exploratory study with special focus on Senegal. Land Degradation and Development 12, 131–142.

Batjes, N.H., Dijkshoorn, J.A., 1999. Carbon and nitrogen stocks in the soils of the Amazon Region. Geoderma 89, 273–286.

Batjes, N.H., Sombroek, W.G., 1997. Possibilities for carbon sequestration in tropical and subtropical soils. Global Change Biology 3, 161–173.

Behrens, T., 2003. Digitale Reliefanalyse als Basis von Boden-Landschaftsmodellen am Beispiel der Verbreitungssystematik periglaziärer Lagen in deutschen Mittelgebirgen. PhD thesis, Justus Liebig University, Gieβen, Germany.

Behrens, T., Schmidt, K., Scholten, T., 2007. Multi-scale digital terrain analysis and feature selection in digital soil mapping. Biannual Conference of Commission 1.5 Pedometrics, Division 1 of the International Union of Soil Sciences (IUSS). Tübingen, Germany. August 27–30.

Bernoux, M., Carvalho, M.D.S., Volkoff, B., Cerri, C.C., 2002. Brazil's soil carbon stocks. Soil Science Society of America Journal 66, 888–896.

Beven, K., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Bulletin of Hydrologic Science 24, 43–69.

Bhatti, A.U., Mulla, D.J., Frazier, B.E., 1991. Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and thematic mapper images. Remote Sensing of Environment 37, 181–191.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Cutler, A., 2004. Random Forest — manual. on-line: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_manual.htm.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth International, Belmont (CA).

Brown, S., Lugo, A.E., 1990. Effects of forest clearing and succession on the carbon and nitrogen-content of soils in Puerto-Rico and Us Virgin Islands. Plant and Soil 124, 53–64.

Brown, S., Iverson, L.R., Prasad, A., Liu, D., 1993. Geographical distribution of carbon in biomass and soils of tropical Asian forests. Geocarto International 4, 45–59.

Bureau, A., Dupuis, J., Hayward, B., Falls, K., Van Eerdewegh, P., 2003. Mapping complex traits using random forests. Bmc Genetics 4. doi:10.1186/1471-2156-4-S1-S64.

Cerri, C.E.P., Coleman, K., Jenkinson, D.S., Bernoux, M., Victoria, R., Cerri, C.C., 2003. Modeling soil carbon from forest and pasture ecosystems of Amazon, Brazil. Soil Science Society of America Journal 67, 1879–1887.

Chaplot, V., Bernoux, M., Walter, C., Curmi, P., Herpin, U., 2001. Soil carbon storage prediction in temperate hydromorphic soils using a morphologic index and digital elevation model. Soil Science 166, 48–60.

Collins, M., Sayer, J.A., Whitmore, T.C., 1991. The Conservation Atlas of Tropical Forests: Asia and the Pacific. International Union of Conservation and Nature. Simon and Shuster, New York.

Daws, M.I., Mullins, C.E., Burslem, D.F.R.P., Paton, S.R., Dalling, J.W., 2002. Topographic position affects the water regime in a semideciduous tropical forest in Panamá. Plant and Soil 238, 79–90.

de Koning, G.H.J., Veldkamp, E., Lopez-Ulloa, M., 2003. Quantification of carbon sequestration in soils following pasture to forest conversion in northwestern Ecuador. Global Biogeochemical Cycles 17. doi:10.1029/2003 GB002099.

Diaz-Uriate, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. Bmc Bioinformatics 7. doi:10.1186/1471-2105-7-3.

Dietrich, W.E., Montgomery, D.R., 1998. A digital terrain model for mapping shallow landslide potential. . on-line: http//socates.berkeley.edu/~geomorph/shalstab.

Dietrich, W.E., Windsor, D.M., Dunne, T., 1982. Geology, climate, and hydrology of Barro Colorado Island. In: Leigh, E.G., Rand, A.S., Windsor, D.M. (Eds.), The Ecology of a Tropical Forest: Seasonal Rhythms and Long-term Changes. Smithsonian Institute Press, Washington D.C., pp. 21–46.

FAO (Food and Agriculture), 1989. Classification and mapping of vegetation types in tropical Asia. Tropical Forest Resources Assessment 1990 Program. Food and Agriculture Organization of the United Nations, Rome.

FAO (Food and Agriculture), 2006. Guidelines for Soil Description. Food and Agriculture Organization of the United Nations, Rome.

Feldwisch, N., 1995. Hangneigung und Bodenerosion. Boden und Landschaft - Schriftenreihe zur Bodenkunde Landeskultur und Landschaftsökologie der Justus-Liebig-Universität Gieβen, vol. 3, p. 152.

Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling and Software 17, 295–311.

Foster, R.B., Brokaw, N.V.L., 1996. Structure and history of the vegetation of Barro Colorado Island. In: Leigh, E.G., Rand, S., Windsor, D.M. (Eds.), The Ecology of a Tropical Forest: Seasonal Rhythms and Long-term Changes. Smithsonian Institution, Washington, DC, pp. 67–82.

Freund, Y., Schapire, R.E., 1996. Game theory, on-line prediction and boosting. Proceedings of the 9th Annual Conference on Computing and Learning Theory, pp. 325–332.

Furlanello, C., Neteler, M., Merler, S., Menegon, S., Fontanari, S., Donini, A., Rizzoli, A., Chemini, C., 2003. GIS and the random forest predictor: Integration in R for tick-borne diseases risk assessment. In: Hornik, K., Leitsch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna, Austria, pp. 1–11.

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modeling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9, 421–432.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27, 294–300.

Gunther, E.C., Stone, D.J., Gerwien, R.W., Bento, P., Melvyn, P.H., 2003. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. Proceedings of the National Academy of Sciences of the United States of America 100, 9608–9613.

Ham, J., Chen, Y.C., Crawford, M.M., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43, 492–501.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. . Springer Series in Statistics. Springer Verlag, New York.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93.

Hou, X., 1982. Vegetation Map of the People's Republic of China (1:4 M). Map Press, Beijing.

Hughes, R.F., Kauffman, J.B., Jaramillo, V.J., 1999. Biomass, carbon, and nutrient dynamics of secondary forests in a humid tropical region in Mexico. Ecology 80, 1892–1907.

IBGE 1988. Mapa de vegetação do Brasil, escala 1:5,000,000. Fundação Instituto Brasileiro de Geografia e Estatstica, Rio de Janeiro.

Jackson, R.B., Mooney, H.A., Schulze, E.D., 1997. A global budget for fine root biomass, surface area, and nutrient contents. Proceedings of the National Academy of Science (USA) 94, 7362–7366.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

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

Johnsson, M.J., Stallard, R.F., 1989. Physiographic controls on the composition of sediments derived from volcanic and sedimentary terrains on Barro Colorado Island, Panama. Journal of Sedimentary Petrology 59, 768–781.

Kahle, M., Kleber, M., Torn, M.S., Jahn, R., 2002. Carbon storage in coarse and fine clay fractions of illitic soils. Soil Science Society of America Journal 67, 1732–1739.

Kinner, D., Mixon, D., Stallard, R., Wahl, S., 2002. BCI GIS version 1.3. Smithsonian Tropical Research Institute, CD-Rom.

Konen, M.E., Burras, C.L., Sandor, J.A., 2003. Organic carbon, texture, and quantitative color measurement relationships for cultivated soils in north central Iowa. Soil Science Society of America Journal 67, 1823–1830.

Kulmatiski, A., Vogt, D.J., Siccama, T.G., Tilley, J.P., Kolesinskas, K., Wickwire, T.W., Larson, B.C., 2004. Landscape determinants of soil carbon and nitrogen storage in southern New England. Soil Science Society of America Journal 68, 2014–2022.

Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sensing of Environment 100, 356–362.

Leigh, E.G., 1999. Tropical Forest Ecology: A View From Barro Colorado Island. Oxford University Press, New York, USA.

Li, Y., Xu, M., Zou, X.M., Shi, P.J., Zhang, Y.Q., 2005. Comparing soil organic carbon dynamics in plantation and secondary forest in wet tropics in Puerto Rico. Global Change Biology 11, 239–248.

Liaw, A., Wiener, M., 2002. Classification and regression by random forests. R News 2/3, 18–22.

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

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

Minasny, B., McBratney, A.B., Mendonca-Santos, M.L., Odeh, I.O.A., Guyon, B., 2006. Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. Australian Journal of Soil Research 44, 233–244.

Ministerio de Comercio e Industrias, 1976. Mapa Geológico de Panama: 1:250,000, 7 sheets.

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphographical and biological applications. Hydrological Processes 5, 3–30.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443–452.

Nogami, M., 1995. Geomorphometric measures for digital elevation models. Zeitschrift für Geomorphologie Supplement 101, 53–67.

Odeh, I.O., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215–226.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26, 217–222.

Parkhurst, D.F., Brenner, K.P., Dufour, A.P., Wymer, L.J., 2005. Indicator bacteria at five swimming beaches — analysis using random forests. Water Research 39, 1354–1360.

Paul, K.I., Polglase, P.J., Nyakuengama, J.G., Khanna, P.K., 2002. Change in soil carbon following afforestation. Forest Ecology and Management 168, 241–257.

Post, W.M., Emanuel, W.R., Zinke, P.J., Stangenberger, A.G., 1982. Soil carbon pools and world life zones. Nature 298, 156–159.

Powers, J.S., 2004. Changes in soil carbon and nitrogen after contrasting land-use transitions in northeastern Costa Rica. Ecosystems 7, 134–146.

Powers, J.S., Schlesinger, W.H., 2002. Relationships among soil carbon distributions and biophysical factors at nested spatial scales in rain forests of northeastern Costa Rica. Geoderma 109, 165–190.

Powers, J.S., Veldkamp, E., 2005. Regional variation in soil carbon and delta C-13 in forests and pastures of northeastern Costa Rica. Biogeochemistry 72, 315–336.

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.

R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rhoades, C.C., Eckert, G.E., Coleman, D.C., 2000. Soil carbon differences among forest, agriculture, and secondary vegetation in lower montane Ecuador. Ecological Applications 10, 497–505.

Schmidt, J., Dikau, R., 1999. Extracting geomorphographic attributes and objects from digital elevation models — semantics, methods, future needs. In: Dikau, R., Sauer, H. (Eds.), GIS for Earth Surface Systems. . Analysis and Modelling of the Nature Environment. Gebrüder Bornträger, Stuttgart.

Schwender, H., Zucknick, M., Ickstadt, K., Bolt, H.M., 2004. A pilot study on the application of statistical classification procedures to molecular epidemiological data. Toxicology Letters 151, 291–299.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27, 171–197.

Shary, P.A., Sharaya, L.S., Mitusov, A.V., 2002. Fundamental quantitative methods of land surface analysis. Geoderma 107, 1–35.

Silver, W.L., Ostertag, R., Lugo, A.E., 2000. The potential for carbon sequestration through reforestation of abandoned tropical agricultural and pasture lands. Restoration Ecology 8, 394–407.

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J.L., Haddix, M.L., 2006. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. Geoderma 132, 471–489.

Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilisation mechanisms of soil organic matter: Implications for C-saturation of soils. Plant and Soil 241, 155–176.

Soil Survey Staff, 1996. Soil survey laboratory methods manual. Soil Survey Investigations Report No. 42, V. 3.0, National Soil Survey Center, Natural Resources Conservation Service. U.S. Department of Agriculture, Lincoln, NE.

Sutton, C., 2005. Classification and regression trees, bagging, and boosting. Handbook of Statistics 24, 303–329.

Svenning, J.C., Kinner, D.A., Stallard, R.F., Engelbrecht, B.M.J., Wright, S.J., 2004. Ecological determinism in plant community structure across a tropical forest landscape. Ecology 85, 2526–2538.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences 43, 1947–1958.

Tarboton, D.G., 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research 33, 309–319.

Taylor, J.R., 1997. An Introduction to Error Analysis — the Study of Uncertainties in Physical Measurements, 2 ed. University Science Books, Sausalito, CA., USA.

Thompson, J.A., Kolka, R.K., 2005. Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Science Society of America Journal 69, 1086–1093.

Thompson, J.A., Pena-Yewtukhiw, E.M., Grove, J.H., 2006. Soil-landscape modeling across a physiographic region: topographic patterns and model transportability. Geoderma 133, 57–70.

Torn, M.S., Trumbore, S.E., Chadwick, O.A., Vitousek, P.M., Hendricks, D.M., 1997. Mineral control of soil organic carbon storage and turnover. Nature 389, 170–173.

USDA (U.S. Department of Agriculture) Soil Conservation Service, 1973. Soil Survey of the Island of Hawaii, State of Hawaii. U.S. Government Printing Office, Washington, D.C., USA.

Van Breemen, N., Feijtel, T.C.J., 1990. Soil processes and properties involved in the prediction of greenhouse gases, with special relevance to soil taxonomic systems. In: Bouwman, A.F. (Ed.), Soils and the Greenhouse Effect. Wiley, Chichester, pp. 195–223.

Veldkamp, E., Becker, A., Schwendenmann, L., Clark, D.A., Schulte-Bisping, H., 2003. Substantial labile carbon stocks and microbial activity in deeply weathered soils below a tropical wet forest. Global Change Biology 9, 1171–1184.

Viscarra-Rossel, R.A., Minasny, B., Roudier, P., McBratney, A.B., 2006. Colour space models for soil science. Geoderma, 133, 320–337.

Wattel-Koekkoek, E.J.W., van Genuchten, P.P.L., Buurman, P., van Lagen, B., 2001. Amount and composition of clay-associated soil organic matter in a range of kaolinitic and smectitic soils. Geoderma 99, 27–49.

Whittaker, R.H., 1975. Communities and Ecosystems. Macmillan, London, UK.

Woodring, W.P., 1958. Geology of Barro Colorado Island, canal zone. Smithsonian Institution Miscellaneous Collections, Washington, D.C., Smithsonian Institution. vol. 135, No.3.

WRB (World reference base for soil classification), 2006. A Framework for International Classification, Correlation and Communication. . World Soil Resources Reports, vol. 103. Food and Agriculture Organization of the United Nations, Rome.

Yavitt, J.B., 2000. Nutrient dynamics of soil derived from different parent material on Barro Colorado Island, Panama. Biotropica 32, 198–207.

Yavitt, J.B., Wright, S.J., 2002. Charge characteristics of soil in a lowland tropical moist forest in Panama in response to dry-season irrigation. Australian Journal of Soil Research 40, 269–281.

Yavitt, J.B., Wieder, R.K., Wright, S.J., 1993. Soil nutrient dynamics in response to irrigation of a Panamanian tropical moist forest. Biogeochemistry 19, 1–25.

Zhong, L., Zhao, Q.G., 2001. Organic carbon content and distribution in soils under different land uses in tropical and subtropical China. Plant and Soil 231, 175–185.