Use of Digital Terrain Analysis and Classification Trees for Predictive Mapping of Soil Organic Carbon in Southern Denmark

R. Bou Kheir[1], M. H. Greve[1], P. K. Bocher[1], M. B. Greve[1]

[1] Department of Agroecology and Environment, Faculty of Agricultural Sciences (DJF), Aarhus University, Blichers Allé 20, P.O. Box 50, DK-8830 Tjele, Denmark
Email: Rania.BouKheir@agrsci.dk

1. Introduction

Soil organic carbon (SOC) is a dynamic component of the terrestrial system, with both internal changes in the vertical and horizontal directions and external changes with the atmosphere and the biosphere. Changes in SOC are attributed to both natural processes and human activities, and reflect the balance between decomposition of organic matter and input from roots and litter (Turner and Lambert 2000). In recent years, the importance of human activities has been widely recognized. Land use changes, including deforestation, biomass burning, draining of wetlands, ploughing, use of fertilisers and other agricultural practices, are regarded as the main factors causing loss of SOC and the emission of CO_2 into the atmosphere. These changes can be significant in grassland and cropland (Conant and Paustian 2002, Schuman et al. 2002) where intensive agricultural activities are carried out.

As part of international efforts to stabilize atmospheric greenhouse gas concentrations, Denmark is committed to establish inventories of the C stock in the frame of the Kyoto protocol. In this context, our study focuses on building a simple, realistic, practical and informative classification-tree model to predict the distribution of spatial patterns and changes in SOC across a study area in southern Denmark from mapped environmental variables.

2. Material and Methods

2.1 Site Description

The chosen study area, covering about 1812 km^2, is located in southern Denmark (Fig. 1). The climate is temperate with mean annual temperature ranging from 0 to 16°C, and a West-East gradient in precipitation oscillating between 900 and 600 mm/year (1961-1990). 95% of parent materials have glacial and fluvio-glacial origin. Approximately 65% of these materials were deposited during the last glacial period (between 10,000 and 100,000 years), and 20% during the previous glacial period (more than 110,000 years ago). However, the deposits from that period were all strongly redistributed by periglacial processes, and evidence of earlier soil formations is extremely rare. The area is representative of a broad region of landscapes in Denmark (i.e. weichsel moraine landscape, glacifluvial plains, saalian landscape, aeolian landscape, and post glacial marine deposits). The elevation varies from 0 m in the western part to 85 m in the eastern part. The area has been intensively cropped since the Middle Ages. Currently, 70% of the area is cultivated, 10% forested and the rest urbanized.

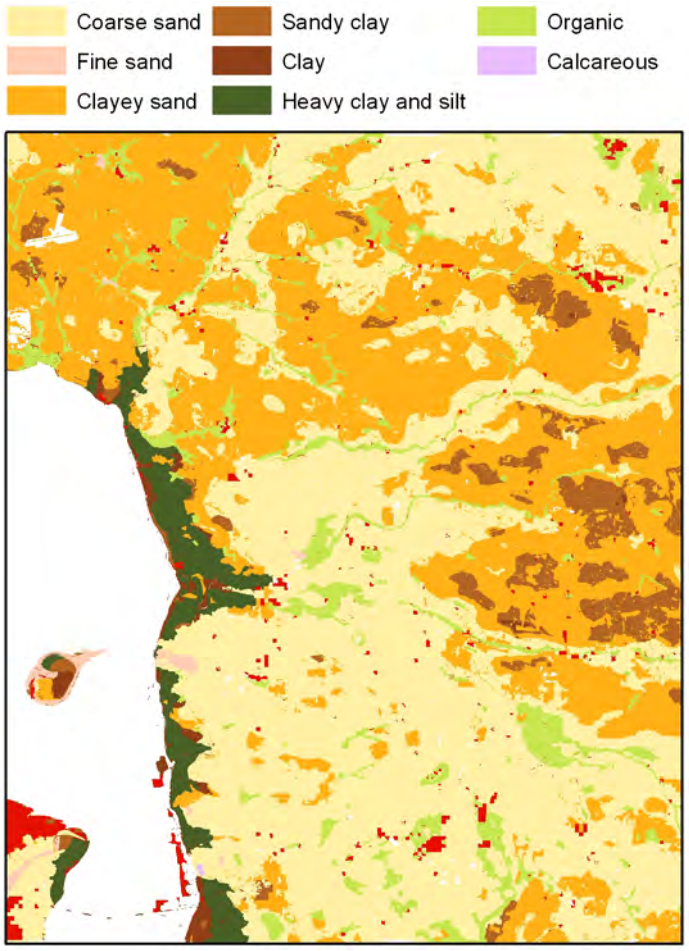

Figure 1. Soil map of the study area within Denmark (Madsen et al., 1992)

2.2 Soil Samples Collection

The soil was sampled at 1541 sites selected by four different surveys to be representative for the area. In order to avoid soil variability on a small scale, 25 bulk soil samples were taken within a radius of 50 m from a depth of 10-20 cm in the Danish Soil Classification (1975) and the Danish Profile Investigation (1990). The collected samples in these two surveys were taken to the laboratory for analysis. These samples were air-dried at room temperature and passed through a 2 mm soil sieve. Concentrations of SOC were determined by the combustion method in a LECO induction furnace, converted to % Soil Organic Matter (SOM). The other two surveys (ochre classification and well database performed in 1985) gave categorical information on parent material (e.g. peat, sand, silt and clay). This parent material information was reclassified into organic and mineral soils. In order to increase the number of samples used in the modeling process, the continuous soil organic matter (SOM) obtained in the former surveys was converted to a categorical variable using 10% SOM as a cut off value. With less than 10% SOM, soils are classified as mineral; and with more than 10% SOM, soils are considered organic.

2.3 Data Acquisition and Pre-Processing

Mapping organic soils can be achieved using decision-tree modelling through incorporating secondary spatial information into prediction (Mueller and Pierce 2003). Available digital geology (with five classes, Jakobsen and Hermansen 2007), soil (with eight classes, Madsen et al. 1992) and landscape (with five geomorphological units,

Smed 1978) maps of the study area at different scales (1:20.000, 1:50.000) were converted to 30 m square grids.

A 30 m digital terrain model (DEM) was developed for the study area. The DEM resolution was chosen to match the spatial resolution of the used remote sensing data. The digital elevation data was acquired using airborn LiDaR, and a DEM of 2-meter resolution in full national scale was produced. This high resolution model was resampled to 30-meter resolution for the purpose of this study. Topographic attributes may aid spatial estimation of soil carbon, because the relief has a great influence on soil formation (McKenzie and Ryan 1999, Bou Kheir et al. 2007, 2008). They may be divided into primary and secondary attributes. Primary terrain attributes can be directly extracted from the DEM. Secondary parameters are calculated from two or more primary attributes. In this work, eight primary attributes (elevation; slope; aspect; plan, profile and tangential curvature; flow accumulation; rate of change of specific catchment area) were calculated with ArcGIS® and TerraSTREAM (Danner et al. 2007). The secondary derived attribute, quasi-dynamic topographic wetness index, was produced using the formula of Barling et al. (1994).

Besides digital elevation models, one of the most interesting sources of secondary information could be remote sensing RS, if a relationship between soil properties and spectral data could be achieved. Remotely sensed data can be useful for improving existing coarse-scale soil survey information at a regional scale (Scull et al. 2005). However, high carbon soils in Denmark can not directly be differentiated from moist soils using satellite imageries, both appearing as dark soils decreasing the spectral reflectance whenever their content increases. For that, several RS indices like the Normalized Difference Vegetation Index NDVI (based on red and NIR bands) and the Normalized Difference Wetness Index NDWI (based on near infrared and short near infrared channels) (Lillesand and Kiefer 1994) were derived from Landsat TM imageries (30 m) acquired in April 1987. NDVI was calculated to integrate vegetation status, while NDWI was extracted to compute surface moisture. The used Landsat TM imageries were chosen for many reasons: (1) being the most recent in the archive, (2) having good radiometric quality, and (3) at a minimum of plant cover shading the bare soils.

2.4 Statistical Analysis

The field survey data were split into two files, one compiling 80% of the field samples (1233 sites) used in the modelling process, and another one comprising 20% used in the validation phase (308 sites). The modelling file integrates x and y fields representing locational coordinates and the z field representing SOC. This file was converted to a square grid that matched the resolution of the constructed DEM. ArcGIS was used to assign topographic, soil, geology, landscape, NDVI and NDWI variables to each of the field survey (sampling) locations.

Spatial prediction of SOC was produced using tree-based classification models. These models are easy to interpret and discuss when a mix of continuous and categorical variables is used as predictors (Gessler 1996, McKenzie and Ryan 1999). They comprise a set of rules facilitating the classification of a categorical (classification tree) or continuous (regression tree) dependent variable based on values of the independent variables. In predictive SOC mapping, the dependent variable (SOC presence/absence) is categorical and the independent variables are both continuous (elevation; aspect; slope; plan, profile and tangential curvature; flow accumulation; rate of change of specific catchment area along the direction of flow; quasi-dynamic topographic wetness index; NDVI; NDWI) and categorical or nominal (soil type;

geological substrate; landscape type). However, the most significant advantage of tree-based models is the capacity to model non-additive and non-linear relationships in a relatively simple way (Scull et al. 2005). This is particularly useful for soil data where interactions between the response variable and environmental explanatory variables are often conditional on other explanatory variables.

Three sets of tree-models were explored based on (1) all of the variables, (2) the primary topographic variables only and (3) selected pairs of variables. Once the tree has been developed, it encodes a set of decision rules that define the range of conditions (values of environmental variables) best used to predict each SOC class.

Pruning the tree is necessary to prevent the model from being over-fit to the sample data, and to reduce tree complexity. Pruning entails combining pairs of terminal nodes into singles nodes to determine how the misclassification error rate changes as a function of tree size. We used cost-complexity pruning with an independent data set (a pruning data set) to produce a plot of training misclassification error rate versus tree size (Safavian and Norvig 1991).

2.5 Construction of SOC Map

Using the preferred classification-tree model (having the highest predictive power, and the lowest number of nodes), a predictive map of SOC was obtained under a GIS environment. This map was validated based on field surveys. An independent dataset has been chosen randomly in all landscape units, consisting of 20% (308 sites) of the total number of field sites, and the total accuracy was calculated.

3. Results and Discussion

Training misclassification error rates for the explanatory trees that were developed using all variables (Model 1) at a time or the primary topographic variables only (Model 2) varied from 23% to 26%, with quasi-identical numbers of terminal nodes (71 nodes for Model 1 and 69 nodes for Model 2). The relative importance of the predictor variables (Gini splitting method) in building those trees and splitting the corresponding nodes is shown in Table 1.

Applying cost-complexity pruning indicated that model 1 (based on all variables) would classify correctly 60% of the tested SOC selecting just seven terrain variables (with their relative importance shown in parentheses): landscape type (100%), soil type (29%), elevation (22.5%), tangent curvature (14%), NDVI (12%), aspect (11%), and slope (9%). Model 2 (based on topographic variables only) did slightly better and classified 63% of the text data accurately using five variables: (1) elevation (100%), (2) slope (36%), (3) aspect (16%), (4) tangent curvature (8%), and (5) profile curvature (5%). The number of the terminal nodes was very similar for both pruned models.

The models based on pairs of variables explained 50-68% of the variation in Soil Organic Carbon (Table 2). The model based on soil type and quasi-dynamic wetness index (TWI) (model 3) showed the highest predictive power, classifying 68% of the data correctly and pruned to fourteen terminal nodes. The TWI is a predictor of zones of soil saturation, and SOC often accumulates in lowland (concave) soils for two reasons: (1) on steep slopes, dry soil conditions prevail due to more rapid removal of water causing an important decrease in SOC, and (2) concave slopes can concentrate more water and sediments indicating the potential accumulation of a large quantity of soil organic carbon.

Predictor variables (%)	Model 1 (explanatory tree)	Model 1 (pruned tree)	Model 2 (explanatory tree)	Model 2 (pruned tree)
Elevation	70%	22.5%	100%	100%
Aspect	50%	11%	54%	16%
Slope	37%	9%	47%	36%
Profile curvature	25%	0%	23%	5%
Tangent curvature	34%	14%	12%	8%
Plan curvature	23%	0%	25%	0%
Flow accumulation	0%	0%	4%	0%
Specific catchment area	0%	0%	0%	0%
Quasi-dynamic wetness index	37%	0%	Not included in building the model	
Geological substrate	31%	0%		
Soil type	39%	29%		
Landscape type	100%	100%		
NDVI	31%	12%		
NDWI	18%	0%		
Tree size – Terminal nodes	71	9	69	10
Error rate – Training (%)	23%	33%	26%	33%
Error rate – Test (%)	-	40%	-	37%

Table 1. Relative importance of predictor variables (%) and misclassification error rates in models 1 (based on all variables) and 2 (based on primary topographic variables only).

Predictor variables	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a	X	62	64	61	61	61	61	61	62	62	63	61	61	61
b		X	58	56	53	56	51	50	54	60	60	60	50	52
c			X	57	58	59	60	60	59	53	62	62	58	58
d				X	54	53	54	54	54	62	60	61	56	55
e					X	55	55	55	56	60	60	62	53	55
f						X	56	56	58	60	61	60	54	54
g							X	51	56	60	60	60	61	56
h								X	56	60	61	60	53	56
i									X	60	68	60	53	54
j										X	62	62	61	61
k											X	62	62	62
l												X	60	60
m													X	53
n														X

a = elevation, b = aspect, c = slope, d = profile curvature, e = tangent curvature, f = plan curvature, g = flow accumulation, h = specific catchment area, i = quasi-dynamic wetness index, j = geological substrate, k = soil type, l = landscape type, m = NDVI, n = NDWI

Table 2. Proportion of accuracy explained (%) for pruned tree-classification models based on pairs of variables (model 3).

Without pruning, this model gave similar results to models 1 and 2 (75% of accuracy explained), but model 3 is preferred because it is easier to understand and faster to use for making predictions. In addition, pruning the trees to their optimal size is a required task because smaller trees may provide greater predictive accuracy for unseen data than large trees. In both models 1 and 3, the predictor variable that was used statistically to generate the split from the parent node was the soil type, indicating its potential role in predicting the geographic location of SOC. The recommended model (model 3) relies on a small number of rules and just two independent predictor

variables, one of which can be easily and quickly constructed whenever a DEM is available, which is the case in most countries (Fig. 2). The produced predictive map of soil organic carbon (Fig. 3) at 1:50,000 cartographic scale indicates that 7.5% of the wetlands in the study area correspond to organic soils, and 92.5% to mineral soils. The confusion matrix between the measured SOC classes and the modelled ones indicates a good overall accuracy of ca. 75%.

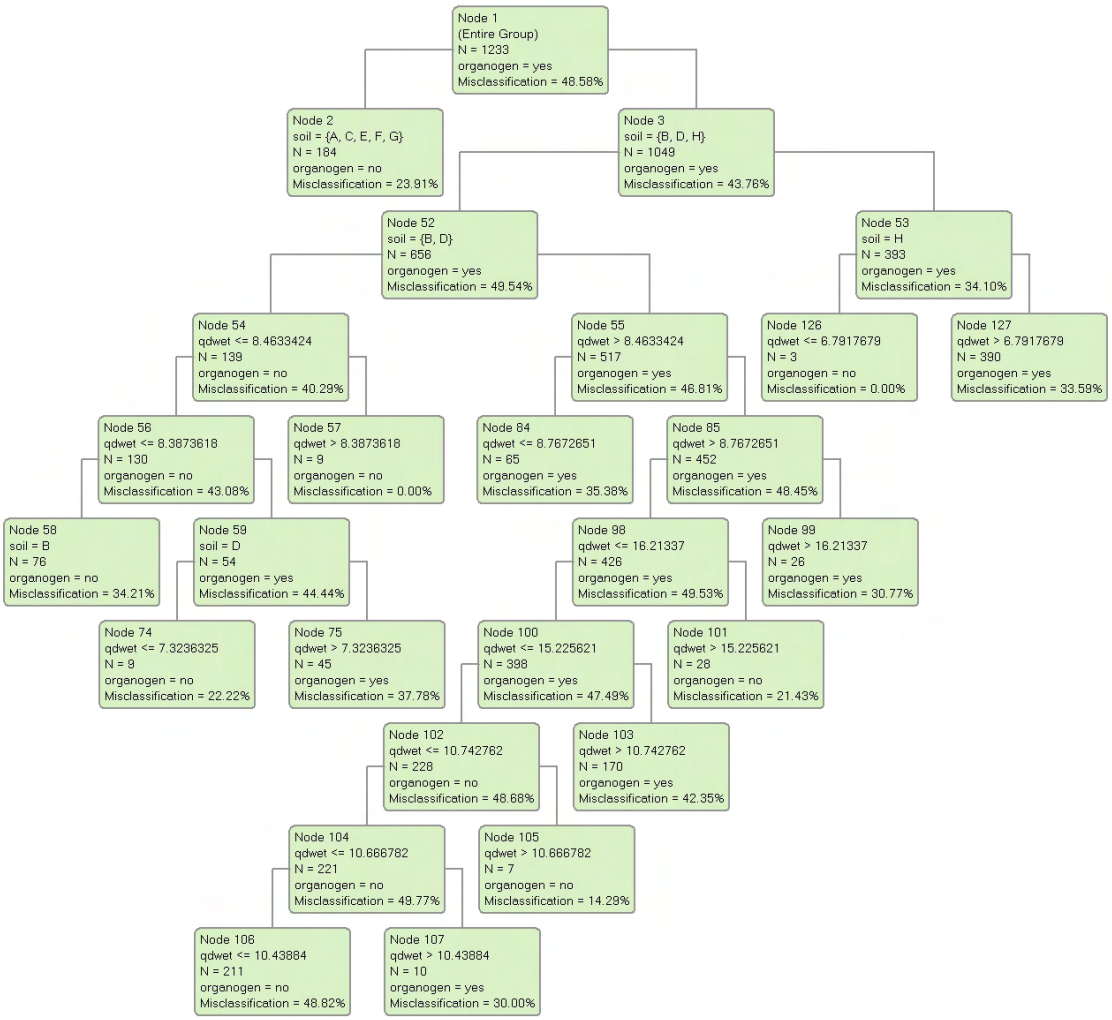

Figure 2. Classification-tree model based on the combination of soil type and quasi-dynamic wetness index for predicting the spatial distribution of SOC.

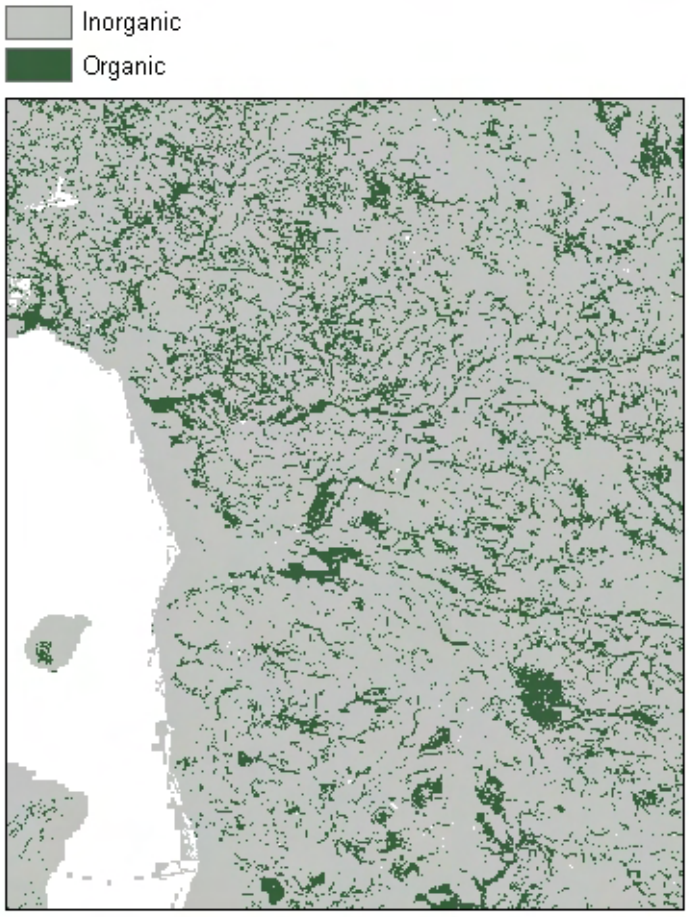

Figure 3. Soil organic carbon predicted using a classification tree model based on the combination of soil type and quasi-dynamic wetness index.

4. Conclusion

Topographic variables derived from DEMs are related to the geographic distribution of SOC. The preferred tree-based models explained 74-77% of the SOC distribution for a series of chosen field sites in southern Denmark. Two terrain variables – soil type and quasi-dynamic topographic wetness index – proved to be the most important variables, indicating that complex or secondary topographic variables show stronger relationships to SOC than primary topographic attributes. This particular pair of secondary topographic variables incorporated the effects of slope and upslope contributing area.

This modelling approach was easily implemented with available GIS (ArcGIS) and statistical (DTREG) software and is suitable for data exploration and predictive SOC mapping. It is explicit and can be critically evaluated and revised when necessary. It has the capacity to integrate easily other primary topographic attributes (e.g. slope length). The inclusion of additional variables might have explained some of the additional variation in the geographic distribution of SOC.

Future work will first compare the results from this study with those from other models (e.g. fuzzy logic, artificial neural networks, etc.), and later seek to gather additional field data so we can examine whether or not finer-scale DEMs can predict the distribution and magnitude of SOC with greater precision and reliability.

References

Barling RD, Moore ID and Grayson RB, 1994, A quasi-dynamic wetness index for characterizing the spatial distribution of zones of surface saturation and soil water content. *Water Resources Research*, 30: 1029-1044.

Bou Kheir R, Wilson J and Deng Y, 2007, Use of terrain variables for mapping gully erosion susceptibility in Lebanon. *Earth Surface Processes and Landforms*, 32: 1770-1782.

Bou Kheir R, Chorowicz J, Abdallah C and Damien D, 2008, Soil and bedrock distribution estimated from gully form and frequency: a GIS-based decision-tree model for Lebanon. *Geomorphology*, 93: 482-492.

Conant RT and Paustian K, 2002, Spatial variability of soil organic carbon in grasslands: implications for detecting changes at different scales. *Environmental Pollution*, 116: S127-S135.

Danner A, Yi K, Mølhave Th, Agarwal PK, Arge L and Mitasova H, 2007, TerraStream: From Elevation Data to Watershed Hierarchies. Proc. 15th International Symposium on Advances in Geographic Information Systems (ACM GIS), 2007.

Gessler PE, 1996, *Statistical Soil-Landscape Modeling for Environmental Management*, Ph.D. thesis, Australian National University, Canberra.

Jakobsen PR and Hermansen B, 2007, Danmarks Digitale Jordartskort. Version 3.0. Danmarks og Grønlands Geologiske Undersøgelse Rapport 2007/84, [Kun på CD-Rom].

Lillesand TM and Kiefer RW, 1994, *Remote sensing and image interpretation*. 3rd ed., John Wiley & Sons Inc., 750 p

Madsen HB, Nørr AH and Holst KA, 1992, The Danish Soil Classification. Altas of Danmark I(3). Reitzel, Copenhagen.

McKenzie NJ and Ryan PJ, 1999, Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89: 67-94.

Mueller TG and Pierce FJ, 2003, Soil carbon maps – Enhancing spatial estimates with simple terrain attributes at multiple scales. *Soil Sci. Soc. Am. J.*, 67, 258-267.

Smed P, 1978, Landskabskort over Danmark. Geografforlaget, 5464 Brendrup.

Safavian SJ and Norvig P, 1991, A survey of tree classifier methodology. *IEEE transactions Syst. Man Cybern*, 21: 660-674.

Schuman GE, Janzen HH and Herrick JE, 2002, Soil carbon dynamics and potential sequestration by rangelands. *Environmental Pollution*, 116: 391-396.

Scull P, Franklin J and Chadwick OA, 2005, The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, 181: 1-15.

Turner J and Lambert M, 2000, Change in organic carbon in forest plantation soils in eastern Australia. *Forest Ecology and Management*, 133, 231-247.