Hydrol. Earth Syst. Sci., 14, 847–857, 2010
www.hydrol-earth-syst-sci.net/14/847/2010/
doi:10.5194/hess-14-847-2010

Hydrology and Earth System Sciences

The application of GIS based decision-tree models for generating the spatial distribution of hydromorphic organic landscapes in relation to digital terrain data

R. Bou Kheir, P. K. Bøcher, M. B. Greve, and M. H. Greve

Department of Agroecology and Environment, Faculty of Agricultural Sciences (DJF), Aarhus University, Blichers Allé 20, P.O. Box 50, 8830 Tjele, Denmark

Received: 21 December 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 18 January 2010
Revised: 11 May 2010 – Accepted: 17 May 2010 – Published: 1 June 2010

Abstract. Accurate information about organic/mineral soil occurrence is a prerequisite for many land resources management applications (including climate change mitigation). This paper aims at investigating the potential of using geomorphometrical analysis and decision tree modeling to predict the geographic distribution of hydromorphic organic landscapes in unsampled area in Denmark. Nine primary (elevation, slope angle, slope aspect, plan curvature, profile curvature, tangent curvature, flow direction, flow accumulation, and specific catchment area) and one secondary (steady-state topographic wetness index) topographic parameters were generated from Digital Elevation Models (DEMs) acquired using airborne LIDAR (Light Detection and Ranging) systems. They were used along with existing digital data collected from other sources (soil type, geological substrate and landscape type) to explain organic/mineral field measurements in hydromorphic landscapes of the Danish area chosen. A large number of tree-based classification models (186) were developed using (1) all of the parameters, (2) the primary DEM-derived topographic (morphological/hydrological) parameters only, (3) selected pairs of parameters and (4) excluding each parameter one at a time from the potential pool of predictor parameters. The best classification tree model (with the lowest misclassification error and the smallest number of terminal nodes and predictor parameters) combined the steady-state topographic wetness index and soil type, and explained 68% of the variability in organic/mineral field measurements. The overall accuracy of the predictive organic/inorganic landscapes' map produced (at 1:50 000 cartographic scale) using the best tree was estimated to be ca. 75%. The proposed classification-tree model is relatively simple, quick, realistic and practical, and it can be applied to other areas, thereby providing a tool to facilitate the implementation of pedological/hydrological plans for conservation and sustainable management. It is particularly useful when information about soil properties from conventional field surveys is limited.

Correspondence to: R. Bou Kheir
(rania.boukheir@agrsci.dk)

1 Introduction

Detailed soil spatial information is indispensable for land resources management and environmental modeling. Distribution patterns of organic/mineral soil occurence have a large potential to affect global climate, and the international efforts for using soils and vegetation as carbon sinks are rapidly increasing (IPCC, 2000). Changes in soil organic distribution are attributed to both natural processes and human activities, the latter being widely recognized in recent years. Land use changes, including deforestation, biomass burning, draining of wetlands (being usually humus-rich), ploughing, use of fertilizers and other agricultural practices, are regarded as the main factors causing loss of soil organic carbon (SOC) and the emission of CO_2 into the atmosphere. These changes can be significant in hydromorphic grasslands and croplands where intensive artificial drainage activities are carried out. This is particularly true in the case of Denmark with an important reduction of the total wetland area during the past 200 years as a result of much drainage activity (digging of drainage ditches and introduction of tile drainage).

As part of international efforts to stabilize atmospheric greenhouse gas concentrations, Denmark (like several other countries) is committed to establish inventories of

organic/mineral soil distribution in the frame of Kyoto protocol. Modeling tools of diverse soil properties (including organic/mineral soil occurence) require more information than available even in detailed soil maps. Digital Soil Mapping has been tested in a wide range of soil mapping contexts at different scales throughout the world (McBratney et al., 2003; Dobos et al., 2006; Grunwald, 2006). It has been used to understand and quantify the relationships between soils and their environmental attributes, mostly derived from exhaustive and easy-to-access datasets such as Digital Elevation Models (DEMs) and remote sensing imagery. According to Bishop and Minasny (2005) and McBratney et al. (2003), in almost 80% of digital soil mapping projects DEMs are used as the most important data source to derive landforms and run predictions of soil properties. Topographic attributes control the differential distribution of water, sediments, and dissolved material, which in turn result in soil differentiation (Pachepsky et al., 2001).

Soil landscape modeling has been successfully applied to predict soil variability in small landscapes of less than 100 ha (Moore et al., 1993; Gessler et al., 2000; Florinsky et al., 2002). These studies have demonstrated that combinations of one to five terrain attributes derived from DEMs can explain 20 to 88% of the variability of selected soil properties. The empirical relationships between soil properties and terrain attributes are unique to each soil property and each soil-forming environment. Recent soil landscape predictive algorithms such as neural networks, fuzzy logic or tree model tools arose mainly from data-mining and machine learning fields, also referred to as knowledge discovery in a database in its overall process (Fayyad et al., 1996). Soil landscape prediction from existing maps involves recovering the mental model used by the soil surveyor to set up the map (Lagacherie et al., 1995; Bui, 2004). This is a reverse soil mapping process and has broad relevance to any other application of knowledge discovery from natural resource maps (Qi and Zhu, 2003). Many researchers have utilized other statistical methods, such as multiple regression, stepwise regression, stepwise principal component regression, and correlation analysis to study the relationships between DEM-derived terrain attributes and different soil attributes, but in most cases for specific, localized landscapes (Moore et al., 1993; Dobos et al., 2000; Gessler et al., 2000; Egli et al., 2006a; Hengl, 2009).

Classification tree analysis (CTA) is a modeling technique that is being used increasingly (Henderson et al., 2004; Lawrence et al., 2004), being dedicated to the prediction of categorical data (classes of soil properties). It significantly enhances the ability of the DEM-derived variables to predict soil attributes (e.g. organic/mineral soil occurrence). CTA has several advantages that seem to suit well soil-landscape modelling applications. One of the most interesting features is that they are non-parametric, which means that no assumption is made regarding variable distribution (Breiman, 2001). Thus, it avoids variable transformation caused, for instance, by bi-modal or skewed histograms, which are frequent in soil class signatures (Lawrence et al., 2004). They are non-sensitive to missing data, perform automatic variable subset selection, are not sensitive to the inclusion of a large number of irrelevant variables, and finally, they can handle quantitative and categorical data, making it possible to integrate DEM-derived variables and indexes together with geology or soil categorical layers (Breiman, 2001; Henderson et al., 2004; Lawrence et al., 2004). However, in the built classification trees, the uncertainties of the classes in each one of their leaves can be explored. Efficiency of using CTA for predictive soil landscape mapping was demonstrated in a few studies at regional and subregional scale (Moran and Bui, 2002; Scull et al., 2005). Recent studies showed their potential for land cover mapping from remote sensing images analysis (Friedl et al., 1999; Lawrence et al., 2004), geomorphological mapping (Luoto and Hjort, 2005) and soil erosion occurrence (Bou Kheir et al., 2008).

As mentioned by Luoto and Hjort (2005), CTA was practically used in two linked but distinct purposes: induction and prediction. Induction-oriented studies used CTA to uncover the relationship between soil units or properties and environmental attributes, to identify the discriminant variables and to compare rules determined by the model with expert knowledge-based rules (McKenzie and Ryan, 1999; Bui et al., 2006). On the other hand, prediction-oriented studies used quantitative relationships between the soil response variables and the environmental soil-forming factors to predict soil landscape patterns over unvisited areas (Lagacherie et al., 1995; Moran and Bui, 2002; Scull et al., 2005).

The purpose of this study is to implement CTA and evaluate its ability to provide accurate soil landscape prediction and more precisely to determine the geographic distribution of hydromorphic organic landscapes (target variable being the organic/mineral soil occurence) depending on the existing field surveys' data collected during the last 60 years at an unsampled area in Denmark from mapped environmental variables. Our hypothesis was that spatial patterns of organic ditribution in hydromorphic landscapes could be predicted from spatial patterns of terrain attributes that have been shown to influence soil-forming processes. Prediction of hydromorphic organic landscapes will have implications for the proper management of marginal and environmentally sensitive areas. Understanding how soil organic distribution varies across landscape positions based on limited field samples has become the focal point of much environmental research nowadays.

2 Study area description

The chosen study area, covering about 1812 km^2, is located in southern Denmark (Fig. 1). It has been selected due to the strong link between land use on historical maps and soil internal drainage (Dalsgaard, 1997), which induces the

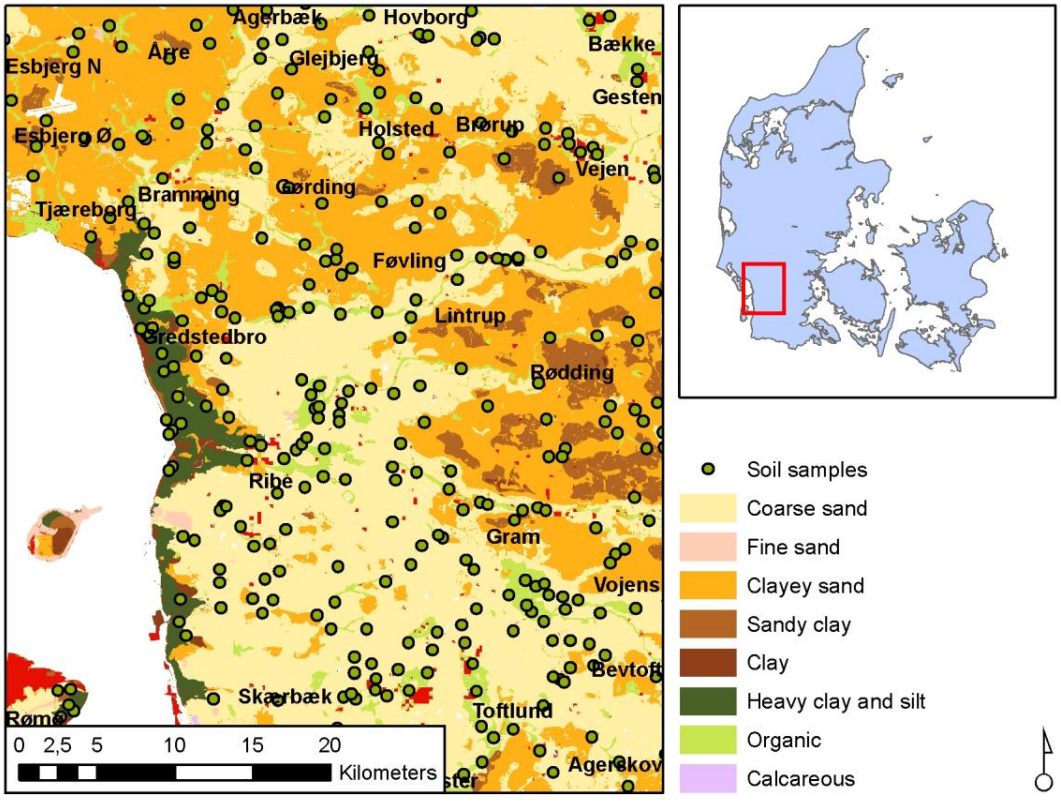

Fig. 1. Soil map of the study area within Denmark (Madsen et al., 1992).

accumulation of organic carbon on poorly drained Danish soils (Madsen et al., 1992). 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.

3 Materials and methods

The spatial prediction of hydromorphic organic landscapes was realized in several steps, combining existing soil survey collection, geomorphometrical analysis and decision tree modeling. Some existing field surveys were collected for specifying organic soils at visited locations. The obtained field samples' layer information (point location) was then intersected with maps of predictor parameters (as extracted from DEMs and other sources). A large number of un-pruned and pruned classification-tree models (186) were explored on the result of this intersection combining field samples locations and the corresponding parameters. The best tree model with the lowest misclassification error and the lowest number of terminal nodes and predictor parameters was used for producing a predictive map of organic/inorganic landscapes' map within the hydromorphic landscapes of the study area using GIS (Geographic Information Systems).

3.1 Soil samples collection and analysis

The soil was sampled at 1541 sites selected by four different existing field 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 0–30 cm (plough layer) in the Danish Soil Classification (1975) and the Danish Profile Investigation (1990). The collected samples in these two existing 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 soil organic carbon (SOC) were determined by the

combustion method in a LECO induction furnace, converted to % Soil Organic Matter (SOM) using a factor of 1.72. 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 (organic/mineral soil occurence) using 10% SOM as a cut off value (commonly used in Denmark). With less than 10% SOM, soils are classified as mineral; and with more than 10% SOM, soils are considered organic.

3.2 Geomorphometrical analysis

It has been postulated in several studies that the occurrence of hydromorphic organic landscapes is dictated by topographic features of the landscape (Moore et al., 1993; Gessler et al., 1995; Bou Kheir et al., 2007). In this study, we consider easy to derive and interpret terrain parameters from Digital Elevation Models.

3.2.1 Generation of Digital Elevation Model

A digital elevation model (DEM) was generated for the chosen area from airborne LIDAR (Light Detection and Ranging) systems. The latter seem effective and reliable means of terrain data collection in relatively large areas with cloudy weather conditions (Baltsavias, 1999; Brian et al., 2007; Schmitt et al., 2007; Liu, 2008). The established triangular irregular network (TIN) was converted using a TOPOGRID algorithm to an ArcGIS grid of 1.6-m pixel resolution. This resolution was chosen to match the planimetric and altimetric accuracies of LIDAR systems. In order to increase the efficiency in terms of storage and manipulation, and to acquire homogeneity and standardization with used ancillary maps, the constructed high-resolution DEM was coarsened in this study to 25-m resolution.

The produced elevation surface (DEM) would still contain several spurious elements, usually classified either as sinks or peaks (one or two cells below or above the local surface). The errors vary between 0.1 m and 4.7 m in a typical 25 m DEM (Tarboton et al., 1991). Although many authors agree that sinks and peaks may actually represent the true nature of topography (Chorowicz et al., 1992), they may act as local barriers that trap water flow and cause a major problem for drainage network extraction. To avoid this problem and before performing any hydrologic analysis, sinks in the DEM were identified and eliminated using TerraStream software (Danner et al., 2007).

3.2.2 Derivation of morphological/hydrological parameters from Digital Elevation Model

The chosen morphological/hydrological predictor parameters may aid spatial estimation of hydromorphic organic landscapes, because the relief had a great influence on soil formation and its physical/chemical properties (McKenzie and Ryan, 1999; Bou Kheir et al., 2007, 2008). They may be divided into primary and compound attributes. In this study, the nine primary parameters, i.e. elevation, slope angle, slope aspect, plan curvature, profile curvature, tangent curvature, flow direction, flow accumulation, and specific catchment area were directly derived from the constructed Digital Elevation Model (DEM) using specific TerraSTREAM (Danner et al., 2007) and ArcGIS (version 9.3) algorithms.

Elevation is useful for classifying the local relief, and locating points of maximum and minimum heights. It had a high correlation with organic/mineral soil occurence (Thompson and Kolka, 2005). At regional scales, several authors found that soil organic carbon content increased with elevation over ranges of $\geq$1000 m, since lower temperatures characterize higher elevations (Bolstad et al., 2001; Egli et al., 2003, 2006b).

Slope, S, characterizing the spatial rate of change of elevation in the direction of steepest descent, affects the velocity of both surface and subsurface flow, and hence the water and organic carbon contents in landscapes.

As for slope aspect, ψ (orientation of the line of steepest descent), is useful for visualizing hydromorphic organic landscapes, and is frequently recorded in pedological/hydrological surveys. Aspect is divided into the eight major directions plus the non-oriented flat areas. Slopes exposed to the south and west are more subject to runoff for two reasons: (1) they are warmer with higher evaporation rates and lower moisture storage capacity, thus less forested than those exposed to the north and east, and (2) rainfall affects slope aspect depending on the direction of winds during rainfall, which commonly has a west and south–west trend in Denmark.

Slope curvature, K, measures the distribution of convex and concave areas; hence the propensity of water to converge or diverge as it flows across the land. Convex surfaces are most likely to be well drained, while for concave surfaces depressions have a higher likelihood of having hydromorphic features. Concave slopes can concentrate more water and sediments indicating the potential accumulation of a large quantity of organic soils. Convex slopes show an inverse effect, dispersing flow and limiting material accumulation, therefore a lesser quantity of soil tends to accumulate than on concave slopes. Flat areas (zero curvature) are without any effect on flow divergence or convergence. Curvature attributes (plan, profile and tangent) are based on second derivatives: the rate of change of a first derivative such as slope gradient or slope aspect, usually in a particular

direction. Curvatures were derived through GIS from the constructed DEM.

The type and the amount of soil organic carbon (SOC) are strongly related to the presence of water. The drainage network provides an important indication of water percolation rate. Special hydrological algorithms were used depending on known matrices (i.e. flow direction matrix, flow accumulation and stream network) to derive the drainage network running over the study area (Tarboton et al., 1991; Chorowicz et al., 1992). A stream network was derived by connecting all pixels that accumulate flow from 100 pixels or more. Flow accumulation grid and digitized outlets from the stream network were used to automatically subdivide the whole area into small watersheds. Each watershed was subdivided into two facets, separated by the streamline passing through the watershed.

The specific catchment area, representing the upslope area per unit width of contour, was calculated using the finite difference slope algorithm and FD8 flow-routing method with a maximum area of $50\,000\,m^2$. FD8 was chosen because it allows flow to be distributed to multiple nearest neighbor nodes in upland areas above different channels, thus modeling flow divergence using flow dispersion. It takes considerably longer to run than the more common D8 algorithm but it avoids many of the problems incurred with D8 and gives much more realistic distributions of contributing area (Gallant and Wilson, 2000).

In addition to primary terrain attributes, a compound topographic index (CTI), often referred to as the steady-state wetness index was also calculated for each pixel using the average upslope contributing area (As) and the slope degree (β), according to the formula (CTI=ln [As/tan β]) (Moore et al., 1993; Wilson and Gallant, 2000).

3.3 Collection of other predictor parameters (soil, parent material and landscape)

Other predictor parameters (soil type, parent material and landscape type) were incorporated also in the constructed decision-tree models for mapping hydromorphic organic landscapes. Soil types were represented by a digital registered form of the available choropleth Danish soil classification map compiled by Madsen et al. (1992) at 1:50 000, and classifying the agricultural areas of Denmark into eight textural classes. Parent material was extracted from scanned and registered national geological maps of Denmark at 1:25 000 cartographic scale (Danmarks Geologiske Undersøgelse, 1978). The major Danish landscape types, considered spatially homogeneous geomorphic units in terms of both environmental characteristics and SOC content, were derived from the existing digital vector landscape map at 1:100 000 scale (Madsen et al., 1992). We did not use climatic data in this study since Denmark is relatively a small country with low topographic relief. Moreover, the main factor controlling soil moisture is local topography and soil conditions, which were retained in the constructed classification tree-models.

3.4 Decision-tree 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 organic/mineral soil occurence. This file was converted to a square grid that matched the resolution of the constructed DEM (25 m). ArcGIS was used to overlay morphology, hydrology, soil, geology, and landscape variables to each of the field survey (sampling) locations.

Spatial prediction of organic/inorganic landscapes was produced using tree-based classification models. The dependent variable is categorical (organic landscapes/inorganiclandscapes) and the independent variables are both continuous (elevation; aspect; slope; plan, profile and tangential curvature; flow accumulation; flow direction; rate of change of specific catchment area along the direction of flow; steady-state topographic wetness index) and categorical or nominal (soil type; geological substrate; landscape type).

Four sets of un-pruned classification tree-models were explored based on (1) all of the variables, (2) the primary morphological/hydrological variables, (3) selected pairs of variables, and (4) excluding each variable at one time from the potential pool of predictor variables. Once the trees have been developed, they encode a set of decision rules that define the range of conditions (values of environmental variables) best used to predict each organic or mineral soil occurence. The process is recursive, growing from the root node (the complete data set) to the terminal nodes in a dentritic fashion (Friedl and Brodley, 1997). The trees created are usually very large with multiple terminal nodes, meaning that the models are intimately fitted on the training data (Lagacherie et al., 1995). Each terminal node is assigned to the label of the majority class (Lees and Ritman, 1991). Splits or rules defining how to partition the data are selected based on information statistics that measure how well the split decreases impurity (heterogeneity or variance) within the resulting subsets (Clarke and Pregibon, 1992). The number of splits to be evaluated is equal to $2^{(k-1)}-1$, where k is the number of categorical classes of predictor parameters (Breiman, 2001). For example, if the soil type with 8 classes is considered, 127 splits are tried; if there are 12 classes (landscape type), 2047 splits are tried. We considered differences in the value of a continuous variable up to 1% of the whole range, which is equivalent to ten thousand classes (Loh and Shih, 1997).

The algorithm used for evaluating the quality of the constructed trees is the Gini splitting method, which is considered as the default method (Breiman, 2001). The Gini

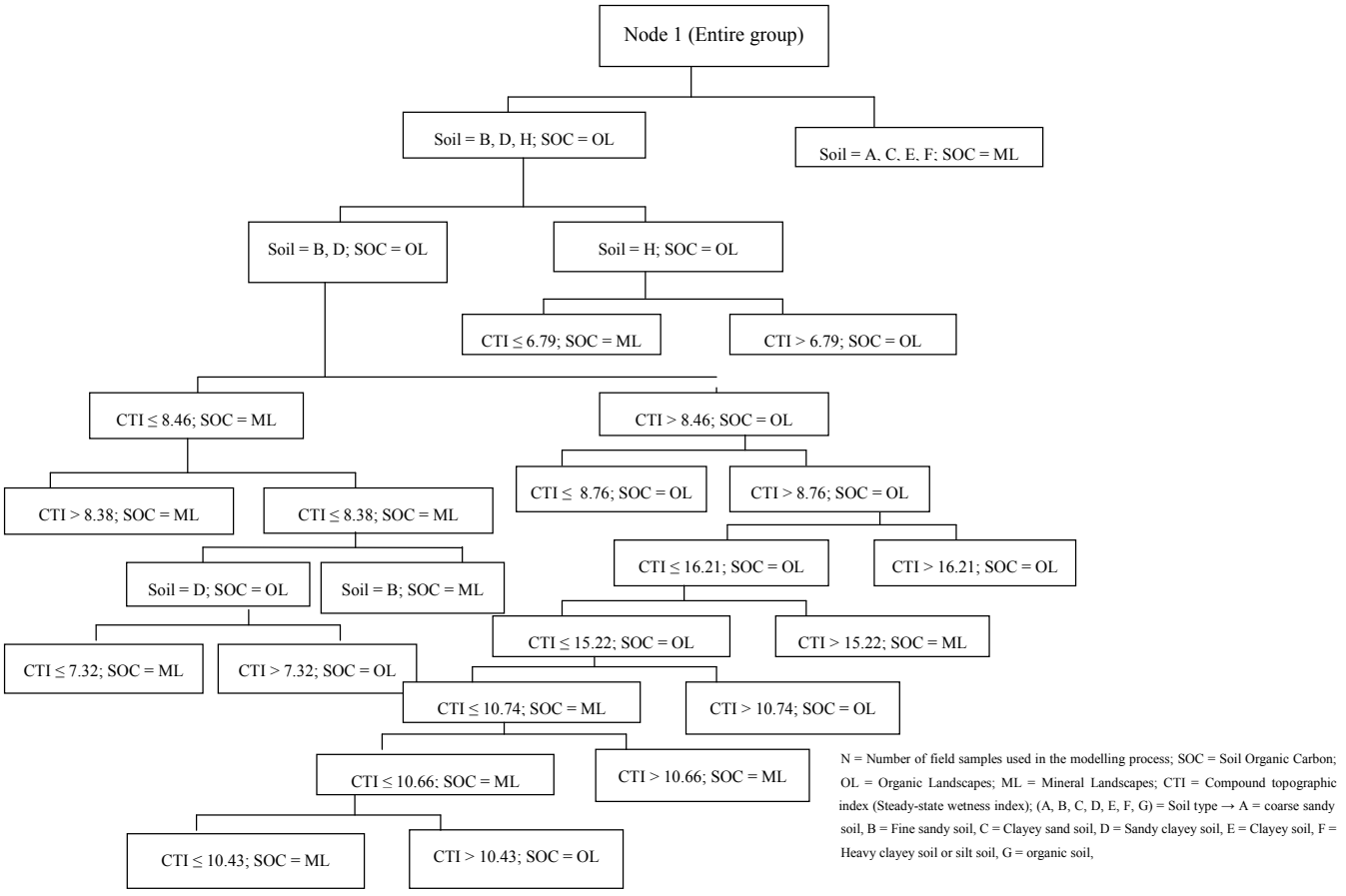

Fig. 2. Classification-tree model based on the combination of soil type and steady-state wetness index for predicting the spatial distribution of hydromorphic organic landscapes.

coefficient is used to measure the degree of inequality of a variable in terms of frequency distribution. It ranges between 0 (perfect equality) and 1 (perfect inequality). The Gini mean difference (GMD) is defined as the mean of the difference between each observation and every other observation (Breiman, 2001) (Eq. 1):

$$\mathrm{GMD} = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{K=1}^{N} \left\{ \left| X_j - X_K \right| \right\} \quad (1)$$

Where X is cumulative percentage (or fractions) and their respective values (j and k) and N is the number of elements (observations).

Pruning the constructed trees is necessary to prevent the models from being overfitted 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). Besides, relatively important variables can be pointed out by counting the times the variable was used in nodes (Bui et al., 2006). However, inconsistencies within the training dataset, such as noise or outliers, can greatly affect the classifier's accuracy (Lagacherie and Holmes, 1997).

3.5 Production of the predictive organic/inorganic landscapes' map

Using the preferred classification-tree model (having the highest predictive power, and the lowest number of terminal nodes and predictor parameters), a predictive map of organic/inorganic landscapes was obtained under a GIS environment through the application of the prediction classification tree rules (shown in Fig. 2). This map was validated based on field surveys.

3.6 Accuracy assessment procedure

The basis of the validation techniques was used to exclude a fraction of the sample from the modeling process and to compare the predicted value of these samples with their reference value. We applied two distinct validation procedures

Table 1. The different terrain parameters (predictors) extracted from DEMs likely to impact on the organic/mineral soil distribution and their corresponding classes.

Variable	Source, description	Range
Elevation	Lidar Digital elevation models (DEMs)	0 to 85 m
Slope	From DEMs by first order finite difference	0 to 64.5°
Aspect	The direction of the steepest downslope slope	0 to 360°
Plan curvature	Curvature of contour drawn through the grid point	−8.5 to 10
Profile curvature	Curvature of the surface in the direction of steepest descendent	−9.9 to 10
Tangent curvature	Plan curvature multiplied by sine of slope angle	−15 to 15
Flow accumulation	Upslope number of grid cells	1 to 59 000
Flow direction	Direction of the steepest drop	1 to 255
Specific catchment area	Upslope area per unit width of contour	32 to 8 268 160
Steady-state wetness index (CTI)	Modeled from DEMs; ln(As/tan β); As is upslope catchment area, β is slope (Moore et al., 1993)	3.05 to 36.39

in order to: (i) assess to what extent the constructed classification tree-models provided accurate prediction of the hydromorphic organic landscapes (internal validation), and (ii) derive the overall accuracy of the produced predictive organic/inorganic landscapes' map (external validation). The former uses training samples (80% of the field data or 1233 sites) that were collected within the training area, whereas the latter uses geographically distinct validation areas (20% of the field data or 308 sites). The internal validation scheme is used to test the efficiency of classification tree analysis (CTA) to predict soil landscape distribution using misclassification errors. The external validation was carried out on the full produced predictive organic/inorganic landsacpes' map. The accuracy assessment used in the external validation is summarized in the error matrix. The matrix shows the overall accuracy rate which is a simple ratio between the correctly allocated number of field samples (confusion matrix diagonal) and the overall number of classified samples.

4 Results and discussion

4.1 Derived terrain attribute maps

Ten primary and secondary topographic attribute grid maps were obtained (Table 1). These maps displayed the surface morphology, zones of soil water saturation and areas likely to have high soil organic carbon content in the study area. The steady-state topographic index (CTI) describes the distribution and extent of zones of soil water saturation. Small values of CTI generally depict upper catenary positions, and large values lower catenary positions with an overall range typically from around 3 to 36. The largest (i.e. high wetness) values are predicted in topographic hollows at higher elevations (i.e. in local areas with convergent flow lines) and immediately above gently sloping areas near channels (i.e. footslopes) in flatted terrain.

4.2 Tree-model evaluation

Training misclassification error rates for the explanatory trees that were developed using all variables (Model 1) at a time or the primary morphological/hydrological 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 2.

Applying cost-complexity pruning indicated that Model 1 (based on all variables) would classify correctly 67% of the tested organic/mineral soil occurence selecting just nine terrain variables (with their relative importance shown in parentheses): landscape type (100%), soil type (29%), elevation (22.5%), steady-state wetness index (20%), flow accumulation (15%), tangent curvature (14%), aspect (11%), and slope (9%). Model 2 (based on morphological/hydrological variables only) slightly reduced the explained accuracy and classified 64% 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 organic/mineral soil occurence (Table 3). The model based on soil type and steady-state topographic wetness index (CTI) (Model 3) showed the highest predictive power, classifying 68% of the data correctly and pruned to fourteen terminal nodes. The CTI proved to have a significant contribution to the estimation of hydromorphic organic landscapes since it is a predictor of zones of soil saturation, and organic carbon 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 soil organic carbon, and (2) concave slopes

Table 2. Relative importance of predictor variables and misclassification error rates in Models 1 (based on all variables) and 2 (based on morphologic/hydrologic variables only).

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	30%	15%	4%	0%
Flow direction	25%	0%	3%	0%
Specific catchment area	0%	0%	0%	0%
Steady-state wetness index	37%	20%		
Geological substrate	31%	0%	Not included in building the tree	
Soil type	39%	29%		
Landscape type	100%	100%		
Misclassification error (%)	23%	33%	26%	36%
Accuracy (%)	77%	67%	74%	64%
Tree size- terminal nodes	71	9	69	10

Table 3. Accuracy explained (%) for pruned classification tree models based on pairs of variables.

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

[a] a = elevation, b = aspect, c = slope, d = profile curvature, e = tangent curvature, f = plan curvature, g = flow direction, h = flow direction, i = specific catchment area, j = steady-state wetness index, k = geological substrate, l = soil type, m = landscape type

can concentrate more water and sediments indicating the potential accumulation of a large quantity of soil organic carbon (SOC).

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 organic landscapes. 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).

Removal of 13 variables one at a time had some effect on training missclassification error rates (decrease or increase) depending on the excluded variable. The three variables that, when excluded from the tree model, caused the greatest increase in error rate were steady state topographic wetness index, specific catchment area and soil type (Table 4).

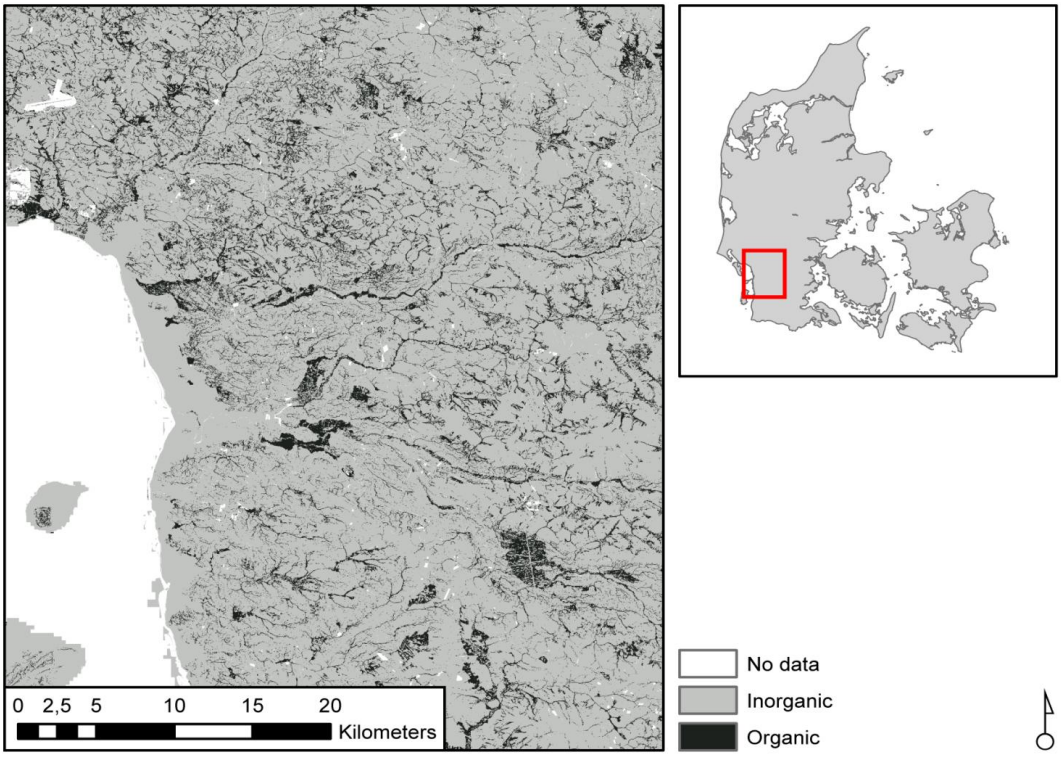

Fig. 3. Map showing the distribution of hydromorphic organic landscapes in the chosen study area within Denmark.

Table 4. Missclassification error rates (%) when excluding each parameter one at a time from the potential pool of predictor parameters.

Predictor parameter	Error rate (%)
Elevation	40
Aspect	34
Slope	40
Profile curvature	40
Tangent curvature	39
Plan curvature	40
Flow direction	40
Flow accumulation	40
Specific catchment area	42
Steady-state wetness index	43
Geological substrate	37
Soil type	43
Landscape type	38

The produced predictive map of organic/inorganic landsacpes (Fig. 3) at 1:50 000 cartographic scale using the tree model based on the combination of soil type and steady-state wetness index, indicates that 7.5% of the wetlands in the study area correspond to organic landscapes, and 92.5% to mineral (inorganic) landscapes. The confusion matrix between the measured organic/mineral soil occurence' classes and the modelled ones indicates a good overall accuracy of ca. 75%. This accuracy value is different from the explained variance of the preferred decision-tree model (68%), since it is dedicated to validate all adopted approaches combining the integration of soil survey collection, geomorphometrical analysis and decision-tree modeling.

5 Conclusions

Topographic variables (either morphologic or hydrologic) derived from DEMs are related to the geographic distribution of organic/inorganic landscapes. The preferred tree-based models explained 68–77% of the organic/mineral distribution for a series of chosen field sites in southern Denmark. Two environmental variables – soil type and steady-state topographic wetness index – proved to be the most important variables, indicating that complex or secondary topographic variables show stronger relationships to organic/mineral soil occurence than primary topographic attributes. This particular secondary topographic variable incorporated the effects of slope and upslope contributing area.

The decision-tree modelling approach was easily implemented with available GIS (ArcGIS) software and is suitable for data exploration and predictive organic/mineral soil occurence 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 organic/inorganic landscapes.

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 quantitative magnitude of soil organic carbon with greater precision and reliability.

Acknowledgements. This research is a part of the 10 million € SINKS project (2009–2012) aiming at improving the Danish estimate of greenhouse gas emission. The project is funded by Ministry of Climate and Energy.

Edited by: R. Purves

References

Baltsavias, E. P.: A comparison between photogrammetry and laser scanning, PRS, 54(2–3), 83–94, 1999.

Bishop, T. F. A. and Minasny, B.: Digital soil-terrain modelling: the predictive potential and uncertainty, Environmental soil-landscape modeling. Geographic information technologies and pedometrics, edited by: Grunwald, S., Taylor & Francis, Boca Raton, Florida, 185–213, 2005.

Bolstad, P. V., Vose, J. M., and McNulty, S. G.: Forest productivity, leaf area, and terrain in Southern Appalachian deciduous forests, Forest Sci., 47, 419–427, 2001.

Bou Kheir, R., Wilson, J., and Deng, Y.: Use of terrain variables for mapping gully erosion susceptibility in Lebanon, Earth Surf. Proc. Land, 32, 1770–1782, 2007.

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

Breiman, L.: Decision-tree forests, Mach Learn, 45(1), 5–32, 2001.

Brian, E., Roth, K., Clint Slatton, M., and Cohen, J.: On the potential for high-resolution lidar to improve rainfall interception estimates in forest ecosystems, Front Ecol. Environ., 5(8), 421–428, 2007.

Bui, E. N.: Soil survey as a knowledge system, Geoderma, 120, 17–26, 2004.

Bui, T.H., Zwieu, J., Poel, M., and Nijholt, A.: Toward affective dialogue modeling using partially observable markov decision processes, Proceedings of workshop emotion and computing, 29th Annual German Conference on Artificial Intelligence, edited by: Reichardt, D., Levi, P., and Meyer, J. C., Bremen, Germany, 47–50, June 2006.

Chorowicz, J., Ichoku, C., Riazanoff, S., Kim, Y. J., and Cervelle, B.: A combined algorithm for automated drainage network extraction, Water Resour., 28(5), 1293–1302, 1992.

Clarke, L. A. and Pregibon, D.: Tree-based models, In: Statistical Models, edited by: Chambers, J. M. and Hastie, T. J., Chapman and Hall, New York, 377–419, 1992.

Dalsgaard, K.: Examples on use of the 1844 land registration maps to illustrate the past environment, Aarhus Geoscience, 7, 141–146, 1997.

Danmarks Geologiske Undersøgelse: Foreløbige geologogiske kort (1:25,000) over Danmark. DGU Serie A, 3, 1978.

Danner, A., Yi, K., Mølhave, Th., Agarwal, P. K., Arge, L., and Mitasova, H.: TerraStream: from elevation data to watershed hierarchies. Proceedings of the 15thInternational Symposium on Advances in Geographic Information Systems (ACM GIS), 2007.

Dobos, E., Micheli, E., Baumgardner, M. F., Biechl, L., and Helt, T.: Use of combined digital elevation model and satellite radiometric data for regional soil mapping, Geoderma, 97, 367–391, 2000.

Dobos, E., Carré, F., Hengl, T., Reuter, H. I., and Toth, G.: Digital soil mapping as a support to production of functional maps – EUR 22123 EN. Office for Official Publication of the European Communities, Luxembourg, 2006.

Egli, M., Mirabella, A., Sartori, G., and Fitze, P.: Weathering rates as a function of climate: results from a climosequence of the Val Genova (Trentino, Italian Alps), Geoderma, 111, 99–121, 2003.

Egli, M., Wernli, M., Kneisel, C., Biegger, S., and Haeberli, W.: Melting glaciers and soil development in the proglacial area Morteratsch (Swiss Alps): II Modelling presentday and future soil state, Arctic, Antarctic, and Alpine Research, 38, 510–522, 2006a.

Egli, M., Mirabella, A., Sartori, G. Zanelli, R., and Bischof, S.: Effect of north and south exposition on weathering and clay mineral formation in Alpine soils, Catena, 67, 155–174, 2006b.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P.: From data mining to knowledge discovery in databases, American Association for Artificial Intelligence Press, Menlo Park, 1–34, 1996.

Florinsky, I. V., Eilers, R. G., Manning, G. R., and Fuller, L. G.: Prediction of soil properties by digital terrain modelling, Environ. Modell. Softw., 17, 295–311, 2002.

Friedl, M. A. and Brodley, C. E.: Decision tree classification of land cover from remotely sensed data, Rem. Sens. Environ., 61, 399–409, 1997.

Friedl, M. A., Brodley, C. E., and Strahler, A. H.: Maximizing land cover classification accuracies produced by decision trees at continental to global scales, IEEE Geosci. Remote, 37, 969–977, 1999.

Gallant, J. C. and Wilson, J. P.: Primary topographic attributes, in: Terrain analysis: principles and applications, edited by: Wilson, J. P. and Gallant, J. C., Wiley, New York, 51–86, 2000.

Gessler, P. E., Moore, I. D., McKenzie, N. J., and Ryan, P. J.: Soil-landscape modeling and spatial prediction of soil attributes, Int. J. Geogr. Inf. Syst., 9, 421–432, 1995.

Gessler, P. E., Chadwick, O. A., Chamran, F., Althouse, L. D., and Holmes, K. W.: Modelling soil-landscape and ecosystem properties using terrain attributes, Soil. Sci. Soc. Am. J., 64, 2046–2056, 2000.

Grunwald, S. (Ed.): Environmental soil-landscape modelling – Geographic Information Technologies and Pedometrics, CRC Press, New York, 2006.

Hengl, T.: A practical guide to geostatistical mapping, 2nd Edt. University of Amsterdam, 291 p, ISBN 978-90-9024981-0, http://spatial-analyst.net/book/GstatIntro, 2009.

Henderson, B. L., Bui, E. N., Moran, C. J., and Simon, D. A. P.: Australia-wide predictions of soil properties using decision trees, Geoderma, 124(3–4), 383–398, 2004.

IPCC: Land-use, land-use change, and forestry, in: Land-use, land-use change, and forestry, edited by: Watson, R. T., Noble, I. R., Bolin, B., Ravindranath, N. H., Verardo, D. J., and Dokken, D. J., A special report to the Intergovernmental Panel on Climate Change (IPCC), Cambridge University Press, Cambridge, UK, 1–51, 2000.

Lagacherie, P. and Holmes, S.: Addressing geographical data errors in a classification tree for soil unit prediction, Int. J. Geogr. Inf. Sci., 11, 183–198, 1997.

Lagacherie, P., Legros, J. P., and Burrough, P. A.: A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area, Geoderma, 65, 283–301, 1995.

Lawrence, R., Bunn, A., Powell, S., and Zambon, M.: Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis, Rem. Sens. Environ., 90, 331–336, 2004.

Lees, B. G. and Ritman, K.: Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in distributed hilly environments, J. Environ. Manage., 15, 823–831, 1991.

Loh, W. Y. and Shih, Y. S.: Split selection methods for classification trees, Stat Sinica, 7, 815–840, 1997.

Liu, X.: Airborne LiDAR for DEM generation: some critical issues, Progress Prog. Phys. Geog., 32(1), 31–49, 2008.

Luoto, M. and Hjort, J.: Evaluation of current statistical approaches for predictive geomorphological mapping, Geomorphology, 67, 299–315, 2005.

Madsen, H. B., Nørr, A. H., and Holst, K. A.: The Danish Soil Clasification. Atlas over Denmark I,3. The Royal Danish Geographical Society, Copenhagen, 1992.

McBratney, A. B., Mendonca, M. L., and Minasny, B.: On digital soil mapping, Geoderma, 117, 3–52, 2003.

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

Moore, I. D., Gessler, P. E., Nielsen, G. A., and Peterson, G. A.: Soil attribute prediction using terrain analysis, Soil Sci. Soc. Am. J., 57, 443–452, 1993.

Moran, C. J. and Bui, E. N.: Spatial data mining for enhanced soil map modeling, Int. J. Geogr. Inf. Sci., 16, 533–549, 2002.

Pachepsky, Y. A., Timlin, D. J., and Rawls, W. J.: Soil water retention as related to topographic variables, Soil Sci. Soc. Am. J., 65, 1787–1795, 2001.

Qi, F. and Zhu, A. X.: Knowledge discovery from soil maps using inductive learning, Int. J. Geogr. Inf. Syst., 17(8), 771–795, 2003.

Safavian, S. J. and Norvig, P.: A survey of tree classifier methodology, IEEE T. Syst. Man. Ct. A, 21, 660–674, 1991.

Scull, P., Franklin, J., and Chadwick, O. A.: The application of classification tree analysis to soil type prediction in a desert landscape, Ecol. Model., 181, 1–15, 2005.

Schmitt, N. P., Rehm, W. F., Pistner, T., Zeller, P., Diehl, H., and Navé, P.: The AWIATOR airborne LIDAR turbulence sensor, Aerospace Sci. Technol., 11(7–8), 546–552, 2007.

Tarboton, D. G., Bras, R. L., and Rodriguez-Iturbe, I.: On the extraction of channel networks from digital elevation data, Hydrol. Processes, 5, 81–100, 1991.

Thompson, J. A. and Kolka, R. K.: Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modelling, Soil Sci. Soc. Am. J., 69, 1086–1093, 2005.

Wilson, J. P. and Gallant, J. C.: Secondary topographic attributes, in: Terrain analysis: principles and applications, edited by: Wilson, J. P. and Gallant, J. C., Wiley: New York, 87–132, 2000.