Geoderma 214–215 (2014) 101–113

Contents lists available at ScienceDirect

Geoderma

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

Constructing a soil class map of Denmark based on the FAO legend using digital techniques

Kabindra Adhikari [a,*], Budiman Minasny [b], Mette B. Greve [a], Mogens H. Greve [a]

Metadata, citation and similar papers at core.ac.uk

ARTICLE INFO

Article history:
Received 16 May 2013
Received in revised form 12 September 2013
Accepted 23 September 2013
Available online 18 October 2013

Keywords:
FAO–Unesco revised legend 1990
Decision tree
Digital soil mapping
Soil classes
Denmark

ABSTRACT

Soil mapping in Denmark has a long history and a series of soil maps based on conventional mapping approaches have been produced. In this study, a national soil map of Denmark was constructed based on the FAO–Unesco Revised Legend 1990 using digital soil mapping techniques, existing soil profile observations and environmental data. This map was developed using soil-landscape models generated with a decision tree-based digital soil mapping technique. As input variables in the model, more than 1170 soil profile data and 17 environmental variables including geology, land use, landscape type, area of wetlands, digital elevation model and its derivatives were compiled. The predicted map showed that Podzols and Luvisols were the most frequent soil groups, covering almost two-thirds of the area of Denmark. Geographically, Podzols occupied a major portion of western Denmark, where the soils have developed on sandy parent material, whereas eastern Denmark mostly contained Luvisols developed on loamy basal till. The occurrence of the predicted soil groups was assigned using several variables, of the most important was clay content in the topsoil and subsoil, elevation, geology and landscape type. The overall prediction accuracy based on a 20% hold-back validation data was 60%, but increased to 76% when prediction accuracy of similar soil groups was considered. Podzoluvisols and Alisols were among the weakly predicted groups (<48% prediction confidence), whereas Podzols and Luvisols had the highest accuracy of prediction (>70%). Overall, the average prediction uncertainty was less than 34%. Compared to the existing conventional soil map, the new map showed promising predictions. Validation of the predicted map with different techniques (point validation, prediction confidence analysis, and map-to-map comparison) confirmed that the output is reliable and can be used in various soil and environmental studies without major difficulties. This study also verified the importance of GlobalSoilMap products and a priori pedological information that improved prediction performance and quality of the new FAO soil map of Denmark.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-SA license.

1. Introduction

Soil mapping in Denmark has a long history and a series of soil maps based on conventional mapping approaches have been produced. However, in the digital age, a fine-resolution soil map for the whole country is needed. In this study, a national soil map of Denmark was constructed based on the FAO–Unesco Revised Legend 1990 using existing soil profile observations and environmental data. This map was developed using soil-landscape models generated with a decision tree-based digital soil mapping (DSM) technique.

Conventionally, soil types are delineated in the field by pedologists or soil surveyors following a tacit mental model (Hudson, 1992), which explores the relationship of soil to its natural surroundings. These conventional maps have several limitations, in particular inadequacy in spatial details and issues concerning the accuracy of soil attributes (Hartemink et al., 2010; McSweeney et al., 1991; Zhu et al., 1997). Moreover, such qualitative mental models are rarely described in a clear manner (Jafari et al., 2012), suffering from personal bias, difficult to replicate, and are inflexible for quantitative studies (Hartemink et al., 2010). Nevertheless, despite the shortcomings of the conventional soil survey, it is rather difficult to be replaced by any mechanical models. However, as an analogy to the conventional mapping, the relationship of soil and environmental variables can be quantified statistically (Dobos and Hengl, 2009; Grunwald, 2006; McKenzie and Ryan, 1999; Minasny et al., 2008; Zhu et al., 2001) and used to spatially predict soil properties including soil class (Bou Kheir et al., 2010; Carre and Girard, 2002; Greve et al., 2012a; Minasny et al., 2013). It has been reported that quantitative empirical modelling can address the limitations of conventional surveys (Bui et al., 1999; Hewitt, 1993; Kempen et al., 2012; McKenzie and Ryan, 1999). Such statistical techniques, which are commonly referred to as DSM techniques (McBratney et al., 2003),

* Corresponding author at: Aarhus University, Department of Agro-ecology, Blichers Allé 20, Postboks 50, 8830 Tjele, Denmark. Tel.: +45 8715 4759; fax: +45 8715 4798.
E-mail address: kabindra.adhikari@agrsci.dk (K. Adhikari).

0016-7061© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-SA license.
http://dx.doi.org/10.1016/j.geoderma.2013.09.023

have been widely used in soil mapping applications during the past decade (Boettinger et al., 2010; Grunwald, 2009; Hartemink et al., 2008; Lagacherie et al., 2007; Minasny et al., 2012). Incorporating new ideas and thoughts through research and development, DSM has been maturing and becoming operational from local to continental or global scale, providing the data and information needed for a new framework for soil assessment to assist in addressing a number of global issues, such as food security and climate change, and also support for environmental policies (Carre et al., 2007; Finke, 2012; Grunwald et al., 2011; Sanchez et al., 2009).

While DSM for mapping soil classes has been trialled at a field, watershed or regional scale, fine resolution mapping at a national extent has not yet been compiled. Denmark as a country with rich soil information is a good application of the digital techniques. Moreover, the availability of a conventional soil class map of Denmark, although constructed at a rather coarse cartographic scale, also provides an opportunity to evaluate the prediction performance of the DSM model. Considering the fact that the FAO soil map legend is the most widely recognised soil mapping basis internationally, our DSM approach to map the FAO soil groups is also justifiable. Based on these premises, we hypothesised that the spatial distribution of soil types in Denmark is influenced by the environmental variables and that it is possible to quantify the relationship between soil types and those variables and map them using DSM techniques. The major objectives of this study were: (i) to spatially predict and map the FAO soil groups in Denmark using decision tree modelling; (ii) to identify the potential environmental drivers for soil type variability in Denmark; and (iii) to evaluate the model prediction performance using the conventional FAO soil map.

2. Materials and methods

2.1. Study area

The study area is Denmark, a country in Northern Europe lying between 54°33′35″–57°45′7″N and 8°4′22″–15°11′55″ E. It has a temperate climate, where the average annual temperature reaches up to 16 °C during summer and 0 °C during winter time. Precipitation is fairly evenly distributed throughout the year, with an average annual total ranging from 500 mm in the east to 800 mm in the west of the country (Danmarks Meteorologiske Institut, 1998). The country covers about 43,000 km^2 area, and nearly two-thirds of this area is used for intensive mechanised agriculture. The topography is relatively flat and smooth (mean elevation 32 m, highest point 171 m asl), but is rather complex in nature, developed by the late glacial and post glacial-marine transgressions and multiple glaciations during the Weichselian geological stage. All these led to a variable distribution and formation of different soil types in Denmark (Schou, 1949). The majority of the eastern and central part of the country is developed in moraine landforms with loamy soils, rich in soil clay content, whereas the western parts consist of older and more strongly eroded landforms and sandy glacio-fluvial flood plains. The northern parts consist of late and post-glacial marine deposits. The major crops grown across the study area include wheat, maize, potato and barley, and livestock production is very common in the west.

2.2. Pedological investigation and soil mapping in Denmark

Denmark has a long history of soil resource assessment, with the information mainly being collected for taxation purposes in the past. For example, King Christian V's Great Danish Land Register of 1688 classified the soils according to their potential yield of various crops (Greve et al., 2001). However, a detailed pedological investigation carried out during the 1980s established the first nation-wide soil profile database in Denmark. During this investigation, 7-km soil monitoring grids were set up and at each 850 grid intersection a detailed profile description and soil classification was made (Madsen and Jensen, 1985). Similarly, the establishment of the main gas pipeline system in Denmark in 1981 provided an opportunity to obtain more detailed soil observations. In 1985, the Commission of the European Communities (EC) published the first soil map of the entire EC at a scale of 1:1,000,000 according to the FAO–Unesco Soil Legend (FAO–Unesco, 1974) (Commission of the European Communities, 1985). However, the Danish part of the EC soil map was based on only a few soil profile data (Madsen and Jensen, 1996). Therefore, Madsen and Jensen (1995) improved on the Danish database and a new map was constructed (Madsen and Jensen, 1995). In 1990, FAO revised the legend to the soil map of the world that was published in 1974 at a 1:5,000,000 scale (FAO–Unesco, 1990). Following the revised legend, the soil map of Denmark was again updated at a 1:1,000,000 scale. The main aim of the conversion was to support soil data harmonisation in Europe, as many of the EC countries have already adopted this new system to update their national soil maps (Madsen and Jensen, 1996).

2.3. Profile observations and data preparation

Soil profile observation and soil classification in Denmark started during 1981–1984 in connection with the establishment of a main gas pipeline system from the North Sea gas fields across Denmark (Madsen and Jensen, 1985). Pedological investigations along these 2 m deep pipeline trenches comprised making 2–3 detailed profile studies per kilometre and taking soil samples from every horizon in each profile for laboratory analysis. In 1986, the Danish Agricultural Advisory Centre established a nation-wide 7-km grid, often called ‘the nitrate grid’, to study and improve nitrogen fertiliser use efficiency in Danish agriculture. This grid consists of about 850 intersections, of which 663 are located on farm land, 106 in forest and 51 on other land uses (Østergaard, 1990). Detailed pedological investigations have been carried out at all the grid intersections, and the soil profiles have been described according to the Bureau of Land Data (ADK) Manual, which in many aspects are very similar to the FAO (1977) guidelines for soil profile description. Soil samples from all the profiles taken according to the genetic horizon sequence have been analysed for texture, organic matter, pH and calcium carbonate contents, and for some selected profiles exchangeable bases, cation exchange capacity, total nitrogen and phosphorus. The analytical methods are described in Madsen and Jensen (1992). On the basis of profile investigations and supporting soil analytical data, it was possible to carry out soil classification according to the FAO–Unesco 1974 classification system and the names were revised according to the FAO–Unesco Revised Legend 1990 (Madsen and Jensen, 1996).

In the classification, more than 40% of the soil profiles in the study area were described as Phaeozems. However, Phaeozems in Denmark are artificially created due to liming and are unevenly distributed between Luvisols and Cambisols. Therefore Phaeozems were further categorised into Luvisols and Cambisols depending on the presence or absence of an argic horizon in the soil profile, according to Madsen and Jensen (1996).

From the entire study area, a total of 1171 soil profiles for which soils were classified were used in the present study. These data consisted of eight soil groups, namely Alisols, Arenosols, Cambisols, Fluvisols, Gleysols, Luvisols, Podzols and Podzoluvisols. The whole set of data was divided into training and validation data sets for model building and validation. A fraction of each soil group (80% of profiles) was first separated randomly, before being combined together into training data, and the remaining 20% of profiles were grouped in the same way to form a validation data. The geographical distribution and the location of soil groups together with the training and validation profiles across the study area are shown in Fig. 1.

2.4. Prediction covariates (‘scorpan’ factors)

The digital soil mapping model used is the so-called ‘*scorpan*’ model (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n) + e$$

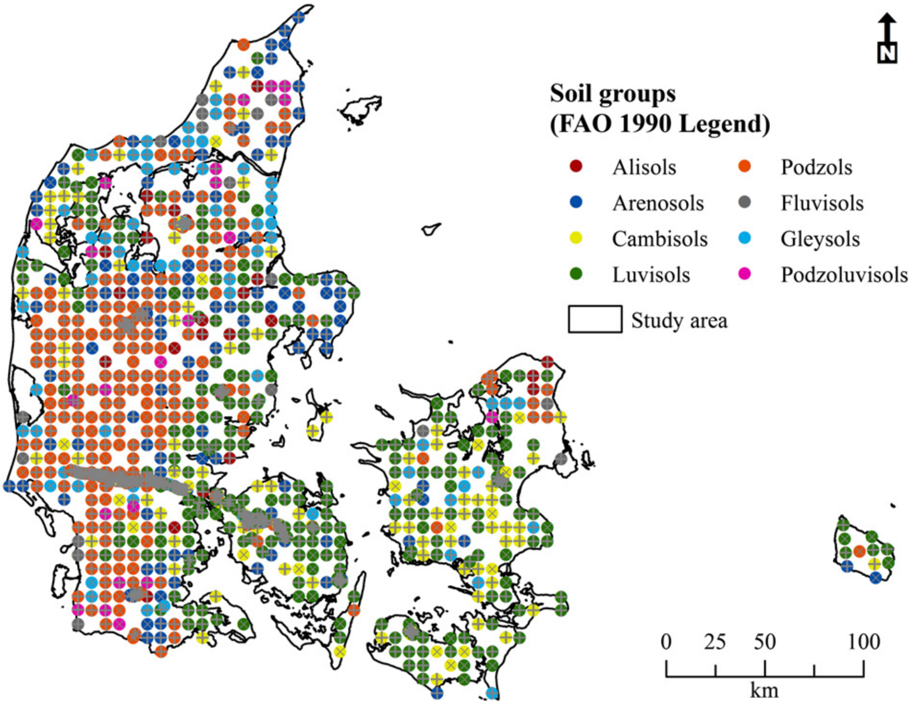


Fig. 1. Location of soil profiles in the study area (training profiles are marked with '+' and validation profiles with '×'). The soil groups for the two data sets are displayed in different colours.

where *S* is the soil class, and the '*scorpan*' factors stand for *s*oil, *c*limate, *o*rganisms, *r*elief, *p*arent material, *a*ge and spatial positio*n*. '*scorpan*' is an extension of the famous soil-forming equation '*clorpt*', by Jenny (1941), which was designed for modelling of soil genesis (standing for *cl*imate, *o*rganisms, *r*elief, *p*arent material and geological *t*ime). The *e* is a spatially correlated error, which is not modelled in this study. The empirical function *f* used in this study is a decision tree model.

A large number of *scorpan* variables can be used in DSM to predict soil types or soil properties. The present study used 17 environmental variables, many of which were derived from the air-borne LiDAR (Light Detection and Ranging)-based Digital Elevation Model (DEM) produced by the National Survey and Cadastre of the Danish Ministry of Environment in 2011. The LiDAR point clouds were converted to a raster of 1.6 m grid size for DEM, which was further aggregated to 30.4 m – a multiple of the original 1.6 m grid size – for use in this study. As far as soil variations in Denmark are concerned, the selected grid size is also comparable to the recent finding of Greve et al. (2012b) where model performance was better with 24 m than 90 m while mapping soil clay content. During post-processing of the DEM, all depressions of ≤50 cm depth in the DEM were filled so that they would not create problems during surface water flow and drainage network extraction. Multiple-flow direction (MFD) or FD8 algorithms (Freeman, 1991) were applied for all flow-related calculations. TerraStream algorithms (Danner et al., 2007) were used to create and process the DEM. From the filled DEM, 10 land surface parameters (LSP), namely slope aspect, direct sunlight insolation, distance to channel network, elevation, flow accumulation, mid-slope position, MRVBF (multi-resolution index of valley bottom flatness), SAGA WI (System of Automated Geoscientific Analyses Wetness Index), slope gradient and valley depth, were extracted in ArcGIS (ESRI, 2012) and SAGA GIS (SAGA GIS).

Maps of geology, geo-regions, landscape, land use and the extent of wetlands in the study area were also used as predictors. The geology map represents the extent and type of parent materials and was extracted from the national geological map (Danmarks Geologiske Undersøgelse, 1978). The map of geo-regions represents 10 distinct regions in Denmark based on climate and geographical settings. The landscape type map shows Danish landforms, mostly referring to quaternary geological developments (Madsen et al., 1992). The land use map corresponds to the land cover types derived from Corine Land Cover data specified for Denmark (Stjernholm and Kjeldgaard, 2004). Furthermore, to include the variability in soil clay content at different soil depths, which is an important parameter in FAO soil group nomenclature, national maps of soil clay content for the 0–30 cm and 60–100 cm soil layers were used as predictors. The clay map for 0–30 cm depth was generated by Greve et al. (2007) and that for 60–100 cm depth by Adhikari et al. (2013). All the predictors were projected to a common projection of ETRS1989 UTM32N and to a similar cell size (grid spacing) of 30.4 m for modelling. Table 1 shows the list of predictors used in the study and gives a brief explanation of each.

2.5. Quantifying the relationship between FAO-soil groups and environmental variables

In recent years, the use of data-driven machine-learning induction methods, such as tree-based methods, has gained popularity in DSM. Friedman and Meulman (2003) outlined several features and advantages of tree-based models and suggested these as a relevant approach for mapping soil properties or soil classes. Tree models are also reported to be capable of exploring the non-linear and complex soil–landscape relationship, which is very common in soil systems (Bui et al., 2006). Moreover, because of their potential and advantages in spatial pattern recognition, decision trees are increasingly used in soil class mapping (Grinand et al., 2008; Minasny and McBratney, 2007; Scull et al., 2005). As decision trees are able to identify the most decisive variables too, the value of the variables can also be evaluated (Bou Kheir et al., 2010). Other applications in which tree-based methods have been applied in soil class mapping include the use of a classification tree to predict soil units in a small area in a French Mediterranean valley (Lagacherie and Holmes, 1997) and the use of a similar tree model to map soil and surficial geology classes in Australia (Bui and Moran,

Table 1
List of environmental variables used in prediction and a brief description of each.

Variable	Mean (range)	Data source/brief description	'*scorpan*' factor[a]
Slope aspect[b]	181.17 (0–360)	DEM/direction of the steepest slope from the North	R
Clay content (0–30 cm)[b]	8.10 (0–67.95)	Clay content (%) for 0–30 cm soil depth	S
Clay content (60–100 cm)[b]	10.72 (0–57.10)	Clay content (%) for 60–100 cm soil depth	S
Direct sunlight insolation[b]	1269.05 (121.91–1706.98)	DEM/potential incoming solar radiation (insolation) calculated for a single year (Böhner and Antonić, 2009)	C
Distance to channel network[b]	559.45 (0–10,041.5)	DEM/calculates distance to channel network	R
Elevation[b]	31.97 (0–171.53)	DEM/LiDAR produced elevation of the land surface	R
Flow accumulation[b]	60.42 (1–110,907)	DEM/number of upslope cells	R
Geology[c]	84 classes	Scanned and registered geological map (*Scale* 1:25,000)	P
Geo-regions[c]	10 classes	Scanned geographical regions map (*Scale* 1:100,000)	C/R
Landscape[c]	12 classes	Landform types (*Scale* 1:100,000)	R
Land use[c]	31 classes	CORINE land cover data adopted in Denmark (*Scale* 1:100,000)	O
Mid-slope position[b]	0.025 (0–1)	DEM/covers the warmer zones of slopes (Bendix, 2004)	R/C
Multi-resolution index of valley bottom flatness[b]	4.26 (2.22–10.9)	DEM/calculates the depositional areas (Gallant and Dowling, 2003)	R
SAGA Wetness Index[b]	14.46 (6.87–19.09)	DEM/calculates slope and specific catchment area based Wetness Index. $\mathrm{WI} = \ln(As / \tan \beta)$: where As is modified catchment area and β is the slope (Böhner et al., 2002)	R
Slope gradient[b]	1.72 (0–90)	DEM/maximum rate of change between the cells and neighbours	R
Valley depth[b]	7.53 (0–89.88)	DEM/extent of the valley depth	R
Wetlands[c]	2 classes	Shows the presence or absence of wetlands (*Scale* 1:20,000)	R

[a] C, climate; O, organisms; R, relief; P, parent material.
[b] Continuous variables.
[c] Categorical variables.

2001, 2003). The boosting procedure has been implemented in the classification tree model in order to improve the prediction accuracy (Friedman, 2001). Boosting generated many decision trees from the same data set, calculated the weights for each tree (based on its accuracy), and combined them into a single prediction in such a way to reduce the bias in prediction. In essence, it constructed a set of 'weak learners' creating a single 'strong learner'. Moran and Bui (2002) applied boosting in the classification tree model to map soil types and reported that the classification error was minimised by the boosting option. Other examples of using a boosted classification tree include the study by Lacoste et al. (2011) and Lemercier et al. (2012), who predicted the distribution of soil parent material types at a regional scale in France.

The relationship between FAO soil groups and the environmental variables was constructed with a decision tree model incorporated in the *See5 2.08* data mining tool, which uses recursive partitioning of the predictors until the intra-subset variation at each node or leaf is minimised (Quinlan, 1993). This concept is similar to the classification and regression tree (CART) methodology developed by Breiman et al. (1984). However, unlike in CART, where Gini or diversity index is used, *See5* uses 'information gain' as a splitting criteria where splitting is based on the field that provides a maximum information gain (Patil et al., 2012; Zhu et al., 2009). This classifier first constructs a fully grown large tree to fit the data and then the tree is pruned back by removing the parts which are predicted with a high error rate. However, in order to improve the predictive performance of the constructed tree, adaptive boosting was applied with 10 iterations or trials. The boosting approach instructs the model to give more attention to the errors generated by the previous classifier and minimise these in the succeeding trials such that the overall performance of the model is improved. It was suggested that boosting with 10 classifiers reduced the error rate by about 25%. In the constructed tree, a global pruning with a standard value of 25% was selected. A value smaller or higher than this standard causes severe or less pruning, respectively. Similarly, a minimum number (n) of cases that must follow at least two of the branches was chosen as $n = 2$. A value higher than this can compromise the tree size by approximate data fitting (Quinlan, 1993).

Moreover, to discuss the specific contribution of two main soil data source used in this study – i) national map of soil clay content from 0 to 30 cm and from 60 to 100 cm soil depths as a GlobalSoilMap product, and ii) a rich a priori pedological knowledge of Danish soils, e.g., a map of peatland areas, small areas in marsh landscape types and coastal dunes – a separate classifier was generated without clay data, and the output was not adjusted with pedological information. The results were evaluated to see whether the GlobalSoilMap database and a rich a priori soil information have some added value in mapping FAO soil groups in Denmark.

2.6. Variables of importance in the prediction

The environmental variables used to predict soil groups showed different levels of contribution to the classifier generated for prediction. These contributions were assessed through the relative importance (RI) of the variables used in the prediction model. Although the *See5* tool does not directly give the RI, it quantifies the utility (expressed as a percentage) of each variable used in the model. It represents the percentage of input data for which the value of the variable is known and used in prediction. These values were considered a measure of the capability of the variables to predict soil groups.

2.7. Confidence of prediction

The programme *See5* calculated the confidence of prediction as an index between 0 (least confidence) and 1 (most confident). If a case is classified by a single leaf of a decision tree, the confidence value is the proportion of training cases at that leaf that belong to the predicted class. If more than one leaf is involved, the value is a weighted sum of the individual leaves' confidences. For a single ruleset, each applicable rule 'votes' for a class with the weight of voting using the rule's confidence value. The confidence value for boosted classifiers is similar, where the individual classifiers vote for a class with the weight equal to their confidence value (Ross Quinlan, personal communication).

2.8. Mapping to the spatial domain

Once the relationships between soil groups and environmental variables had been established, the decision tree model was applied to the whole set of covariate data and FAO soil groups were mapped across the study area. Programmes written in FORTRAN were used to convert the model output to the grid.

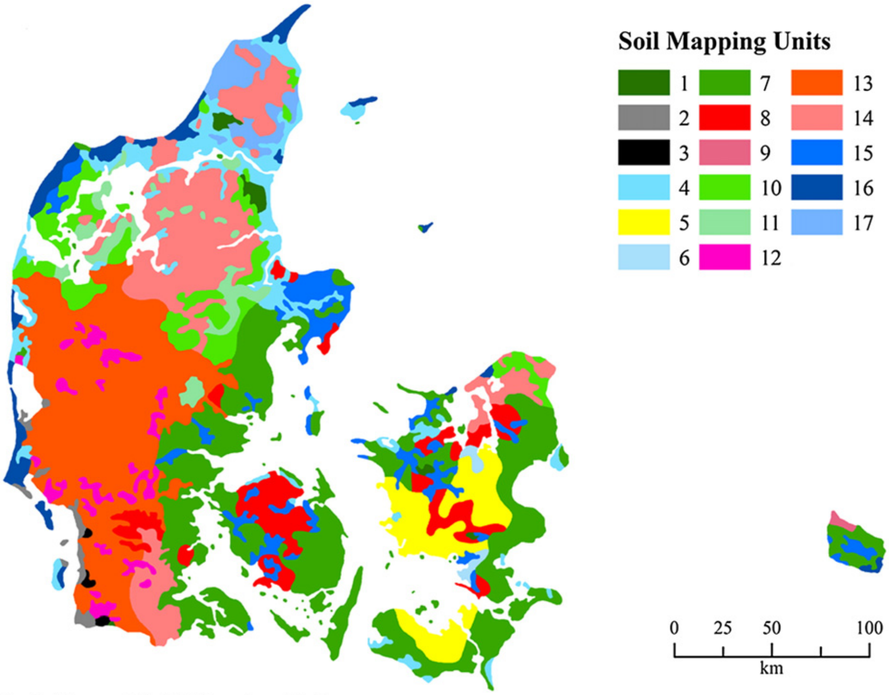
Similarly, the error associated with the prediction was assessed. The model confidence to predict soil groups in each leaf was derived and a continuous map was generated. This map was expressed as an indicator of uncertainty associated with the prediction, as it shows how certainly each soil group was predicted. An average prediction confidence for all the pixels of each soil group throughout the study area was also evaluated. Besides, confidences of correctly classified and misclassified soil groups in validation locations were also assessed. This provides a response of misclassification when prediction confidence changes. Moreover, a *t*-test was performed to check whether the mean prediction confidence between misclassified and correctly classified soil groups was significantly different.

2.9. Pedological adjustments in the prediction

To improve the map reliability, predicted classes from some specific areas were reclassified to the defined classes considering an exhaustive knowledge on pedogenesis and soil–landscape interaction prevailing in those areas. Young sand dunes along the coastline especially in the west and that from some small island area were recognised as Arenosols. The spatial extent of Histosols was defined using the recent map of Danish peatlands. Similarly, soils developed in reclaimed areas were named as Gleysols, and that from the Marsh areas as Fluvisols. The procedure included burning of the predicted raster map with the reclassified soil groups in those specific areas. The map of landscape types was considered to identify such specific areas. However, no such pedological adjustments were adopted when the model was run without soil clay data from both the top and subsoil depths. This suggests whether exclusion of such a rich a priori soil information during prediction compromises with map quality.

2.10. Model validation using training and validation data sets

The performance of the decision tree model to predict soil groups was evaluated on 20% validation profiles, which were not used in the model building. The performance was also checked internally, following



Definition of Soil Mapping Units

1: HSf (70%); HSt (20%); HSs (10%)

2: FLe-2 (70%); FLt-2 (30%)

3: FLt-3 (70%); FLe-3 (30%)

4: GLe-1 (40%); GLm-1 (20%); GLd-1 (20%); GLt-1 (10%); ARa-1 (10%)

5: CMe-2 (40%); CMg-2 (20%);; LVg-2 (10%); LVh-2 (10%); CMk-2 (10%); ARb-1 (5%); GLm-2 (5%)

6: CMe-1 (60%); Lo-1 (10%);; Lg-1 (10%); CMg-1 (10%); ARb-1 (5%); GLm-1 (5%)

7: LVh-2 (50%); LVg-2 (25%); CMe-2 (10%); CMg-2 (5%); ARb-1 (5%); GLm-2 (5%)

8: LVh-1 (60%); LVg-1 (15%); CMe-1 (15%); ARb-1 (5%); GLm-1 (5%)

9: LVh-2 (50%); LVg-2 (15%); LPq-1 (10%); LPd-1 (10%); ARb-1 (5%); GLd-2 (5%); CMd-2 (5%)

10: LVg-2 (25%); LVh-2 (20%); ALh-2 (15%); PDd-2 (15%); CMe-2 (10%); GLd-2 (10%); PZh-1 (5%)

11: LVh-1 (25%); LVg-1 (20%); ALh-1 (15%); PDd-1 (15%); CMe-1 (10%); GLd-1 (5%); ARb-1 (5%); PZh-1 (5%)

12: ALh-1 (25%); PDd-1 (25%); CMd-1 (20%); CMg-1 (10%); GLd-1 (10%); PZh-1 (10%)

13: PZh-1 (75%); PZg-1 (10%); ARb-1 (10%); HSs (5%)

14: PZh-1 (50%); ARb-1 (25%); PZg-1 (10%); ARg-1 (5%); GLd-1 (5%); HSs (5%)

15: ARb-1 (80%); PZh-1 (15%); GLe-1 (5%)

16: ARa-1 (90%); ARg (10%)

17: ARb-1 (50%); PZh-1 (15%); GLd-1 (15%); CMd-1 (10%); PDd-1 (5%); HSs (5%)

Fig. 2. Soil mapping units defined in the conventional soil map of Denmark compiled according to the FAO–Unesco revised legend 1990.
Source: Madsen and Jensen (1996).

a leave-one-out cross-validation approach, where soil group at each site was predicted using $n - 1$ observations and compared with the observed group at the same site. Two confusion matrices, each for training and validation data sets, were constructed and three validation indices, namely User Accuracy (UA), Producer Accuracy (PA) and Overall Accuracy (OA), were calculated according to Taghizadeh-Mehrjardi et al. (2012) on a point-by-point basis (Eqs. (1)–(3)). UA is the probability that describes how a predicted soil group matches that being observed, whereas PA suggests how well the observed soil group was predicted by the model. The OA is the mean of correctly classified soil groups. The same indices were used to evaluate the accuracy of the model to predict similar soil groups (SSG). For each of the soil groups observed, the corresponding SSG were identified based on general pedogenetic similarities. This allowed user accuracy including similar soil groups (UASSG), producer accuracy including similar soil groups (PASSG) and overall accuracy including similar soil groups (OASSG) to be calculated for each predicted soil group.

$$UA_j = \frac{X_{ii}}{\sum_{i=1}^{C} X_{ij}} \tag{1}$$

$$PA_j = \frac{X_{jj}}{\sum_{i=1}^{C} X_{ij}} \tag{2}$$

$$OA = \frac{\sum_{i=1}^{C} E_{ii}}{N} \tag{3}$$

where X_{ii} is the diagonal value for each class in one row, X_{jj} is the diagonal value for each class in one column, X_{ij} is the sum of values in one row or column, E_{ii} is the sum of diagonal elements, N is the number of observations and C is the number of soil groups predicted.

2.11. Consistency assessment using the existing FAO soil map

To check the overall quality and consistency of the predicted map, it was also compared with the conventional FAO soil map of Denmark, which was compiled during the early 1990s (Fig. 2). This map consists of 17 soil mapping units (SMU) where the area covered by the soil groups in each SMU is estimated and expressed as a percentage of the total area of that SMU. Before comparison, the paper sheet map (the only available format of the old map) was digitised and projected to the same coordinate system as the new map. The area covered by the predicted soil groups in each of the 17 SMUs was calculated and compared with the estimated area of the corresponding groups from the old map.

3. Results

3.1. Distribution of FAO Soil groups in the profile data

The soil profiles consisted around 60% of the national monitoring grid and the rest were from the gas pipeline transect. The grid profiles were well distributed throughout the study area, whereas the transect profiles only represented a narrow strip of it, although high density profile data were recorded from there. Instead of a simple division of a whole set of data into two parts, an attempt was made to individually divide each soil group into two sub-parts (i.e. 80% and 20% each) and merge the corresponding sub-parts together into training and validation data sets. This division ensured a good representation of the soil groups in terms of geography and feature space. Table 2 outlines the proportionate division of the data, where each group makes an equal contribution to the training and validation data sets used in this study.

Table 2
FAO soil groups in the Danish soil profiles.

FAO soil group	Whole data set	Training data set	Validation data set
Alisols	35	28	7
Arenosols	140	112	28
Cambisols	150	120	30
Luvisols	368	294	74
Podzols	326	261	65
Fluvisols	24	19	5
Gleysols	105	84	21
Podzoluvisols	23	18	5
Total	1171	936	235

The highest number of profiles in the study area was classified as Luvisols (368 profiles) followed by Podzols and Cambisols, whereas the Fluvisols and the Podzoluvisols were the least observed soil groups. The majority of the soil groups in the central and the eastern part of the study area were Luvisols and Cambisols, whereas the western part was predominantly Podzols. Arenosols were distributed throughout the study area, the majority along the coastline. Gleysols, on the other hand, were mostly confined to the north, while Fluvisols were found along the south-west coast and also towards the north. Although 29 Histosols were also observed, they were not considered in the study because a recent better resolution peat map of Denmark was used as a mask for the areas with Histosols. Similarly, very few Rendzinas and Leptosols were present and they were excluded from the mapping, together with Anthrosols as they were not representative.

3.2. Decision tree modelling

The classifier generated a large tree model using different combinations of environmental variables to predict soil groups in the study area. The model was run for 10 consecutive trials which constructed 10 separate decision trees, that were combined in a boosting approach. The first tree (at trial zero), which is identical to the tree that could be produced without boosting, consisted of 152 leaves to which the soil groups were predicted. Of 936 cases studied, this trial mis-classified 194, giving a prediction error of about 21%. When the tree was 'boosted', the error was reduced to nearly 4%. But when the tree was run without clay content from the top and subsoil depths, this error increased to 8%. Table 3 lists the classification error associated with each of the trials, along with the number of tree leaves where soil groups were predicted.

The classifier generated a large tree model but as an example, one of several branches of the constructed tree is described. In the first run (trial zero), topsoil clay content was used as a first variable to initiate branching, followed by elevation, which was again partitioned based on the landscape type, geo-regions and geology of the study area. The first split identified the area with topsoil clay content ≤7.6% or higher. Where the topsoil clay content was lower, elevation with a cutoff value of ≤11.2 m was used for the next split. The last split to this branch predicted Podzols, where subsoil clay content was less than 6.8%. Of 936

Table 3
Classification error associated with each trial in the prediction of soil groups in the study area (total number of cases 936).

Number of trials	Number of leaves	Classification error (%)
0	152	20.7
1	155	27.0
2	181	26.2
3	177	26.9
4	198	26.0
5	193	25.5
6	193	26.0
7	192	26.0
8	188	29.0
9	174	26.9
Boost	–	3.8

total cases, 298 were mapped to this leaf, where 91 cases were misclassified. Areas with subsoil clay content >6.8% were further divided based on landscape type and geo-regions. The last node in this branch considered subsoil clay content (>14%) again and predicted two Luvisols where the slope gradient was less than 1.7°.

3.3. Identifying the important variables

Not all the variables used by the decision tree model were equally important in predicting soil groups. Variables such as clay content and geology were used in all the splits, while variables such as the mid-slope position and slope aspect were used less frequently. Clay content from both topsoil and subsoil depths and geology were among the top three predictors, which had a RI of 100%. This indicated that the 'soil' and 'parent material' factors in *scorpan* are the most important predictors. Meanwhile relief factors such as MRVBF and slope aspect were among the variables which had a lower importance of less than 50% RI. Table 4 lists the variables used by the model with their corresponding RI values. Among the most frequently used top five predictors, clay content from topsoil and subsoil depths and elevation were the only continuous variables, while the other two were categorical or class variables. Similarly, for the model without clay information included, geology and landscape types appeared to be the most influencing variables (100% RI for both) suggesting a high importance of parent material and geomorphology to predict soil groups in Denmark. This result also suggested that incorporation of prior pedological knowledge as covariates is important.

3.4. Predicted map in the spatial context

The map showing the spatial distribution of different FAO soil groups predicted across Denmark is presented in Fig. 3. The majority of the soils in central and eastern Denmark were predicted as Luvisols. Cambisols were scattered all over the study area as small patches where Luvisols were present, but a higher concentration of Cambisols was noticed towards the south-east corner. A significant amount of Arenosols was mapped towards the north, where they were present side by side with Podzols. A few Arenosols were also noticed along the western coast. Fluvisols, on the other hand, were mainly present towards the south-west of Denmark, whereas most of the Gleysols covered a relatively larger area in the north. Podzols covered a huge area in the west that was clearly distinguished from the rest of the study area. A few small islands of Podzoluvisols in the main body of Podzols in the west and a channel-like pattern of Arenosols, especially towards the south-west and the border between eastern and western Denmark, were also noticed. Although Alisols were mapped in northern Denmark, they covered only a small portion of land. The island to the extreme east (Bornholm) was also found to be rich in Luvisols, surrounding a central block of Arenosols.

Table 4
Relative importance of environmental variables used by the decision tree model to predict FAO soil groups in the study area.

Environmental variable	Scorpan factors	Relative importance (%)
Clay 0–30[a]	S	100
Clay 60–100[a]	S	100
Geology	P	100
Land use	O	92
Elevation	R	88
Landscape types	R	85
Geo-regions	R/C	81
Direct insolation	C	73
Flow accumulation	R	65
Slope gradient	R	64
Valley depth	R	62
SAGA WI	R	57
Mid-slope position	R	54
Distance to channel network	N	53
Wetlands	P	53
MRVBF	R	50
Slope aspect	R	47

SAGA WI, system for automated geoscientific analyses wetness index; MRVBF, multi-resolution index of valley bottom flatness.

[a] Clay 0–30, topsoil clay content; Clay 60–100, subsoil clay content.

The area coverage calculated for each of the predicted soil groups in Denmark is shown in Table 5. Luvisols covered a maximum area of about 35% followed by Podzols, which shared another 32% of the total predicted area. These two soil groups appeared to cover more than two-thirds of Denmark. Podzoluvisols, on the other hand, occupied the least area (1.7%), whereas Alisols and Fluvisols covered a similar extent of >2%. Based on the new peat map of Denmark, Histosols occupied about 2.5% of the total area.

3.5. Evaluation of model performance

3.5.1. Validation with the training and validation data

The performance of the tree model to predict different soil groups is summarised in Tables 7 and 8 for the training and validation data sets, respectively. Of the 28 Alisols in the training data, 24 were predicted to the same group as observed (UA 86%) and the rest as Luvisols and Podzols. The two predicted Alisols were apparently classified as Cambisols and Luvisols. Due to this prediction error, Alisols had 92% PA. Podzoluvisols showed the highest PA, 100% but the UA was found to be 94%, as one of the observed Podzoluvisols was misclassified as a Podzol. Some of the Cambisols and Alisols, on the other hand, were incorrectly classified, whereas all Fluvisols were correctly classified (UA 100%). Considering all errors, the overall prediction accuracy of the training data was 96%. However, the model performance based on the validation data (OA = 60%) was comparatively lower than that of the training data. The highest UA was again observed for the Podzols (80%), whereas the lowest was for Fluvisols (0%). Of five Fluvisols observed in the validation data set, none was accurately predicted by the model. Similarly, among five predicted Alisols, only one was correctly classified as observed (PA 20%). On the other hand, the classifier without clay data and no a priori soil information included in mapping, model performance based on OA reduced to 92% for training data and to 51% for validation data.

This validation also provided an opportunity to evaluate the model performance in predicting similar or related soil groups. Soils with a comparable pedogenetic process of profile development were categorised as similar soil groups. As an example, Alisols and Luvisols were put in one group because both these soils have a higher subsoil clay content and argic subsoil horizon, together with a cation exchange capacity (CEC) of more than 25 cmol$^+$/kg clay. However, Luvisols had a high BS (Base Saturation) compared with Alisols. Similarly, for Cambisols, a considerably similar group could be the Luvisols, although the former lack a well-developed argic horizon. Moreover, both soil groups were derived from Phaeozems. The suggested similar soil groups for each of the predicted FAO groups from the study area are listed in Table 6.

The prediction accuracy of the FAO soil groups including SSG is shown in Table 8. It suggested that the overall accuracy increased from 60% to 76% when including the SSG. Although Fluvisols had a UA of 0%, considering Gleysols as a SSG of Fluvisols, the UASSG increased to 40%. Similarly, the PA of Podzoluvisols increased from 20% to 80% when Podzols were included as a SSG of Podzoluvisols.

3.5.2. Comparison with existing FAO soil map for consistency

The results of the comparison between the soil groups in two different maps (i.e. existing and newly predicted maps) are shown in Table 9. The calculated area covered by the predicted soil groups in most of the SMUs was comparable to the corresponding estimated areas occupied by the same soil groups in the existing map. For example, SMU 7 had an area coverage of about 9405 km^2, where Luvisols were assumed to

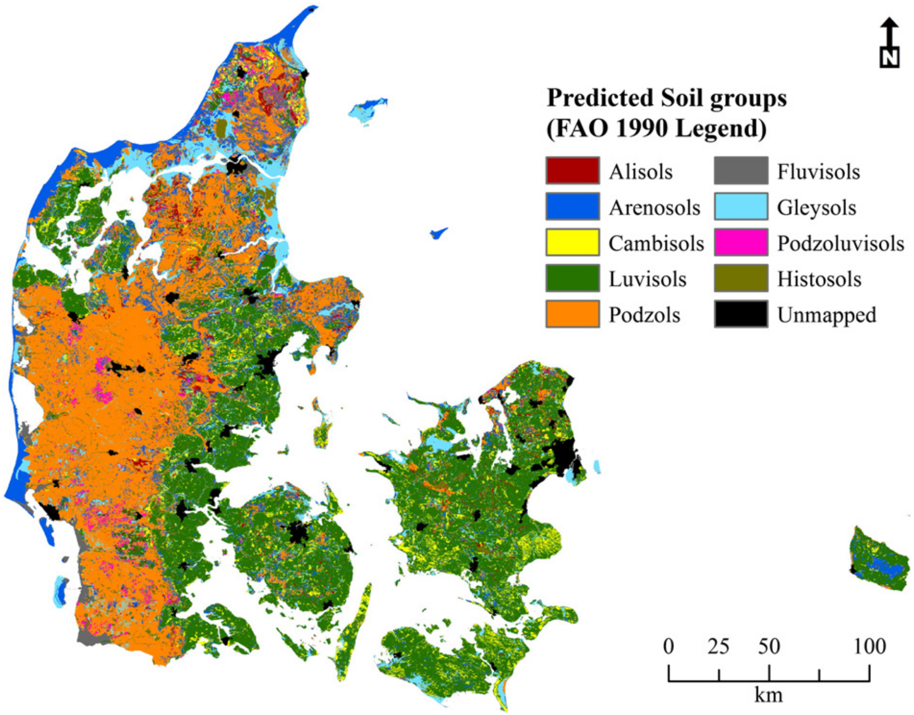


Fig. 3. FAO soil groups in the study area as predicted by boosted-decision tree modelling.

occupy 75%, Cambisols 15%, Arenosols and Gleysols both covering 5% area each. For that SMU, the calculated area covered by Luvisols was 70%, Cambisols 12%, Arenosols 5% and Gleysols 4%, suggesting a good match between the two maps. However, some of the groups, e.g. Arenosols from SMU 15, were poorly predicted. The predicted area was only 17%, whereas the map showed 80% Arenosols for that SMU. Similarly, for Podzoluvisols and Alisols the area predicted by the model was much lower than the estimated area on the existing map. However, the Podzols and Gleysols from SMUs 13 and 14, Cambisols from SMUs 11 and 9 and Arenosols from SMUs 4 and 6 were well predicted and comparable (refer to Table 9 for the detailed comparison). Arenosols in SMU 16 were found to cover about 76% where the conventional map suggested 100% coverage. Although Leptosols were reported to cover 20% area of SMU 9 in the existing map, our prediction was 0% because it was not predicted.

3.6. Mapping prediction uncertainty

The map of prediction uncertainty (Fig. 4) shows the confidence of the tree model in predicting soil groups in the study area. The confidence values ranged between 0.2 (least confident) and 1 (most confident). Pixel-by-pixel evaluation of the predicted maps of the entire study area showed that Podzols were predicted with the highest confidence, with a mean value of about 0.72, whereas Podzoluvisols and Alisols were predicted with the lowest confidence (0.48). The model confidence was about 0.64 when predicting Fluvisols and 0.70 when predicting Luvisols. But for the classifier that excluded clay content data as predictors and the output not adjusted pedologically, average confidences were lower for all predicted soil groups except Cambisols and Podzoluvisols which had slightly higher values. With clay maps used as predictors, Luvisols exhibited a higher confidence of 0.70 but when no such maps were used, the value reduced to 0.66. Table 10 lists the corresponding mean values of prediction confidence with their standard deviation calculated for each soil group predicted throughout the study area. Similarly, model confidence for each of the 17 SMUs was also investigated. It was found that SMU 9 which had an estimated area of 65% Luvisols, had the highest confidence (0.77), while SMU 17 had the lowest (0.55). SMU 17 represented 50% Arenosols and 15% Podzols and Gleysols each. SMU 2, which was predominantly Fluvisols (100% Fluvisols) was also predicted with a relatively higher confidence (0.71). A similar confidence (0.73) was also observed for the SMU 13 which was assumed to have 85% Podzols in it. The remaining SMUs had an intermediate value of confidence (Table 9).

Fig. 5 displays the distribution of prediction confidence for correctly classified and misclassified soil groups in validation locations. It was observed that misclassifications were less frequent when model confidence increased. The average confidence between these two classification groups was also found significantly different (p-value = 0.0009; $\alpha = 0.05$).

Table 5
Area covered by the predicted soil groups in Denmark.

FAO soil group	Area in km^2	% of total predicted area
Alisols	920.86	2.22
Arenosols	3581.90	10.62
Cambisols	2901.27	6.97
Fluvisols	878.67	2.09
Gleysols	3303.92	7.70
Histosols	1030.59	2.48
Luvisols	14,492.44	34.85
Podzols	13,731.60	31.40
Podzoluvisols	693.34	1.67

Table 6
Similar soil groups from the study area.

FAO soil group	Similar soil groups
Alisols	Luvisols
Arenosols	Podzols
Cambisols	Luvisols
Luvisols	Alisols, cambisols
Podzols	Podzoluvisols, arenosols
Fluvisols	Gleysols
Gleysols	Fluvisols
Podzoluvisols	Podzols

Table 7
Confusion matrix for training data (N[a] = 936).

Observed soil group	Predicted soil group									
	Alisols	Arenosols	Cambisols	Luvisols	Podzols	Fluvisols	Gleysols	Podzoluvisols	Row total	UA[a]
Alisols	24	–	–	3	1	–	–	–	28	86
Arenosols	–	105	–	–	7	–	–	–	112	94
Cambisols	1	–	110	3	3	1	2	–	120	92
Luvisols	1	–	1	291	–	–	1	–	294	99
Podzols	–	1	–	–	259	1	–	–	261	99
Fluvisols	–	–	–	–	–	19	–	–	19	100
Gleysols	–	–	–	1	2	1	80	–	84	95
Podzoluvisols	–	–	–	–	1	–	–	17	18	94
Column total	26	106	111	298	273	22	83	17	–	–
PA[a]	92	99	99	98	95	86	96	100		OA[a]–96

[a] N, the number of profiles; UA, user accuracy; PA, producer accuracy, OA, overall accuracy as percentages.

4. Discussion

4.1. Prediction model

The decision tree model was able to partition the environmental variables based on their quantified relationship with FAO soil groups. The importance of variables selected during the prediction showed their level of influence in the model when mapping soil groups in the study area. In most cases, the model split started with the soil clay content. As the amount of clay and its translocation in the profile are key parameters when classifying soil groups in the field, this was also reflected in our prediction, which found the highest RI (100%) for topsoil and subsoil clay maps. Information on geology was also highly considered by the model (RI 100%). As geology determines the type of parent material on which the soil develops, the higher influence of this factor during the prediction of soil groups was expected. Thus 'soil' and 'parent material' are the most important predictors for soil group maps. Arenosols and Podzols, which were predominantly found along coastal areas and in glacial-flood plains, were among the soil groups that developed on sandy parent material, whereas Luvisols and Cambisols were developed on a glacial basal till in the moraine landscapes covering a major part of central and eastern Denmark. Land use also played an important role during the split (RI 92%). In areas with sandy parent material, most of the soils under heath vegetation and coniferous or mixed forests were predicted as Podzols. The soils from the post-glacial marine landscapes developed at a lower elevation were predominantly Gleysols. Similarly, a large number of profiles from agricultural land in the kettled moraine landscape were predicted as Luvisols.

The boosting approach applied in the prediction model proved to minimise classification error by considering the possible source of errors in each trial and subsequently allocating the corresponding weight to each of the trees while combining them together. This also managed to reduce over-fitting (Lacoste et al., 2011) which is a common case in most tree-based models.

4.2. Pedological significance of the environmental variables

Some of the predicting variables were able to clearly distinguish the soil groups from each other in the study area (Fig. 6). Subsoil clay content separated Podzols and Arenosols from the rest of the soils because the former groups have less clay in their subsoil depths. On the other hand, Luvisols have a higher clay content in the subsoil due to clay illuviation (Bt horizon) and the model was able to identify this using the subsoil clay content map. Similarly, elevation managed to isolate soil groups that are normally formed in similar elevation settings. Examples include Gleysols and Fluvisols, which are normally expected in low slopes or flat areas. Soils developed under coniferous forest or heath were predominantly Podzols, because the environmental conditions created by such land use types are favourable for the development of Podzols. Arenosols from the aeolian deposits, Podzols from the sandy glacial-flood plains, Fluvisols from the marsh areas and Gleysols mostly from the post-glacial marine deposits also indicated a strong pedological significance of landscape type in classifying different soil groups in Denmark. These results were very much in line with the suggestions made by different

Table 8
Confusion matrix for test data (N[†] = 235).

Observed soil group	Predicted soil group										
	Alisols	Arenosols	Cambisols	Luvisols	Podzols	Fluvisols	Gleysols	Podzoluvisols	Row total	UA[†]	UASSG[†]
Alisols	1	–	1	4	1	–	–	–	7	15	71
Arenosols	1	8	4	4	6	3	2	–	28	29	50
Cambisols	–	2	10	14	4	–	–	–	30	34	80
Luvisols	2	2	8	58	–	–	3	1	74	79	89
Podzols	1	3	2	1	52	–	3	3	65	80	89
Fluvisols	–	–	1	1	1	0	2	–	5	0	40
Gleysols	–	1	2	6	2	–	10	–	21	48	48
Podzoluvisols	–	1	–	1	2	–	–	1	5	20	60
Column total	5	17	28	89	68	3	20	5	–	–	–
PA[†]	20	47	36	66	77	0	50	20		OA[†]– 60	
PASSG[†]	60	65	64	80	88	0	60	80			OASSG[†]– 76

[†] N, number of soil profiles; UA, user accuracy; PA, producer accuracy, OA, overall accuracy, expressed as percentages; UASSG, PASSG and OASSG for the user accuracy, prediction accuracy and overall accuracy including the similar soil groups expressed as percentages. Shaded values are reserved for the similar soil groups.

Table 9
Soil mapping unit (SMU) based comparison of the area covered by different soil groups in the existing and newly predicted soil map.

SMU[a] code	SMU[a] area (km^2)	Soil group	% of SMU area covered		Prediction confidence for the SMU	
			Existing map	Predicted map	Mean	Std. dev.
1	274.13	Histosols	100	38	0.58	0.14
2	296.71	Fluvisols	100	73	0.71	0.16
3	80.73	Fluvisols	100	55	0.63	0.17
4	2769.80	Gleysols	90	50	0.64	0.18
		Arenosols	10	13		
5	2278.40	Cambisols	70	15	0.69	0.16
		Luvisols	20	73		
		Arenosols	5	4		
		Gleysols	5	4		
6	169.36	Cambisols	70	9	0.68	0.18
		Luvisols	20	77		
		Arenosols	5	3		
		Gleysols	5	5		
7	9405.35	Luvisols	75	70	0.68	0.17
		Cambisols	15	12		
		Arenosols	5	5		
		Gleysols	5	4		
8	2487.30	Luvisols	75	62	0.63	0.17
		Cambisols	15	11		
		Arenosols	5	7		
		Gleysols	5	5		
9	60.62	Luvisols	65	72	0.77	0.19
		Leptosols	20	0		
		Arenosols	5	10		
		Gleysols	5	1		
		Cambisols	5	5		
10	2541.78	Luvisols	45	58	0.62	0.18
		Alisols	15	2		
		Podzoluvisols	15	1		
		Cambisols	10	11		
		Gleysols	10	6		
		Podzols	5	8		
11	1050.05	Luvisols	45	41	0.59	0.17
		Alisols	15	3		
		Podzoluvisols	15	1		
		Cambisols	10	10		
		Gleysols	5	7		
		Arenosols	5	13		
		Podzols	5	16		
12	1055.57	Alisols	25	2	0.64	0.17
		Podzoluvisols	25	9		
		Cambisols	30	3		
		Gleysols	10	4		
		Podzols	10	60		
13	9001.25	Podzols	85	78	0.73	0.19
		Arenosols	10	6		
		Histosols	5	3		
14	5818.61	Podzols	60	53	0.60	0.18
		Arenosols	30	14		
		Gleysols	5	4		
		Histosols	5	4		
15	2233.93	Arenosols	80	17	0.63	0.18
		Podzols	15	28		
		Gleysols	5	5		
16	1165.90	Arenosols	100	76	0.66	0.17
17	936.64	Arenosols	50	25	0.55	0.17
		Podzols	15	30		
		Gleysols	15	7		
		Cambisols	10	14		
		Podzoluvisols	5	9		
		Histosols	5	4		

[a] SMU, Soil Mapping Units. The SMU codes follow the definition by Madsen and Jensen (1996).

authors while studying soil and landscape developments in Denmark (Jacobsen, 1984; Madsen and Jensen, 1996; Schou, 1949).

This study also verified the importance of GlobalSoilMap output (used as clay content maps from 0 to 30 cm and from 60 to 100 cm soil depth) while mapping soil groups in Denmark. It also showed the added value of a priori pedological knowledge and soil information that improved the quality and reliability of the predicted map. Without this information used, overall prediction accuracy was decreased from 60% to 51% in the validation locations.

4.3. Quality of the predicted map

The reliability of the predicted soil map can also be evaluated considering validation, prediction uncertainty and map comparison results.

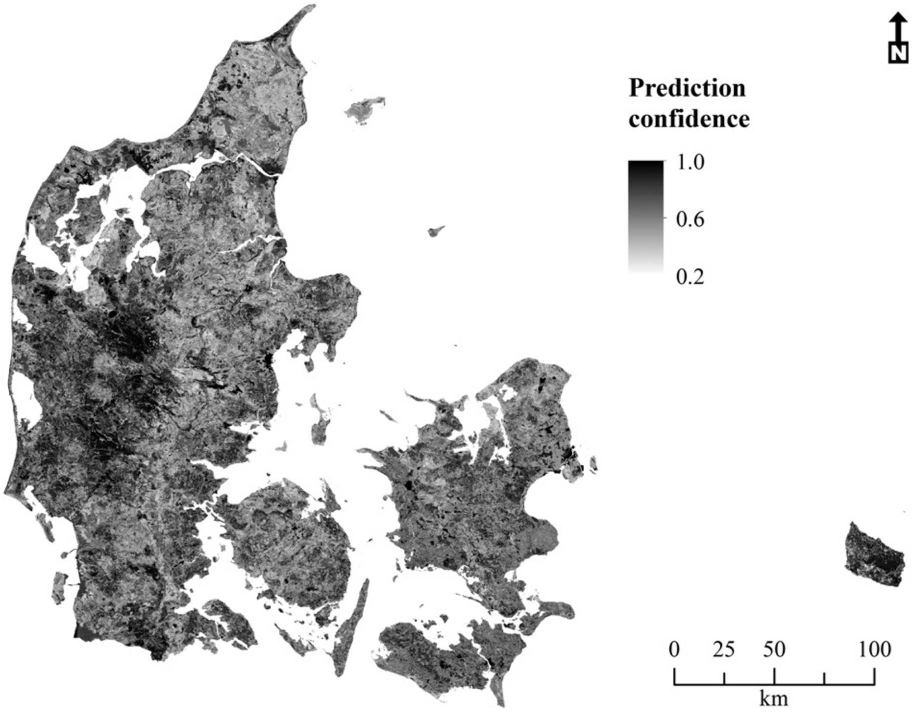


Fig. 4. Prediction uncertainty map for FAO soil groups in Denmark.

The approximately 60% overall accuracy of our prediction is comparable to that reported by Grinand et al. (2008) and Lacoste et al. (2011), who predicted soil parent material and landscape types using similar mapping principles. The predictive nature of the model was also found to be interesting as most of the mis-classifications affected the similar soil groups. Including such similar soil groups in the assessment suggested increased overall accuracy (76%), which indicates a reliable prediction not crossing similar soil group boundaries. Moreover, the results of the map comparison seemed promising, as some of the soil groups and SMU were predicted very well with great confidence. The average prediction uncertainty of less than 34% (66% confidence) while predicting all soil groups from the entire study area was also very convincing. Although some SMUs were weakly predicted, we still trust our results for two main reasons: 1) we calculated the uncertainty of each prediction, which helped assess the reliability and acceptance; and 2) we compared our new product to one for which mapping quality has not been reported but still the results were in a good agreement in most cases. However, in the prediction in some minor areas such as on Læsø island where no Luvisols can be expected, our mapping suggested a thin strip of it. It might be an error introduced due to the subsoil clay map which predicted some amount of clay that was enough for the model to assign Luvisols in this area.

Table 10
Prediction confidence associated with different soil groups.

Predicted soil group	Prediction confidence			
	Classifier 1[a]		Classifier 2[a]	
	Mean	SD[a]	Mean	SD[a]
Alisols	0.48	0.10	0.48	0.09
Arenosols	0.58	0.17	0.56	0.13
Cambisols	0.55	0.14	0.57	0.13
Fluvisols	0.64	0.19	0.61	0.17
Gleysols	0.63	0.17	0.61	0.19
Luvisols	0.70	0.17	0.66	0.17
Podzols	0.72	0.18	0.69	0.18
Podzoluvisols	0.48	0.13	0.49	0.13

[a] Classifier 1, model that used soil clay content from top and subsoil depths as predictors and pedological adjustments applied; Classifier 2, clay data not used as predictors, and pedological adjustments not applied; SD, Standard deviation.

5. Conclusions

This study made a national level prediction of soil groups from the FAO–Unesco Revised Legend in Denmark, using the information derived from point soil observations and environmental data as predictors. The relationship between FAO soil groups and the predictors was derived by applying a boosted decision tree-based DSM model. A considerable

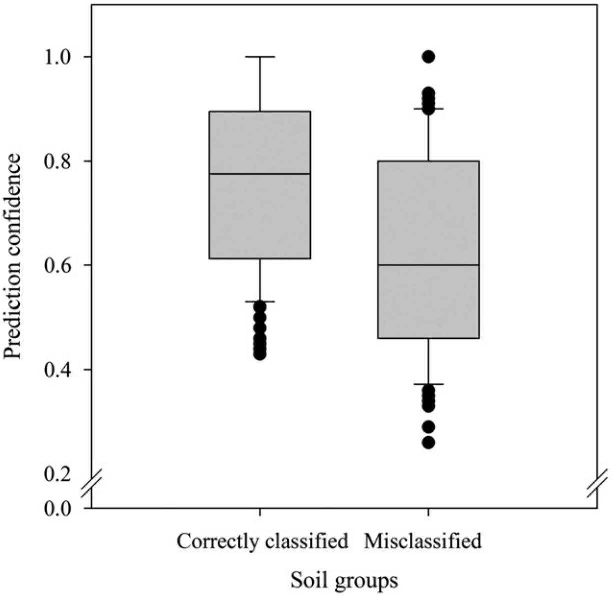


Fig. 5. Prediction confidence of correctly classified and misclassified soil groups derived for validation profiles.

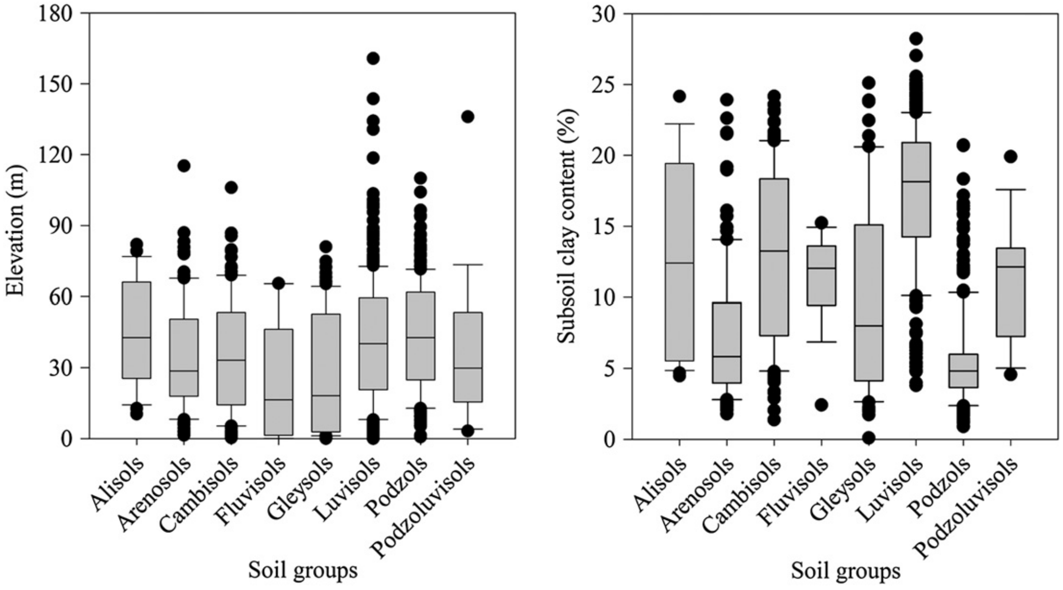


Fig. 6. Relationship of different soil groups with elevation (left) and clay content in the 60–100 cm soil layer (right).

reduction in classification error was obtained after the trees were boosted. The overall prediction accuracy based on 20% validation data was increased from 60% to 76% when the prediction accuracy of similar soil groups was considered. For the whole prediction, average prediction confidence of about 66% seems to be a convincing result. Noticeable similarities when comparing predicted soil groups with the same groups from an existing soil map also increased the reliability of our predictions. Clay content in topsoil and subsoil layers, geology, landscape type and elevation were found to be the most important drivers influencing the spatial distribution of FAO soil groups in Denmark. In spite of the fact that some predictions were rather weak, the results can be considered reliable as we showed the mapping uncertainty, based on which end-users can decide whether to or not to use the map. Thus the predicted soil group map, which is at a high resolution (30.4 m spacing), can now be used in various soil management and environmental studies in Denmark, including soil erosion, mapping soil properties, and as support in soil and land use policy development decisions. It can also be used as a reference map for future soil mapping activities in Denmark and beyond. This study also concludes that prediction performance and soil map quality can be improved by using GlobalSoilMap products and a priori pedological information.

In the future, the similarity (or taxonomic distance) between the soil classes needs to be considered when making the model (Minasny and McBratney, 2007). Error propagation from the covariate data will also be considered during the prediction process when we apply the similar model to predict soil classes based on the universal classification system, for example — World Reference Base in Denmark. As a reference to other nations lacking enough legacy soil information and GlobalSoilMap products, other methods such as map disaggregation or homosoil need to be explored (Minasny and McBratney, 2010). Freely available environmental data (e.g., DEM from the Shuttle Radar Topography Mission and Landsat images) can be incorporated to help the mapping process.

Acknowledgments

The study was supported by the SINKS Project (2009–2012) with funding from the Danish Ministry of Climate and Energy. The authors kindly thank the editor and two anonymous reviewers for their constructive comments on the manuscript. We also thank Henrik Breuning-Madsen, Niels H. Jensen and Thomas Balstrøm for their valuable comments and suggestions. The assistance of Mary McAfee with language editing is also gratefully acknowledged.

References

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-resolution 3-D mapping of soil texture in Denmark. Soil Sci. Soc. Am. J. 77, 860–876.

Bendix, J., 2004. Gelandeklimatologie. Gebruder Borntraeger, Berlin.

Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S., 2010. Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer-Verlag, Dordrecht, the Netherlands.

Böhner, J., Antonić, O., 2009. Land surface parameters specific to topo-climatology. In: Hengl, T., Reuter, H.I. (Eds.), Geomorphometry: Concepts, Software, Applications. Elsevier, New York, pp. 195–226.

Böhner, J., Köthe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalization by means of terrain analysis and process parameterization. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), Soil Classification 2001. Eur. Soil Bur., Res. Rep. No. 7, EUR 20398 EN, Luxembourg, pp. 213–222.

Bou Kheir, R., Greve, M.H., Bøcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: the case study of Denmark. J. Environ. Manag. 91 (5), 1150–1160.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Pacific Grove, CA.

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103 (1–2), 79–94.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111 (1–2), 21–44.

Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil–landscape rules from previous soil surveys. Aust. J. Soil Res. 37 (3), 495–508.

Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Model. 191 (3–4), 431–446.

Carre, F., Girard, M.C., 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110 (3–4), 241–263.

Carre, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: beyond DSM. Geoderma 142 (1–2), 69–79.

Commission of the European Communities, 1985. Soil map of the European Communities, 1:1,000,000, Office for official publications of the European Communities, Luxembourg.

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

Danmarks Meteorologiske Institut, 1998. Danmarks Klima 1997. Danmarks Meteorologiske Institut, Copenhagen.

Danner, A., Mølhave, T., Yi, K., Agarwal, P.K., Arge, L., Mitasova, H., 2007. TerraStream: from elevation data to watershed hierarchies. 212–219.

Dobos, E., Hengl, T., 2009. Soil mapping applications. In: Reuter, H.I., Hengl, T. (Eds.), Geomorphometry: Concepts, Software and Applications. Elsevier, New York, pp. 461–479.

ESRI, 2012. ArcGIS Desktop: Release 10.1. Environmental Systems Research Institute, Redlands, CA.

FAO, 1977. Guidelines for Soil Profile Description. FAO, Rome, Italy.

FAO–Unesco, 1974. Soil Map of the World, Legend. FAO, Rome, Italy.

FAO–Unesco, 1990. Soil Map of the World, Revised Legend. FAO, Rome, Italy.

Finke, P.A., 2012. On digital soil assessment with models and the pedometrics agenda. Geoderma 171, 3–15.

Freeman, T.G., 1991. Calculating catchment-area with divergent flow based on a regular grid. Comput. Geosci. 17 (3), 413–422.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381.

Gallant, J.C., Dowling, T.I., 2003. A multi-resolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12), 1347–1359.

Greve, M.H., Mount, H., Hudson, B., Breuning-Madsen, H., 2001. History of land value assessment and establishment of benchmark soils in Denmark. Soil Surv. Horiz. 42 (1), 19–23.

Greve, M.H., Greve, M.B., Bocher, P.K., Balstrom, T., Breuning-Madsen, H., Krogh, L., 2007. Generating a Danish raster-based topsoil property map combining choropleth maps and point information. Dan. J. Geogr. 107 (2), 1–12.

Greve, M.H., Kheir, R.B., Greve, M.B., Bocher, P.K., 2012a. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: the case study of Denmark. Ecol. Indic. 18, 1–10.

Greve, M.H., Kheir, R.B., Greve, M.B., Bøcher, P.K., 2012b. Using digital elevation models as an environmental predictor for soil clay contents. Soil Sci. Soc. Am. J. 76 (6), 2116–2127.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1–2), 180–190.

Grunwald, S., 2006. Environmental Soil–Landscape Modeling: Geographic Information Technologies and Pedometrics. In: Grunwald, S. (Ed.), CRC Press, New York.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152 (3–4), 195–207.

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Sci. Soc. Am. J. 75 (4), 1201–1213.

Hartemink, A.E., McBratney, A.B., Mendonca-Santos, M.L. (Eds.), 2008. Digital Soil Mapping with Limited Data. Springer-Verlag, Dordrecht, the Netherlands.

Hartemink, A.E., Hempel, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., MacMillan, R.A., Minasny, B., Montanarella, L., Mendonça Santos, M.L., Sanchez, P., Walsh, M., Ghang, G.L., 2010. GlobalSoilMap. net—a new digital soil map of the world. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer Science, Dordrecht, pp. 423–428.

Hewitt, A., 1993. Predictive modelling in soil survey. Soils Fertil. 56 (3), 305–314.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56 (3), 836–841.

Jacobsen, N.K., 1984. Soil map of Denmark according to the FAO–UNESCO legend. Dan. J. Geogr. 84, 93–98.

Jafari, A., Finke, P.A., Van de Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA—great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63 (2), 284–298.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G., de Vries, F., 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Sci. Soc. Am. J. 76 (6), 2097–2115.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133 (1–2), 90–99.

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geogr. Inf. Sci. 11 (2), 183–198.

Lagacherie, P., McBratney, A.B., Voltz, M., 2007. Digital Soil Mapping: An Introductory Perspective, 31. Elsevier, Amsterdam, The Netherlands.

Lemercier, B., Lacoste, M., Loum, M., Walter, C., 2012. Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171–172, 75–84.

Madsen, H.B., Jensen, N.H., 1985. The establishment of pedological soil databases in Denmark. Dan. J. Geogr. 85, 1–8.

Madsen, H.B., Jensen, N.H., 1992. Pedological regional variations in well-drained soils, Denmark. Dan. J. Geogr. 92, 61–69.

Madsen, H.B., Jensen, N.H., 1995. The Elaboration of a Revised EU Soil Map of Denmark. European Land Information Systems for Agro-environmental Monitoring. Joint Research Center, European Commission.

Madsen, H.B., Jensen, N.H., 1996. Soil map of Denmark according to the revised FAO legend 1990. Dan. J. Geogr. 96, 51–59.

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

McBratney, A.B., Santos, M.L.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89 (1–2), 67–94.

McSweeney, K., Gessler, P.E., Slater, B.K., Hammer, R.D., Bell, J.C., Petersen, G.W., 1991. Towards a new framework for modeling the soil–landscape continuum. Factors of soil formation. Proc. symposium, Denver, pp. 127–145.

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142 (3–4), 285–293.

Minasny, B., McBratney, A.B., 2010. Methodologies for global soil mapping. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Applications, and Operation. Progress in Soil Science. Springer, London.

Minasny, B., McBratney, A.B., Lark, R.M., 2008. Digital soil mapping technologies for countries with sparse data infrastructures. In: Hartemink, A.E., McBratney, A.B., Mendonca-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer, Australia, pp. 15–30.

Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), 2012. Digital Soil Assessments and Beyond. CRC Press/Balkema, Leiden, the Netherlands.

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Adv. Agron. 118, 1–47.

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16 (6), 533–549.

Østergaard, H.S., 1990. Kvadratnettet for nitratundersøgelser i Denmark 1986–89. Landbrugets Rådgivningscenter, Landskontoret for Planteavl, Skejby, Århus.

Patil, N., Lathi, R., Chitre, V., 2012. Comparison of C5. 0 & CART classification algorithms using pruning technique. Int. J. Eng. Res. Technol. 1 (4), 1–5.

Quinlan, J.R., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, USA.

SAGA GIS, S. System for automated geoscientific analyses http://www.saga-gis.org.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital soil map of the world. Science 325 (5941), 680–681.

Schou, A., 1949. Atlas of Denmark I: The Landscapes. The Royal Danish Geographical Society, Copenhagen.

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

Stjernholm, M., Kjeldgaard, A., 2004. CORINE Landcover Update in Denmark—Final Report. National Environment Research Institute (NERI), Denmark.

Taghizadeh-Mehrjardi, R., Minasny, B., McBratney, A.B., Triantafilis, J., Sarmadian, F., Toomanian, N., 2012. Digital soil mapping of soil classes using decision trees in central Iran. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessment and Beyond. CRC, London, pp. 197–202.

Zhu, A.X., Band, L., Vertessy, R., Dutton, B., 1997. Derivation of soil properties using a soil land inference model (SoLIM). Soil Sci. Soc. Am. J. 61 (2), 523–533.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65 (5), 1463–1472.

Zhu, X., Wang, J., Yan, H., Wu, S., 2009. Research and application of the improved algorithm C4.5 on Decision tree. International Conference on Test and, Measurement, pp. 184–187.