Geoderma 341 (2019) 148–160

Contents lists available at ScienceDirect

Geoderma

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

Improved disaggregation of conventional soil maps

Anders Bjørn Møller[a,*], Brendan Malone[b], Nathan P. Odgers[b,c], Amélie Beucher[a], Bo Vangsø Iversen[a], Mogens Humlekrog Greve[a], Budiman Minasny[b]

[a] *Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark*
[b] *Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW 2015, Australia*
[c] *Soils and Landscapes Team, Manaaki Whenua – Landcare Research, PO, Box 69040, Lincoln 7640, New Zealand*

ARTICLE INFO

Keywords:
Decision trees
Denmark
Landscapes
Models
Soil group
Wetlands

ABSTRACT

The disaggregation of conventional soil maps is an alternative for producing high-quality soil maps when point observations are not available. Previous studies developed the DSMART algorithm ("Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees") for this purpose. In this study, we tested the sensitivity of DSMART towards the input data by using two different conventional soil maps covering Denmark at scales of 1:1,000,000 and 1:2,000,000. As a potential way to improve the algorithm, we tested an implementation of soil-landscape relationships, using maps of wetlands and soil texture. We also tested two different sampling schemes, generating either a set number of virtual samples per polygon in the input map or a number of virtual samples in proportion to the areas of the polygons. Thirdly, we tested the replacement of the resampling procedure and decision tree model with Random Forest. The original procedure repeated the generation of the virtual samples 50 times, fitting a decision tree in each repetition. We modified it by sampling only once and fitting a Random Forest model. The area-proportional sampling scheme and soil-landscape relationships both improved the accuracy. Random Forest yielded a lower accuracy than the original resampling and decision tree procedure, but was far more computationally efficient. The accuracy also depended strongly on the input maps. In the best case, the algorithm predicted soil types with 18% accuracy and soil groups with 47% accuracy. The results demonstrated that there are several ways to improve the disaggregation of conventional soil maps, and that a suitable approach can provide reliable soil maps at a national extent.

1. Introduction

Several innovative agricultural practices and measures to protect the environment take place at local levels, which has reinforced the demand for high-quality, high-resolution maps of soil properties (Kovacic et al., 2000; Auernhammer, 2001). Researchers have developed several approaches to produce accurate soil maps from point observations (McBratney et al., 2003; Scull et al., 2003). Unfortunately, the fieldwork to acquire soil observations is expensive and time consuming, and it may not be a practical option across large areas.

However, in many areas, conventional soil maps exist, and these may serve as an alternative source of input data (Arrouays et al., 2017). Soil surveyors usually produce these maps using soil observations in combination with expert knowledge, following a tacit mental model (Hudson, 1992; Bui, 2004). Several studies have sampled conventional soil maps with the aim to produce new maps with a higher level of detail (Cialella et al., 1997; Scull et al., 2005; Giasson et al., 2011).

Situations when the polygons of the map contain more than one soil type require a more complex approach. When soil observations are available, they can be used to disaggregate the polygons (Schmidt et al., 2008), and if the survey report includes soil-landscape relationships or soil toposequences, they can also be used for this purpose (Bui and Moran, 2001; Nauman and Thompson, 2014).

Odgers et al. (2014) proposed a different approach in the form of the DSMART algorithm ("Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees"). The algorithm works by generating a set number of virtual samples within each of the polygons in the input map and assigns soil types to the samples according to the proportions of the soil types in the polygons, which can be gathered from the associated soil reports. The algorithm then trains a single decision tree model from the virtual samples and geographic data layers of environmental variables and uses it to produce a map of soil types. The algorithm repeats the process for a specified number of times, generating new virtual samples in every repetition. Afterwards, the algorithm summarizes the results, and produces maps of the most probable soil types, second-most probable soil-types etc. and their

[*] Corresponding author.
E-mail address: anbm@agro.au.dk (A.B. Møller).

https://doi.org/10.1016/j.geoderma.2019.01.038
Received 14 May 2018; Received in revised form 14 January 2019; Accepted 14 January 2019

associated probabilities.

The study which presented DSMART covered a 68,000 km^2 area of in Queensland, Australia (Odgers et al., 2014). Holmes et al. (2015) tested the method in Western Australia, combining input maps from several surveys. The study showed that the results were most accurate for the most common soil types and in areas where the soil distribution was comparatively homogenous. Furthermore, the results were poor in areas with detailed original surveys, as the smaller surveys differed in purposes and methods.

Chaney et al. (2016) used DSMART to disaggregate the SSURGO map (Soil Survey Staff, 2016) into a soil series map of the conterminous United States at a 30 m resolution, by dividing the area into 12,474 subareas. Chaney et al. (2016) replaced the default C5.0 decision trees (Quinlan, 1993) with Random Forest models (Breiman, 2001), but did not quantify the effect of this alteration. Furthermore, Chaney et al. (2016) only sampled SSURGO once for each area and used the bootstrap procedure of the Random Forest algorithm to generate resampled decision trees for predictions.

Vincent et al. (2016) successfully implemented soil-landscape relationships from expert knowledge into the algorithm by specifying rules for the assignment of soil types to the virtual samples. The authors mapped the soil types of Brittany, France with soil-landscape relationships implemented, but did not state if the change improved the accuracy. Furthermore, Vincent et al. (2016) suggested that area-proportional sampling would improve the results, as the sampling scheme of the original algorithm results in an uneven sampling density. They also stated that single decision trees are prone to overfitting and hypothesized that a more robust predictive model would improve the results.

In this study, we implement an area-proportional sampling scheme and Random Forest models into the algorithm. We test how both factors affect the accuracy of the output maps, with and without soil-landscape relationships. Furthermore, we test the algorithm's sensitivity towards the input maps by using two different conventional soil maps.

2. Methods

2.1. Study area

Denmark, located in northern Europe at 54–58° latitude and 8–15° longitude, has an area of approximately 43,000 km^2 (Fig. 1). The terrain is generally flat with a maximum elevation 171 m above sea level. The climate is temperate coastal with temperatures ranging from 1 °C in January to 17 °C in July. Precipitation varies from 700 mm per year in the eastern part of the country to 875 mm in the western part of the country (Wang, 2013). The geology varies from loamy Weichselian moraines in the eastern part of the country to sandy Saalian moraines and glacial outwash plains in the western part of the country. Raised seabeds are present in the northern part of country. Agriculture is the dominant land use (61%) followed by forests and natural vegetation (21%) and urban areas (13%) (Statistics Denmark, 2017).

2.2. Input maps

This study used two different conventional soil maps of Denmark as input data. Both maps used the FAO-Unesco, 1974 classification system (FAO-Unesco, 1974). The system consists of 26 soil groups divided into 106 soil types. In the system, each soil type has a two-part name: The last part indicates the soil group, and the first part specifies the soil type. For example, the soil types *Gleyic Acrisols* and *Orthic Acrisols* both belong to the soil group *Acrisols*.

Jacobsen (1984) made the first map at a 1:2,000,000 scale. The map contains 869 polygons, representing 14 map units and 23 soil types (Fig. 2A). The number of soil types in each map unit varied from 2 to 7, with a mean value of 4.4.

The Commission of European Communities (CEC, 1985) made the second map at a 1:1,000,000 scale. The map covered the member countries of the European Communities at the time by combining contributions from the member countries. In this study, we use the contribution for Denmark made by Professor K. Rasmussen. Despite its finer scale, the map contains only 323 polygons representing 11 map units with 18 soil types in total (Fig. 2B). The map units contained 3–5 soil types each, with a mean value of 4.1.

The maps stated the proportions of the soil types in the map units with the labels 'dominant soils' (50%–100%) 'associations' (20%–50%) and 'inclusions' (0%–20%). As the DSMART algorithm requires explicit percentages, we converted these classes into percentages by assigning weights to the soil types according to their labels. We assigned the weights 0.75 to dominant soils, 0.35 to associations and 0.10 to inclusions. Subsequently, we scaled the sum of weights for each map unit to 100% (Table 1, Table 2). Some map units listed entries that referred to entire soil groups rather than specific soil types. In these cases, we replaced the listed soil group with the most common soil type within the soil group, if the map unit listed other soil types in the same soil group. Therefore, the same soil type can appear more than once in a map unit. If the map unit did not list other soil types within the group, we split the share evenly amongst the soil types of the group.

2.3. Covariate layers

We included 42 covariate layers in the study: nine variables related to the soil or the parent material, 20 variables from a digital elevation model, and 11 satellite-derived images and layers of land use and precipitation (Table 3). The layers were the same as in Møller et al. (2018), who described their derivation. The original study resampled the layers to a common spatial resolution of 30.4 m × 30.4 m.

McBratney et al. (2003) developed the *scorpan* approach for describing the relationship between the soil and other spatially distributed variables. The approach treats soil classes or properties as a function of other soil information (S), the climatic properties of the environment (C), organisms, especially vegetation, but also human influences, (O), relief and topographic variables (R), the parent material or lithology (P), the age of the soil, i.e. the time factor, (A) and spatial position (N). The authors emphasized that the *scorpan* function was devised for quantitative description of soil-landscape relationships rather than inference about pedogenesis, unlike the *clortp* approach of Jenny (1941).

We elucidated the relationship between the covariates and the *scorpan* approach by listing the relevant *scorpan* factors for the covariates (Table 3). Most of the covariates were associated with the factors R, O and S, while a few related to the factors C, P and A. The horizontal distance to waterbodies was the only covariate that related to the spatial position, N.

2.4. Experiments

2.4.1. Soil-landscape relationships

We used the two input maps in original and modified forms. In the modified maps, we split and reshaped the original map units into new units to take into account soil-landscape relationships. We then assigned the relevant soil types to the new map units. Specifically, we split the map units according to the extent of wetlands and soil texture classes.

Firstly, specific soil types are present in wetlands, so we used a map of wetland areas (Kheir et al., 2010) (Fig. 3A). Secondly, both input maps assign the soil types of each map unit to a number of textural classes for the depth interval 0–30 cm. Jacobsen (1984) used the texture classes 'fine', 'medium' and 'coarse' as defined by FAO-Unesco (1974). The CEC (1985) subdivided the 'fine' and 'medium' texture classes, which increased the number of classes to five. We produced maps of these texture classes from soil texture maps of Denmark made by Adhikari et al. (2013). In the resulting map, fine-textured soils were rare, and a single subclass dominated the medium-textured soils. We

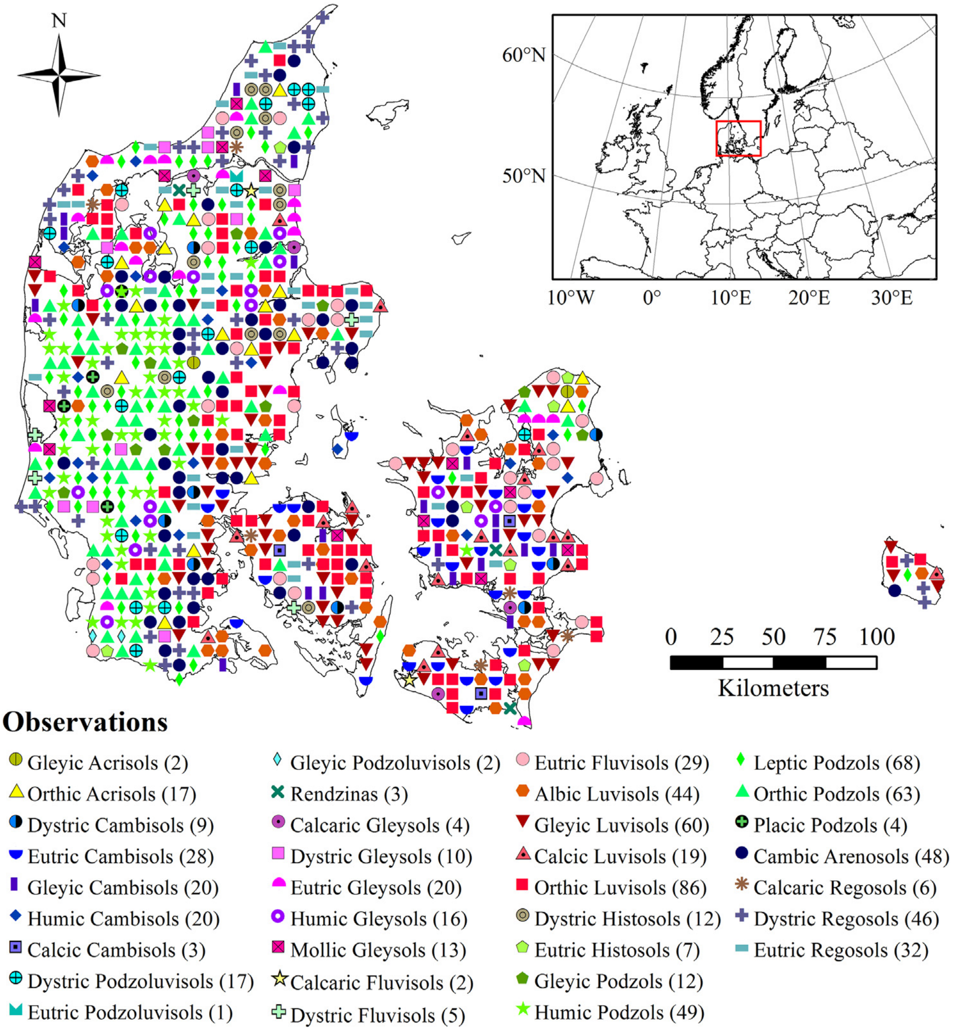

Fig. 1. Soil profile observations of the FAO-Unesco, 1974 soil types (FAO-Unesco, 1974) used for evaluating the maps produced by the disaggregation of conventional soil maps. Numbers in parentheses indicate the number of soil profiles for each soil type. The insert in the upper right corner shows the location of Denmark in Europe (red box). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

therefore grouped fine and medium-textured soils into one class. Therefore, the map only contained two texture classes (Fig. 3B). We maintained areas with peat, mapped by Kheir et al. (2010), as a separate class in both maps.

Vincent et al. (2016) assigned soil types to the training samples generated by the algorithm using explicit rules. However, in this study we aimed at maintaining the overall shares of the soil types from the original maps. Therefore, we used a different approach compared to Vincent et al. (2016).

If a map unit contained both wetland and non-wetland soils, or if the soil types in the map unit belonged to more than one textural class, we split the map unit into two or more new map units. If the soil types in a map unit were exclusively wetland or non-wetland soil types, or if the soil types belonged to only one textural class, we reshaped the map unit by constraining its extent to the corresponding areas.

We allocated soil types to the new map units by preference while maintaining the assigned shares. If the share of soil types of a specific wetland and texture class within a map unit was larger than the area matching the classes, we allocated the excess shares of the soil types to the other parts of the map unit. For example, if the shares of wetland soil types were too large to be accommodated in the areas with wetlands, the excess wetland soil types would be allocated to the non-wetland parts of the map unit. Peat areas were the only exception to this procedure, as we assumed that they were identical to the locations of Histosols. Therefore, we did not maintain the relative shares between Histosols and other soil groups. Instead, we assigned Histosols to areas with peat independently from their shares.

These modifications increased the number of polygons to 2278 and number of map units to 44 for the map by Jacobsen (1984). For the map by the CEC (1985), they increased the number of polygons to 585 and the number of map units to 27. At the same time, the number of soil types in the map units decreased to 1–6 (mean 2.3) for the map by Jacobsen (1984) and to 1–5 (mean 2.2) for the map by the CEC (1985). The modifications did not change the total numbers of soil types in the input maps.

2.4.2. Sampling scheme

The default sampling scheme generates the same number of virtual samples for all polygons. We tested a sampling scheme, in which the number of virtual samples was proportional to the area of the polygon in question. We refer to the default sampling scheme as per-polygon sampling and the new scheme as area-proportional sampling. We adjusted the number of virtual samples per polygon or unit area in each experiment, depending on the sampling scheme, so the number of virtual samples was slightly above 10,000 for each repetition. Afterwards, we reduced the sample size to 10,000, dropping cases for the most frequent soil types first. Therefore, all models used the same total number of virtual samples.

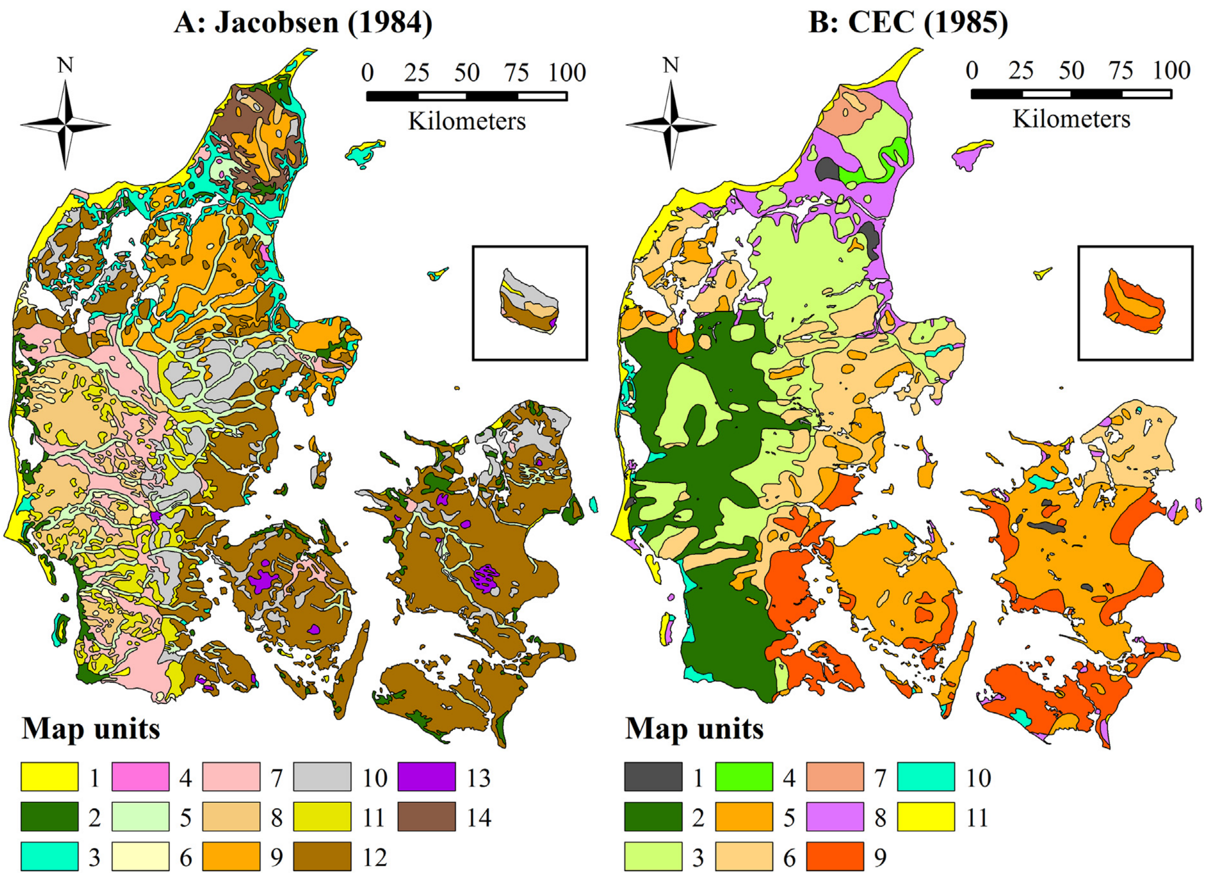

Fig. 2. The two input maps used in the study, produced by (A) Jacobsen (1984) and (B) the Commission of European Communities (CEC, 1985). Soil types for the map units of the two maps are listed in Table 1 and Table 2 respectively.

2.4.3. Resampling procedure and decision tree models

In its original design, DSMART generates new virtual samples in every repetition and trains a C5.0 decision tree (Quinlan, 1993) from the samples. To test the effect of changing the resampling procedure and decision tree algorithm, we replaced it with the Random Forest algorithm (Breiman, 2001), as implemented in the R package *ranger* (Wright and Ziegler, 2015). The Random Forest algorithm trains several decision trees by drawing bootstrap samples from the training dataset. The algorithm also implements randomness in the splitting process, as only a number of randomly selected covariates are available for each split. The parameter *mtry* sets the number of available covariates.

In this study, the individual Random Forest models contained 100 trees. For each forest, we tested ten values of *mtry* and two different splitting rules and selected the values that yielded the highest accuracy. We tested the accuracy by random 90/10% splits on the data repeated ten times for each Random Forest model. In the experiments using Random Forest, we sampled the input maps only once and used the soil type probabilities predicted by a single Random Forest in an approach similar to Chaney et al. (2016). In the experiments using C5.0, we used 50 sampling and prediction repetitions per experiment.

We tested all four input maps (original and modified) in combination with both sampling schemes and both predictive models. As a result, we carried out 16 experiments (Fig. 4).

2.5. Evaluation

Earlier studies using the DSMART algorithm used the term ‘validation’ when assessing the reliability of the output maps (Odgers et al., 2014; Chaney et al., 2016; Vincent et al., 2016). However, a model of a complex natural system such as the soil cannot truly be considered as validated (Oreskes, 1998). We therefore use the term ‘evaluation’.

We evaluated the generated maps using 777 soil profiles located in a 7 km grid (Fig. 1). These profiles were described in the years 1987–1990 using the FAO-Unesco, 1974 classification system (Madsen et al., 1992). Therefore, they are entirely independent of the input maps. The surveyors originally classified 179 profiles as soil types within the soil group Phaeozems. However, Phaeozems do not appear in the input maps. This is most likely because they do not form under natural conditions in Denmark. Instead, they form as a consequence of agricultural practices including tillage and additions of manure and lime (Madsen and Jensen, 1996). We therefore reclassified them as soil types within the soil groups Luvisols or Cambisols, depending on the presence of an argic B-horizon.

In each experiment, we calculated the predictive accuracy as the proportion of profiles correctly predicted in three ways. First, we calculated the accuracy of the output map with the most probable soil types, the ‘first soil type’. Second, we calculated the sum of the accuracy of the output maps using the three most probable soil types, named the ‘first three soil types’, summing the accuracy of the three output maps, following Odgers et al. (2014) and Vincent et al. (2016). Third, we calculated the accuracy of the soil groups in the output map of the most probable soil types, referred to as the ‘first soil group’. We also calculated the three accuracy measures on the unmodified input maps to serve as a baseline, using the same 777 soil profiles. In this operation, we used the shares of the soil types within the map units in place of probabilities. For example, we regarded the dominant soil type in each map unit as the most probable soil type. In cases where a map unit contained equal shares of several soil types, we calculated the attribution to each of the soil types and divided the results by the number of soil types that had equal shares.

To aid the interpretation of the results, we calculated the importance of the covariates in the models and averaged the results for

Table 1
Number of polygons, total area and soil types with associated percentages for the map units in the map by Jacobsen (1984).

Map unit	Number of polygons	Area (km^2)	Soil type	Area (%)
1	41	1091	Dystric Regosol	46.9
			Eutric Regosol	46.9
			Dystric Histosol	6.3
2	100	1634	Eutric Fluvisol	68.2
			Eutric Histosol	31.8
3	83	1987	Eutric Gleysol	45.5
			Dystric Histosol	21.2
			Dystric Fluvisol	21.2
			Eutric Fluvisol	6.1
			Mollic Gleysol	6.1
4	3	39	Dystric Histosol	50.0
			Eutric Histosol	50.0
5	65	3923	Dystric Histosol	25.9
			Eutric Histosol	25.9
			Dystric Fluvisol	12.1
			Eutric Fluvisol	12.1
			Eutric Gleysol	12.1
			Mollic Gleysol	12.1
6	55	523	Orthic Podzol	48.4
			Gleyic Podzol	22.6
			Humic Podzol	22.6
			Dystric Histosol	6.5
7	47	3734	Orthic Podzol	71.4
			Dystric Histosol	9.5
			Gleyic Podzol	9.5
			Placic Podzol	9.5
8	60	3656	Orthic Podzol	42.9
			Orthic Acrisol	20.0
			Humic Podzol	20.0
			Gleyic Luvisol	5.7
			Gleyic Podzol	5.7
			Dystric Histosol	5.7
9	52	3657	Orthic Podzol	31.3
			Luvic Arenosol	31.3
			Orthic Luvisol	14.6
			Albic Arenosol	14.6
			Gleyic Luvisol	4.2
			Dystric Histosol	4.2
10	101	3863	Cambic Arenosol	35.7
			Dystric Cambisol	16.7
			Orthic Luvisol	16.7
			Orthic Podzol	16.7
			Gleyic Cambisol	4.8
			Gleyic Luvisol	4.8
			Dystric Histosol	4.8
11	94	2280	Orthic Acrisol	45.5
			Dystric Cambisol	21.2
			Humic Cambisol	21.2
			Gleyic Acrisol	6.1
			Dystric Histosol	6.1
12	145	14,158	Orthic Luvisol	53.6
			Eutric Cambisol	25.0
			Orthic Podzol	7.1
			Gleyic Luvisol	7.1
			Eutric Histosol	7.1
13	14	336	Dystric Cambisol	51.7
			Gleyic Cambisol	24.1
			Eutric Histosol	24.1
14	5	950	Dystric Cambisol	51.7
			Eutric Gleysol	24.1
			Gleyic Cambisol	24.1

each experiment. The C5.0 algorithm calculates covariate importance as the number of cases included in splits using the covariate. On the other hand, the Random Forest algorithm calculates the change in the proportion of cases correctly predicted, when the covariate is perturbed. In both cases, the algorithm scales the importance to 100 for the most important covariate.

Table 2
Number of polygons, total area and soil types with associated percentages for the map units in the map by the CEC (1985).

Map unit	Number of polygons	Area (km^2)	Soil type	Area (%)
1	8	321	Dystric Histosol	62.5
			Eutric Histosol	29.2
			Humic Gleysol	8.3
2	2	7363	Humic Podzol	48.4
			Orthic Podzol	22.6
			Dystric Histosol	22.6
			Gleyic Podzol	6.5
3	31	7423	Orthic Podzol	45.5
			Humic Podzol	21.2
			Dystric Cambisol	21.2
			Gleyic Podzol	6.1
			Dystric Histosol	6.1
4	1	251	Orthic Podzol	53.6
			Humic Gleysol	25.0
			Eutric Cambisol	7.1
			Gleyic Podzol	7.1
			Humic Podzol	7.1
5	104	8873	Eutric Cambisol	62.5
			Orthic Luvisol	29.2
			Orthic Podzol	8.3
6	38	6916	Eutric Cambisol	62.5
			Orthic Podzol	29.2
			Eutric Histosol	8.3
7	1	428	Eutric Cambisol	45.5
			Calcic Cambisol	21.2
			Calcaric Gleysol	21.2
			Dystric Regosol	6.1
			Eutric Regosol	6.1
8	51	2593	Humic Gleysol	45.5
			Eutric Fluvisol	21.2
			Eutric Gleysol	21.2
			Dystric Histosol	6.1
			Humic Podzol	6.1
9	47	5626	Orthic Luvisol	45.5
			Eutric Cambisol	21.2
			Gleyic Luvisol	21.2
			Rendzina	6.1
			Eutric Regosol	6.1
10	21	537	Eutric Fluvisol	57.7
			Dystric Fluvisol	26.9
			Dystric Histosol	7.7
			Eutric Histosol	7.7
11	15	1221	Dystric Regosol	62.5
			Eutric Regosol	29.2
			Orthic Podzol	8.3

3. Results

3.1. Predictive accuracy

The map produced by Jacobsen (1984) (Fig. 2) had the highest baseline accuracy for the first soil type, while the map by the CEC (1985) (Fig. 2B) had the highest baseline accuracy for the first three soil types and the first soil group (Fig. 5). In the experiments which used the map by the CEC (1985), the accuracy of the output maps was often lower than the accuracy of the input map. On the other hand, this was only the case in two of the experiments, which used the map by Jacobsen (1984). The experiments that used the map by the CEC (1985) generally achieved a slightly higher accuracy for the first three soil types. On the other hand, the experiments that used the map by Jacobsen (1984) achieved a higher accuracy for the first soil type. They also achieved a higher general accuracy for the first soil group, although the input map had a lower baseline accuracy. In the experiments that used both soil-landscape relationships and area-proportional sampling, the accuracies were invariably higher than the baseline.

The accuracy of the first soil type and the first soil group were significantly positively correlated ($R = 0.75$, $n = 16$, $p < 0.05$), while the accuracy of the first three soil types was uncorrelated to the two

Table 3
Covariate layers used in the study, including their name, an explanation of the layer, the mean value and the range of values for numeric variables and the number of classes for categorical variables. The table also lists the scorpan factors related to the variables.

Covariate	Explanation	Mean (range)/number of classes	Scorpan factor
Soil and parent material			
clay_a	Clay content, 0–30 cm (%)	8.2 (0.0–51.2)	S
clay_b	Clay content, 30–60 cm (%)	10.1 (0.0–62.7)	S
clay_c	Clay content, 60–100 cm (%)	11.2 (0.0–59.1)	S
clay_d	Clay content, 100–200 cm (%)	10.9 (0.0–57.1)	S
dc	Soil drainage class from 1 (Very well-drained soils) to 5 (Very poorly drained soils)	2.9 (1–5)	S
geology	Scanned and registered geological map (Scale 1:25,000)	10 classes	P
georeg	Scanned geographical regions map (Scale 1:100,000)	7 classes	C, P, A
landscape	Landscape types (Scale 1:100,000)	12 classes	P, R, A
wetlands	Shows the presence of non-wetlands (0), wetlands (1), central wetlands (2) and peat (3) (Scale 1:20,000)	0.3 (0–3)	S, R
Topographic variables			
asp_cos	Cosine of the surface aspect	0.01 (−1.00–1.00)	R
asp_sin	Sine of the surface aspect	−0.03 (−1.00–1.00)	R
bluespot	Depth of sinks (m)	0.1 (0.0–92.5)	R
curv_plan	Plan curvature	0.0 (−5.1–6.0)	R
curv_prof	Profile curvature	0.0 (−7.3–6.1)	R
demdetrend	Elevation minus the mean elevation in a 4 km radius (m)	1.0 (−57.9–105.4)	R
dirinsola	Direct insolation (kWh/year)	1269 (122–1707)	C
elevation	Elevation above sea level (m)	30.9 (−39.5–170.5)	R
flowaccu	Number of upslope cells	60 (1–110,908)	R
gwd_intp	Depth to groundwater table interpolated from well observations and surface water	6.8 (0.0–144.3)	R
gwd_model	Depth to groundwater table from hydrological model	5.8 (0.0–126.0)	R
hdtochn	Horizontal distance to the nearest waterbody	231 (0–3238)	R
msp	Mid-slope position	0.27 (0.00–1.00)	C, R
mrvbf	Multi-resolution index of valley bottom flatness	4.3 (0.0–10.9)	R
sagawi	SAGA wetness index	14.5 (6.9–19.1)	R
slpdeg	Surface slope gradient (degrees)	1.6 (0.0–90.0)	R
slptochn	Downhill gradient to the nearest waterbody (degrees)	1.1 (0.0–52.6)	R
twi	Topographic wetness index; Calculated as TWI = ln(a/tan b): where a is flow accumulation, and b is local slope gradient	5.9 (−15.8–63.3)	R
valldepth	Valley depth (m)	7.5 (0.0–89.9)	R
vdtochn	Vertical distance to the nearest waterbody (m)	4.1 (0.0–115.4)	R
Satellite imagery			
LS8_band1	Landsat 8 Band 1 surface reflectance, March 2014 (Ultra blue)	356 (−676–15,471)	O
LS8_band2	Landsat 8 Band 2 surface reflectance, March 2014 (Blue)	421 (−407–15,769)	O
LS8_band3	Landsat 8 Band 3 surface reflectance, March 2014 (Green)	623 (−406–15,843)	O
LS8_band4	Landsat 8 Band 4 surface reflectance, March 2014 (Red)	665 (−673–16,000)	O
LS8_band5	Landsat 8 Band 5 surface reflectance, March 2014 (Near infrared)	2191 (−114–15,955)	O
LS8_band6	Landsat 8 Band 6 surface reflectance, March 2014 (Shortwave infrared 1)	1879 (−74–16,051)	S, O
LS8_band7	Landsat 8 Band 7 surface reflectance, March 2014 (Shortwave infrared 2)	1252 (−29–17,082)	S, O
ndmi	Normalized difference moisture index; (Band 5 − Band 6) / (Band 5 + Band 6)	0.08 (−1.00–1.00)	S, O
ndvi	Normalized difference vegetation index; (Band 5 − Band 4) / (Band 5 + Band 4)	0.52 (−1.00–1.00)	S, O
ndwi	Normalized difference water index (Band 5 − Band 3) / (Band 5 + Band 3)	−0.54 (−0.99–1.00)	O
savi	Soil-adjusted vegetation index; (Band 5 − Band 4) * (1 + 0.5) / (Band 5 + Band 4 + 0.5)	0.29 (−0.29–0.72)	O
Other layers			
lu	CORINE land cover data adopted in Denmark (Scale 1:100,000)	4 classes	O
precipitation	Mean annual precipitation in the period 1961–1990 interpolated from point data (mm)	708 (452–964)	C

other accuracy measures ($p > 0.05$).

The implementation of soil-landscape relationships increased the accuracy in nearly all cases. Likewise, area-proportional sampling generally increased the accuracy, especially for the first soil group. On the other hand, the use of Random Forest models generally decreased the accuracy, especially for the first soil type. For the first three soil types, there was no general difference in the accuracy achieved with the two decision tree algorithms.

The experiment with the highest accuracy for the first soil type (18%) used the map by Jacobsen (1984), soil-landscape relationships, *per*-polygon sampling and C5.0 decision trees (Fig. 6). On the other hand, the experiment with the highest accuracy for the first three soil types (36%) used the map by the CEC (1985), soil-landscape relationships, area-proportional sampling and a Random Forest model. Lastly, the experiment with the highest accuracy for the first soil group (47%) used the map by Jacobsen (1984), soil-landscape relationships, area-proportional sampling and C5.0 decision trees.

If we rank each measure of accuracy, the experiment that achieved the highest accuracy for the first soil type also had the highest mean rank. This experiment achieved the sixth highest accuracy for the first three soil types (33%) and the third highest accuracy for the first soil group (43%).

The supplementary materials present maps of the three most probable soil types produced in each experiment.

3.2. Covariate importance

The most important covariate was the map of geographical regions (*georeg*), followed by the map of landscape elements. The clay contents in all four depth intervals also had a high importance. *Precipitation*, *elevation*, *geology*, *mrvbf* and *wetland* were generally important as well (Table 4).

Most variables derived from the digital elevation model had an intermediate importance, and the land use map (*lu*) had an intermediate importance as well. The satellite-derived images all had a low importance. Some topographic variables also had a low importance, namely the sine and cosine of the aspect (*asp_sin*, *asp_cos*), the flow accumulation (*flowaccu*), the topographic wetness index (*twi*) and the

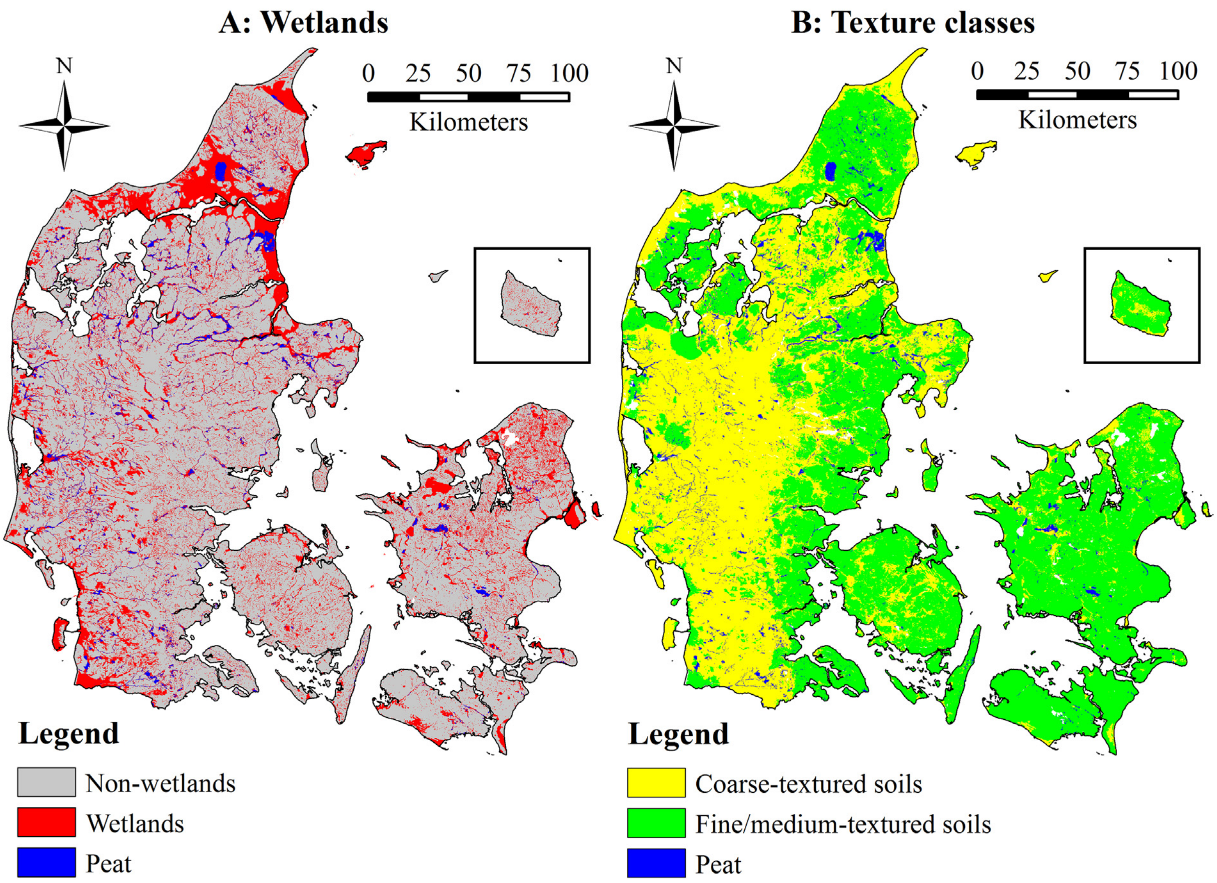

Fig. 3. The maps of (A) wetlands and (B) soil texture used for modifying the map units of the input maps.

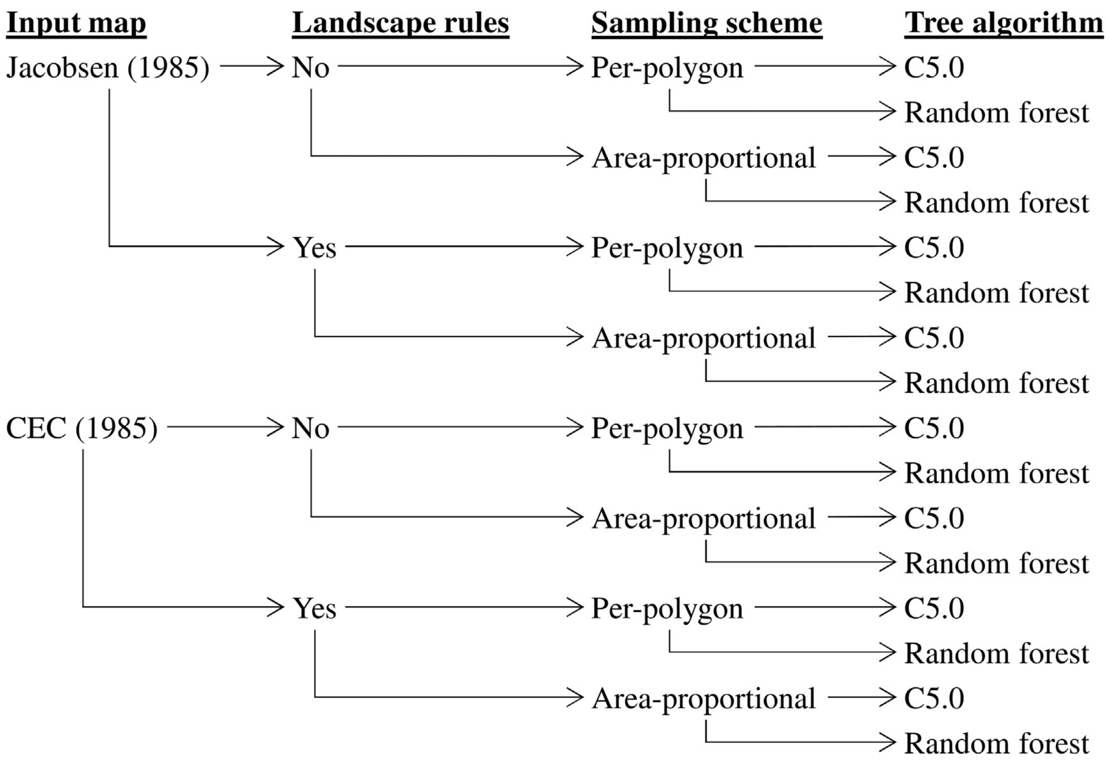

Fig. 4. Overview of the experiments carried out, including input map, implementation of soil-landscape relationships, sampling scheme and tree algorithm. The figure uses 'landscape rules' for the implementation of soil-landscape relationships, although we implemented them by modifying the map units.

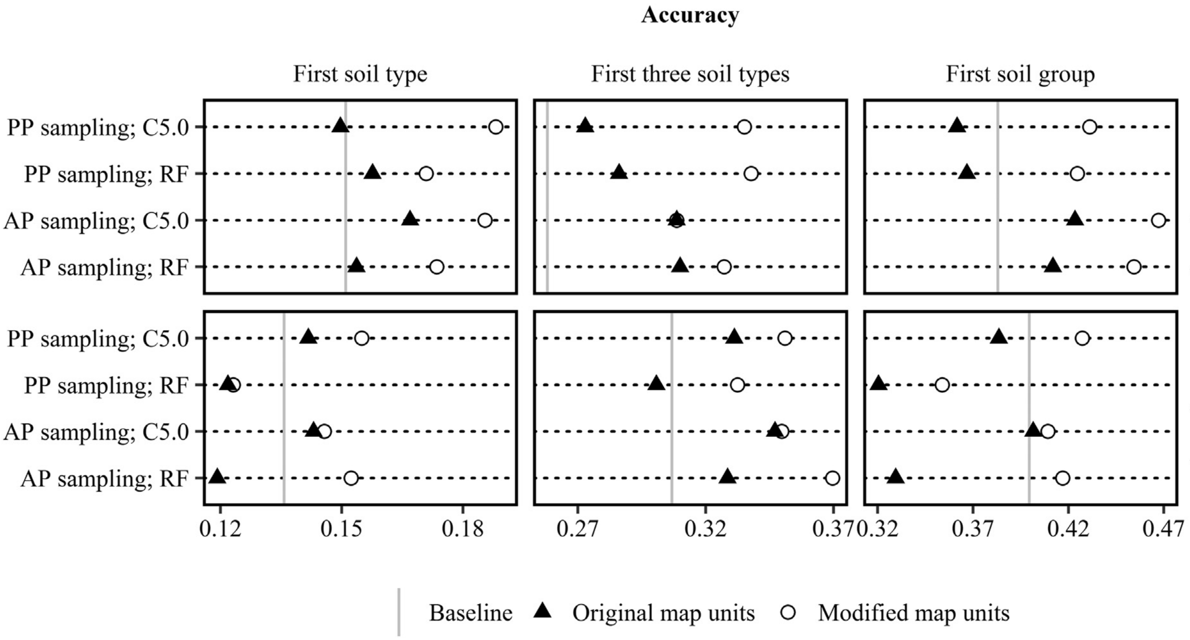

Fig. 5. Accuracy achieved in the experiments on disaggregating conventional soil maps. The boxes in the first row show the results obtained with the map by Jacobsen (1984), while the boxes in the second row show the results obtained with the map by the CEC (1985). The three columns of boxes show the accuracy measured on the first soil type, first three soil types and the first soil group. Triangles show the accuracy achieved with the original map units, while circles show the accuracy achieved with the map units modified to accommodate soil-landscape relationships. Vertical lines indicate the baseline accuracy of the input maps. Labels on the y-axes indicate experiments using per-polygon (PP) sampling, area-proportional (AP) sampling, C5.0 and Random Forest (RF) models, respectively.

plan and profile curvature (*curv_plan*, *curv_prof*).

Covariate importance was very similar for the maps by Jacobsen (1984) and the CEC (1985). The mean ranked covariate importance was highly correlated between the experiments using the two input maps ($R = 0.96$, $n = 42$, $p < 0.05$). There were no outliers to the correlation (Fig. 7A). Covariate importance was also similar with and without soil-landscape relationships, as the mean ranked covariate importance was highly correlated between the experiments ($R = 0.96$, $n = 42$, $p < 0.05$). The map of wetland areas was an outlier, as it was very important with soil-landscape relationships implemented but of intermediate importance without them (Fig. 7B). Covariate importance was also similar for per-polygon sampling and area-proportional sampling. The correlation in the mean ranked covariate importance was also high between the experiments using the two sampling schemes ($R = 0.94$, $n = 42$, $p < 0.05$). The mid-slope position (*msp*) was an outlier from the distribution, as its importance ranked 10 places higher on average with per-polygon sampling than with area-proportional sampling (Fig. 7C). Lastly, there was a lesser degree of similarity in the covariate importance of the C5.0 and the Random Forest models. Correlation between the mean ranked covariate importance in the two model types was moderate ($R = 0.65$, $n = 42$, $p < 0.05$). There were no outliers to the correlation (Fig. 7D).

4. Discussion

4.1. Predictive accuracy

The accuracy of the output maps depended heavily on the input maps, as the map by Jacobsen (1984) yielded higher accuracies for the first soil type and the first soil group, while the map by the CEC (1985) yielded higher accuracies for the first three soil types. Soil-landscape relationships increased accuracies in nearly all cases, and area-proportional sampling also generally increased the accuracy of the predictions. On the other hand, the use of Random Forest models generally decreased the accuracy.

In some experiments, the accuracy of the disaggregated outputs was lower than the accuracy of the input maps. However, the experiments that combined soil-landscape relationships and area-proportional sampling all achieved accuracies that were higher than the accuracies of the input maps.

It is difficult to compare the accuracies obtained in studies using DSMART, due to the diversity of approaches, sizes of the areas covered and the numbers of map units, polygons and soil types in the input maps (Table 5). However, it is interesting to note that the accuracy for the most probable soil types is in the range 17–23% in other studies, while in this study it was in the range 12–18% depending on the input map and the methods. This shows that the input map and the specific implementation of DSMART has a large effect on the accuracy of the outputs, even within the same area.

4.1.1. Input maps

The accuracy depended heavily on the input maps. This stresses the necessity to evaluate the input maps before disaggregation. In some cases, the disaggregated results had a lower accuracy than the input maps, which suggests that some input maps are not suitable for disaggregation or require careful consideration.

Both input maps had a coarse map scale, but the map by Jacobsen (1984) had a higher level of detail than the map by the CEC (1985) despite its coarser scale. The outputs from the map by Jacobsen (1984) nearly all had a higher accuracy than the input map. However, many of the outputs from the map by the CEC (1985) had a lower accuracy than the input map. This suggests that maps with a high level of detail (number of soil types, number of polygons, number of map units) are a better input for disaggregation than maps with a lower level of detail. Furthermore, the results suggest that the level of detail is more important than the nominal scale of the input maps. The most likely advantage of using input maps with a high level of detail is that it allows DSMART to detect more detail in the relationships between soil types and covariates.

These results contrast with the findings of Holmes et al. (2015) who found the lowest accuracies in areas with detailed soil maps. However, this was largely because Holmes et al. (2015) combined several soil

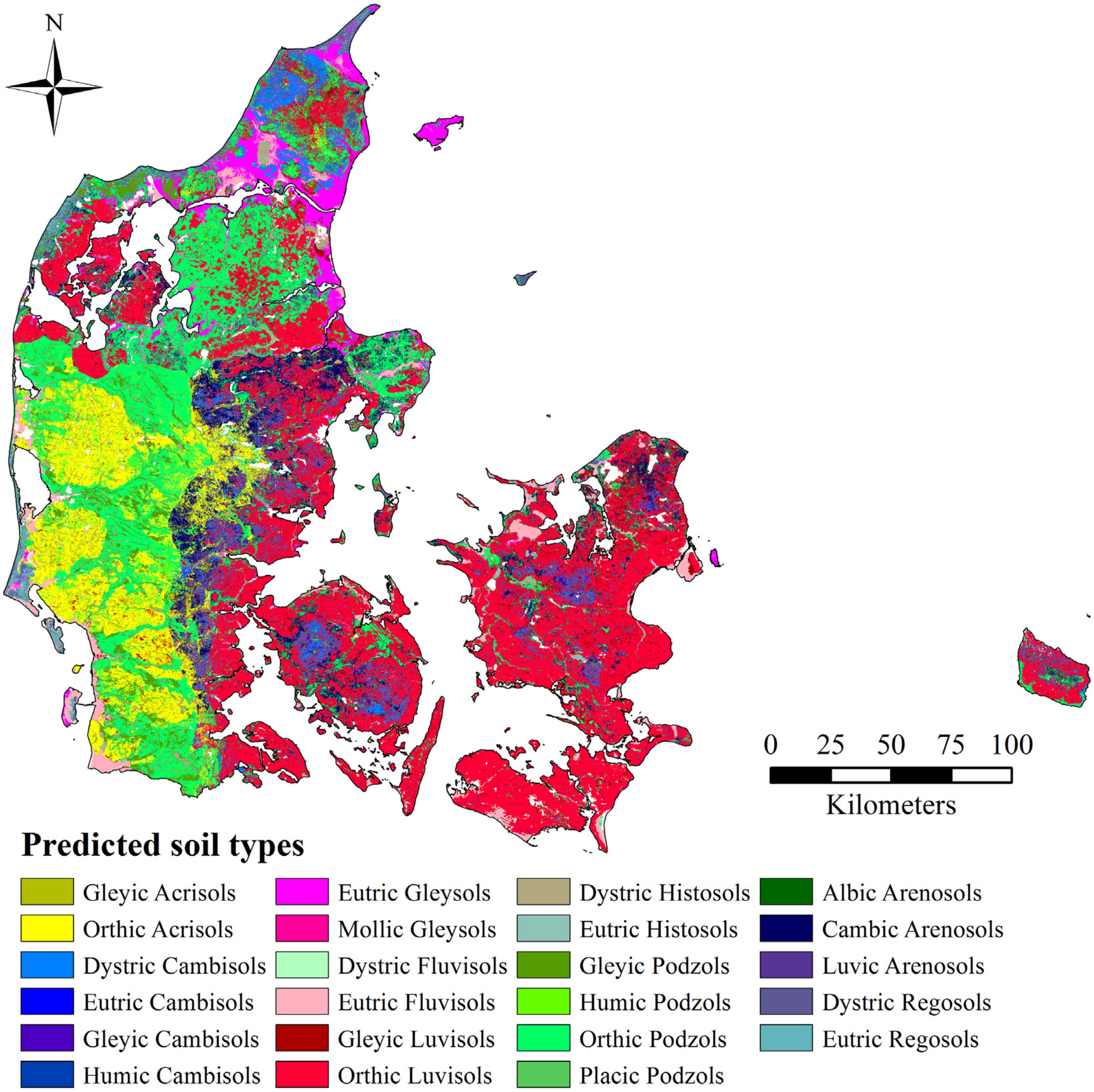

Fig. 6. Most probable soil types predicted in the experiment using the map produced by Jacobsen (1984), soil-landscape relationships, per-polygon sampling and C5.0 decision trees. This experiment had the highest mean rank for the three measures of accuracy. The supplementary materials show the second and third most probable soil types for the same experiment. The map has an accuracy of 18% for the first soil type, 33% for the first three soil types and 43% for the first soil group.

maps in these areas, as differences in survey method and intent prevented DSMART from finding relationships between soil types and covariates. The level of detail in itself was therefore not the cause of the lower accuracy. We will add that the input maps in our study had a very low degree of cartographic detail compared to other studies (Table 5). The advantage of using detailed input maps may therefore be relative, as an extremely high level of cartographic detail might lower the accuracy of the outputs. For example, an extremely large number of soil types would decrease the chance of predicting the correct soil type. Moreover, a highly detailed conventional soil map is likely to be very

Table 4
Importance of the covariates across experiments. We sorted the covariates by importance in each experiment and calculated the mean rank across the experiments.

Covariate	Mean rank	Covariate	Mean rank	Covariate	Mean rank	Covariate	Mean rank
georeg	1.8	lu	12.6	Bluespot	24.4	LS8_band3	32.7
Landscape	2.8	clay_d	13.1	Slptochn	24.4	asp_cos	33.1
clay_a	4.3	demdetrend	13.3	Dirinsola	26.5	LS8_band2	33.4
Precipitation	4.7	valldepth	14.8	LS8_band1	28.8	LS8_band7	34.1
Geology	5.0	sagawi	16.6	curv_prof	29.0	flowaccu	34.5
Wetland	5.9	gwd_intp	16.9	curv_plan	30.3	LS8_band4	35.4
Elevation	7.3	gwd_model	17.9	LS8_band5	30.9	ndvi	35.4
mrvbf	9.3	msp	21.6	asp_sin	31.3	ndwi	35.5
clay_b	10.2	hdtochn	22.9	LS8_band6	32.3	savi	35.8
clay_c	10.4	vdtochn	23.6	twi	32.4		
dc	11.8	slpdeg	24.0	ndmi	32.5		

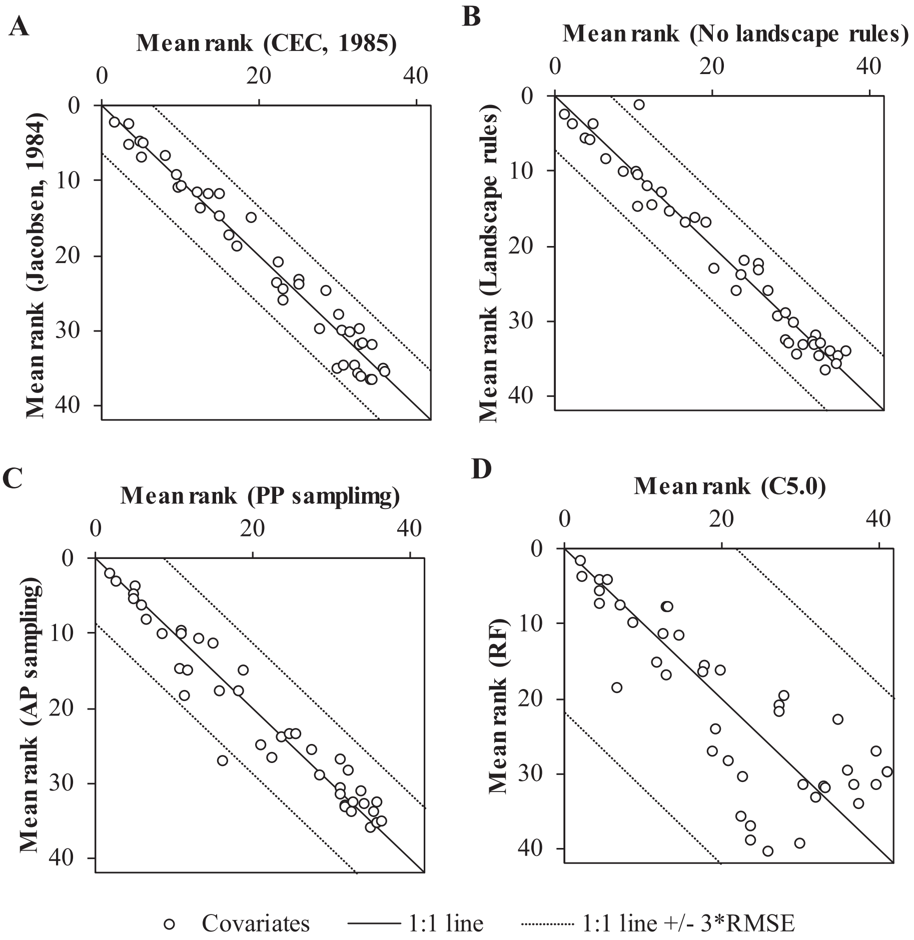

Fig. 7. Comparison of ranked covariate importance in different experiments. A: Mean covariate ranks in the experiments using the map by Jacobsen (1984) and the CEC (1985), respectively. B: Mean covariate ranks in the experiments with and without soil-landscape relationships, referred to as 'landscape rules', implemented. C: Mean covariate ranks in the experiments using per-polygon (PP) sampling and area-proportional (AP) sampling. D: Mean covariate ranks in the experiments using the C5.0 or the Random Forest (RF) algorithms. Solid lines marks the 1:1 line. Dotted lines mark the limits for outlier detection, the 1:1 line ± 3*RMSE of the correlation.

accurate, if it is the product of an intensive survey, and it will therefore be difficult to improve the accuracy of the map through disaggregation.

It is also possible that the representation of soil types influenced the accuracy achieved with the two input maps used in this study. In the map by Jacobsen (1984), Luvisols were the dominant soil group in eastern Denmark, while Cambisols were the dominant soil group in the map by the CEC (1985). The difference is not immediately explicable, but it may be due to a scarcity of observations, which forced the surveyors to rely on theoretical judgements. Regardless of the cause, it shows that the surveyors disagreed strongly on this issue.

In addition to the scale and the relative shares of soil types in the input maps, the two input maps also show different spatial structures (Fig. 2). The map by Jacobsen (1984) shows a much larger emphasis on hydrological networks, while the map by the CEC (1985) has a larger number of map units in the eastern part of the country. The differences in the accuracies achieved with the two input maps may therefore be a combination of several factors, including level of detail, shares of soil types and spatial structures.

The input maps used in this study were coarse and largely based on expert knowledge rather than observations, as no national-level investigations on soil types existed at the time (Greve and Madsen, 1999). These circumstances probably explain the modest accuracies achieved in this study. Adhikari et al. (2014) mapped the revised FAO-Unesco soil groups (FAO-Unesco, 1988) in Denmark based on 936 soil profiles and achieved an accuracy of 60%. In comparison, the disaggregated results in this study predicted soil groups with an accuracy up to 47%. This suggests that soil observations are a better input than coarse soil maps for digital soil mapping, when a sufficient number are available. However, it also shows that disaggregation is a viable alternative for producing soil maps in areas with few or no observations. This will, of

Table 5
Comparison between the sizes of the areas covered, numbers of polygons, map units and soil types in the input maps and the achieved accuracy reported in this and other studies using DSMART.

Study	Area (km^2)	Map units	Polygons	Soil types	Accuracy (%)
Odgers et al. (2014)	68,000	1,110	3,058	72	23
Holmes et al. (2015)	2,500,000	5,069	127,626	73	20–22
Chaney et al. (2016)	–	–	–	–	~17
Vincent et al. (2016)	27,040	341	~2,000[a]	320	20–23
This study	43,000	11–14	323–869	18–23	12–18

[a] Calculated based on numbers reported by the authors.

course, require particular attention to methods and input data. Some predictions in this study had lower accuracies than the input maps, and it is difficult to assess the accuracy of the predictions in areas with a low number of observations. A thoroughly tested approach is therefore the best option in these areas.

4.1.2. Soil-landscape relationships

Vincent et al. (2016) stated that the implementation of soil-landscape relationships would not necessarily increase the overall accuracy of the generated map. However, in this study it lead to a clear increase in accuracy in nearly all cases. Furthermore, while Vincent et al. (2016) used three geographic datasets (parent material, Topographic Position Index, and a waterlogging index) for landscape rules, the present study used only two datasets (wetlands and soil texture). This shows that soil-landscape relationships can improve the results even with relatively few but relevant geographic datasets.

4.1.3. Sampling scheme

Area-proportional sampling increased both the accuracy of the output maps and their agreement with the input maps. This confirms the expectations of Vincent et al. (2016). The cause of the improvement is clearly visible from areas and numbers of polygons for the map units. For example, in the map by the CEC (1985), Map Unit 2 has an area of 7,391 km^2 but consists of only two polygons, while Map Unit 5 has an area of 9,056 km^2 but consists of 106 polygons (Table 2). Therefore, with per-polygon sampling, DSMART would generate a disproportionately large number of virtual samples for Map Unit 5 relative to its area and vice-versa for Map Unit 2. With per-polygon sampling, the output maps contained very large areas of Eutric Cambisols, the dominant soil type of Map Unit 5, and nearly no Humic Podzols, the dominant soil type of Map Unit 2 (Fig. 8). With area-proportional sampling, the extent of these soil types corresponded more closely with the areas of the map units. In effect, the implementation of area-proportional sampling ensured that the output maps accorded more closely with the soil surveyors' intentions.

The experiment with the highest accuracy for the first soil type used per-polygon sampling. However, this finding goes against a general trend, as area-proportional sampling resulted in a higher accuracy than per-polygon sampling in most other cases. Furthermore, the difference between per-polygon sampling and area-proportional sampling was only 0.3% in this particular case.

4.1.4. Resampling procedure and decision tree models

The outputs generated with Random Forest models generally had a lower accuracy. The experiments using Random Forest had twice as many decision trees as the experiments using C5.0 (100 versus 50). However, in the experiments using C5.0, the algorithm generated 10,000 new virtual samples in each of the 50 repetitions. On the other hand, in the experiments using Random Forest, the algorithm generated 100 decision trees from bootstrap samples of the same 10,000 virtual samples. The experiments using C5.0 therefore sampled the input maps and the covariate layers more thoroughly, which would explain the higher accuracy achieved in the experiments using C5.0.

However, for the first three soil types, the difference in accuracy between the two decision tree methods was only 0.1% on average. Furthermore, an experiment using Random Forest achieved the highest accuracy for the first three soil types. Random Forest may therefore predict rare soil types more effectively, which may compensate for the

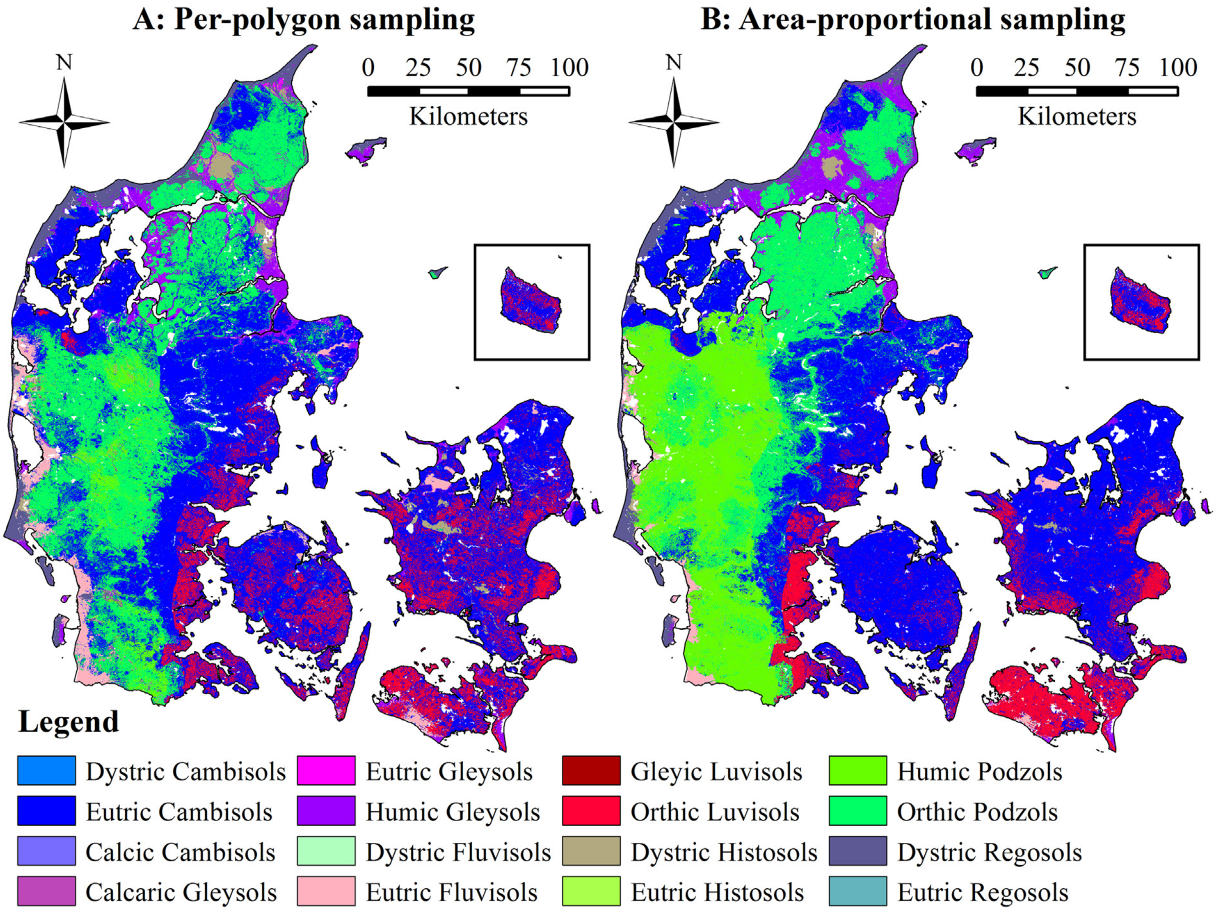

Fig. 8. Most probable soil types predicted by DSMART with per-polygon sampling and area-proportional sampling. In both cases the input was the map by the CEC (1985) without landscape rules in use, and the predictive models were C5.0 decision trees.

less intensive sampling. Per default, DSMART generates a map of soil types in every repetition. The algorithm then calculates the probabilities for each soil type by counting the number of times that it appears in a grid cell. On the other hand, in the experiments using Random Forest, the algorithm calculated soil type probabilities for each decision tree as the share of each soil type in the leaf nodes in the tree. It then averaged the probabilities across the decision trees. This procedure, also known as "probabilistic bagging", is therefore less likely to outvote rare soil types. The approach has increased the accuracy of decision tree predictions in other cases (Bauer and Kohavi, 1999).

The use of Random Forest with a single sampling was also more computationally efficient than the original resampling procedure. On average, the processing time for the experiments using Random Forest was 16 times shorter than for the experiments using C5.0 (data not shown). The shorter processing time makes it feasible to increase the accuracy of the Random Forest approach by increasing the number of virtual samples. It may be possible to achieve a higher accuracy with Random Forest than with the default resampling procedure with a sufficiently high number of virtual samples.

4.1.5. Measures of accuracy

The results showed that the optimal approach depended on the specific measure of accuracy. The map by Jacobsen (1984) provided the most accurate results for the first soil type (18%) and the first soil group (47%), while the map by the CEC (1985) provided more accurate results for the first three soil types (36%). In fact, three different experiments achieved the highest accuracy for the first soil type, the first three soil types and the first soil group, respectively. Therefore, there is no universally optimal output map, as the choice depends entirely on the measure of accuracy.

Researchers should therefore choose a measure of accuracy that reflects the intended use of the outputs. Consequently, the choice of the optimal map will rely on the end users' needs. For example, if the end users request a map of the most probable soil type, accuracy should only include the first soil type. However, for some purposes, the accuracies of all soil types are relevant. For example, Odgers et al. (2015) used a soil type probability map generated with DSMART to predict soil properties by means of weighted averages. For similar uses, researchers should calculate accuracy in a way that considers class probabilities. For other uses, the accuracy could take into account taxonomic differences between soil types in an approach similar to Rossiter et al. (2017).

4.2. Covariate importance

The covariate importance was very similar for the two input maps. This suggests that the surveyors worked within similar frameworks. It is not surprising, as they produced the two maps only a year apart in the same country. It is likely that the similarity reflects a consensus on the factors that affect soil formation in Danish soil science at the time.

In both maps, the factors soil (S), relief (R) and parent material (P) had a large importance. Climate (C) was also important, as precipitation had a high importance for both maps, while direct insolation had intermediate importance. The explanatory text for the map by Jacobsen (1984) also focuses on these four factors.

The high importance of soil properties (S) is logical, as they are the basis for classifying soils. For example, Luvisols and Acrisols have argic B horizons, while Arenosols have a sandy texture (FAO-Unesco, 1974). In Denmark, the parent material (P) has a large impact on the soil texture, which explains some of its high importance (Madsen et al., 1992; Adhikari et al., 2013). Additionally, the parent material defines some soil types, such as Histosols and Fluvisols.

The most important covariates related to the relief were variables that describe differences in elevation (*elevation*, *mrvbf*, *demdetrend*, *valldepth*). The surveyors apparently aimed to separate valleys from uplands and hills. In fact, the explanatory text for the map by Jacobsen (1984) describes Map Unit 5 as "valley soils" and Map Unit 9 as "hilly, sandy soils".

Climate (C) influences the distribution of some soil types in Denmark, as the western parts receive larger amounts of precipitation (Wang, 2013). The larger amounts of precipitation increases the rates of leaching, but it is difficult to isolate their effect, as the geology of western Denmark also differs from the eastern parts. Sandy glacial outwash plains and Saalian moraines dominate the western parts, while loamy Weichselian moraines are common in the eastern parts of the country (Jacobsen, 1984; Madsen et al., 1992).

Unlike the factors S, C, R and P, covariates relating to organisms (O) generally had a low importance. Land use, the most important covariate relating to organisms, had an intermediate importance. The reason for their low importance may be that the scale of the input maps cannot contain the detailed patterns in the vegetation. This circumstance may also explain the low importance of some of the topographic covariates, such as aspect, curvature, flow accumulation and the topographic wetness index. Some of these covariates have large variations within a short range, which the surveyors could not include due to the coarse scale of the maps. This is a clear disadvantage of using coarse scale soil maps for disaggregation.

It is also possible that the surveyors omitted vegetation as a conscious decision. Denmark is mostly agricultural, and the land use may have little correlation with soil types. Furthermore, the surveyors may have aimed to describe the soil in its "natural state". The two maps did not include Phaeozems, despite their high frequency in the observations. As stated earlier, the most likely reason for this is that they do not form under natural conditions in Denmark (Madsen and Jensen, 1996).

The horizontal distance to waterbodies, the only covariate relating to spatial position (N) had an intermediate importance. Other spatial trends may have played a role in the making of the maps, but we could not assess them, as we did not include any other spatial covariates. In fact, Holmes et al. (2015) was the only study using DSMART to include a purely spatial coordinate in the form of the distance to coastline. Alternatively, the x- and y-coordinates could be used as covariates to include spatial relationships.

The use of soil-landscape relationships increased the importance of the map of wetland areas, which we expected because we used it to modify the map units of the input maps. However, soil-landscape relationships did not increase the importance of the maps of the clay content to the same degree. This is possibly because the clay content was also important without soil-landscape relationships implemented.

5. Conclusions

In this study, we aimed to test the sensitivity of DSMART towards the conventional soil maps used as input data. We tested if soil-landscape relationships and area-proportional sampling improved the accuracy of the generated maps. Lastly, we tested the effect of replacing the default resampling procedure and C5.0 models with Random Forest models.

The accuracy of the outputs obtained with DSMART depend very strongly on the input maps. In this study, most of the dependence was due to differences in the shares of the soil types in the maps and the different levels of detail. The results suggest that detailed maps are most useful, even when they have a nominally coarser scale.

The inclusion of soil-landscape relationships and area-proportional sampling generally increased the accuracy of the results. These changes to DSMART are therefore a clear recommendation for future studies. This is highly relevant, as in some experiments, the accuracy of the disaggregated maps was lower than the accuracy of the input maps. However, the combination of soil-landscape relationships and area-proportional sampling resulted in output maps with higher accuracies than the input maps.

Changing the resampling procedure and decision tree models also affected the predictive accuracy of the outputs. Random Forest generally decreased the accuracy of the output maps. However, it was far more computationally efficient than the original procedure, so it may be possible to compensate for the lower accuracy by increasing the number of virtual samples. Potentially, other model types than decision trees may be useful, and testing the effects of model types should be an object of further study.

Software

The latest version of DSMART is available as an R package at https://bitbucket.org/brendo1001/dsmart/overview. This version implements area-proportional sampling and allows the user to specify the model type.

Acknowledgements

The authors are obliged to Dr. Gabor Löwei for providing feedback on an earlier version of the manuscript.

The authors acknowledge the Sydney Informatics Hub and the University of Sydney's high performance computing cluster Artemis for providing the high performance computing resources that have contributed to the research results reported within this paper.

The authors thank the two anonymous reviewers for their insightful feedback, which has improved the article over the original manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2019.01.038.

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 (3), 860–876. https://doi.org/10.2136/sssaj2012.0275.

Adhikari, K., Minasny, B., Greve, M.B., Greve, M.H., 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma 214-215, 101–113. https://doi.org/10.1016/j.geoderma.2013.09.023.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M.H., Grundy, M., Guerrero, E., Hempel, J.W., Hengl, T., Heuvelink, G.B.M., Batjes, N.H., Carvalho, E., Hartemink, A.E., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A.B., McKenzie, N.J., Vasquez, G.M., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J.A., Smith, S., Sulaeman, Y., Vintila, R., Rossel, R.V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Ibelles Navarro, A.R., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., van Liedekerke, M., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S.K., Moussadek, R., Badraoui, M., Da Silva, M., Paterson, G., Gonçalves, M.d.C., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19. https://doi.org/10.1016/j.grj.2017.06.001.

Auernhammer, H., 2001. Precision farming — the environmental challenge. Comput. Electron. Agric. 30 (1–3), 31–43. https://doi.org/10.1016/s0168-1699(00)00153-8.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36 (1–2), 105–139. https://doi.org/10.1023/A:1007515423169.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120 (1–2), 17–26. https://doi.org/10.1016/j.geoderma.2003.07.006.

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. https://doi.org/10.1016/S0016-7061(01)00070-2.

CEC, 1985. Soil map of the European Communities at Scale 1:1,000,000, Luxembourg.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W., Odgers, N.P., 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274, 54–67. https://doi.org/10.1016/j.geoderma.2016.03.025.

Cialella, A.T., Dubayah, R., Lawrence, W.T., Levine, E., 1997. Predicting soil drainage class using remotely sensed and digital elevation data. Photogramm. Eng. Remote. Sens. 63 (2), 171–178.

FAO-Unesco, 1974. Soil Map of the World, Paris.

FAO-Unesco, 1988. UNESCO soil map of the world, revised legend. World Resources Report 60, 138.

Giasson, E., Sarmento, E.C., Weber, E., Flores, C.A., Hasenack, H., 2011. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci. Agric. 68 (2), 167–174. https://doi.org/10.1590/S0103-90162011000200006.

Greve, M.H., Madsen, H.B., 1999. Soil Mapping in Denmark. European Soil Bureau Research Report.

Holmes, K.W., Griffin, E.A., Odgers, N.P., 2015. Large-area spatial disaggregation of a mosaic of conventional soil maps: evaluation over Western Australia. Soil Res. 53 (8), 865. https://doi.org/10.1071/sr14270.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56 (3), 836. https://doi.org/10.2136/sssaj1992.03615995005600030027x.

Jacobsen, N.K., 1984. Soil map of Denmark according to the FAO-UNESCO Legend. Dan. J. Geogr. 84, 93–98. https://doi.org/10.1080/00167223.1984.10649206.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. Dover Publications Inc., New York.

Kheir, R.B., Bøcher, P.K., Greve, M.B., Greve, M.H., 2010. The application of GIS based decision-tree models for generating the spatial distribution of hydromorphic organic landscapes in relation to digital terrain data. Hydrol. Earth Syst. Sci. 14 (6), 847–857. https://doi.org/10.5194/hess-14-847-2010.

Kovacic, D.A., David, M.B., Gentry, L.E., Starks, K.M., Cooke, R.A., 2000. Effectiveness of constructed wetlands in reducing nitrogen and phosphorus export from agricultural tile drainage. J. Environ. Qual. 29 (4), 1262. https://doi.org/10.2134/jeq2000.00472425002900040033x.

Madsen, H.B., Jensen, N.H., 1996. Soil map of Denmark according to the revised FAO legend 1990. Dan. J. Geogr. 96 (1), 51–59.

Madsen, H.B., Nørr, A.H., Holst, K.A., 1992. The Danish Soil Classification. The Royal Danish Geographical Society, Copenhagen, Denmark.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52. https://doi.org/10.1016/s0016-7061(03)00223-4.

Møller, A.B., Beucher, A., Iversen, B.V., Greve, M.H., 2018. Predicting artificially drained areas by means of a selective model ensemble. Geoderma 320, 30–42. https://doi.org/10.1016/j.geoderma.2018.01.018.

Nauman, T.W., Thompson, J.A., 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213, 385–399. https://doi.org/10.1016/j.geoderma.2013.08.024.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214-215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024.

Odgers, N.P., McBratney, A.B., Minasny, B., 2015. Digital soil property mapping and uncertainty estimation using soil class probability rasters. Geoderma 237-238, 190–198. https://doi.org/10.1016/j.geoderma.2014.09.009.

Oreskes, N., 1998. Evaluation (not validation) of quantitative models. Environ. Health Perspect. 106 (Suppl. 6), 1453–1460. https://doi.org/10.1289/ehp.98106s61453.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

Rossiter, D.G., Zeng, R., Zhang, G.-L., 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma 292, 118–127. https://doi.org/10.1016/j.geoderma.2017.01.012.

Schmidt, K., Behrens, T., Scholten, T., 2008. Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma 146 (1–2), 138–146. https://doi.org/10.1016/j.geoderma.2008.05.010.

Scull, P., Franklin, J., Chadwick, O., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27 (2), 171–197. https://doi.org/10.1191/0309133303pp366ra.

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. https://doi.org/10.1016/j.ecolmodel.2004.06.036.

Soil Survey Staff, 2016. Soil Survey Geographic (SSURGO) Database. Natural Resources Conservation Service. United States Department of Agriculture. https://sdmdataaccess.sc.egov.usda.gov, Accessed date: 1 January 2016.

Statistics Denmark, 2017. . Statistical Yearbook 2017.

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2016. Spatial disaggregation of complex soil map units at the regional scale based on soil-landscape relationships. Geoderma. https://doi.org/10.1016/j.geoderma.2016.06.006.

Wang, P.R., 2013. Referenceværdier: døgn-, måneds- og årsværdier for regioner og hele landet 2001–2010, Danmark for temperatur, relativ luftfugtighed, vindhastighed, globalstråling og nedbør. In: Teknisk Rapport 12–24. Danish Meteorological Institute.

Wright, M.N., Ziegler, A., 2015. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1). https://doi.org/10.18637/jss.v077.i01.