Geoderma 352 (2019) 351–359

Contents lists available at ScienceDirect

Geoderma

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

Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark

A. Beucher*, A.B. Møller, M.H. Greve

Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Soil
Mapping
Artificial neural networks
Decision tree classification
Modelling
Denmark

ABSTRACT

Soil drainage constitutes a substantial factor affecting plant growth and various biophysical processes, such as nutrient cycling and greenhouse gas fluxes. Consequently, soil drainage maps represent crucial tools for crop, forest and environmental management purposes. As extensive field surveys are time- and resource-consuming, alternative spatial modelling techniques have been previously applied for predicting soil drainage classes. The present study assessed the use of Artificial Neural Networks (ANN) for mapping soil drainage classes in Denmark and compared it to a Decision Tree Classification (DTC) technique. 1702 soil observations and 31 environmental variables, including soil and terrain parameters, and spectral indices derived from satellite images, were utilized as input data. Based on a 33% holdback validation dataset, the best performing ANN and DTC models yielded overall accuracy values of 54 and 52%, respectively. DTC models benefited from the use of all variables, but ANN models performed better after variable selection. Notably, ANN and DTC model performances were comparable although differential costs for misclassification were only implemented for DTC modelling. Nevertheless, both methods produced predictive drainage maps in accordance with one another and demonstrated promising classification abilities over a large study area (c. 43,000 km^2).

1. Introduction

Soil drainage can be described as the degree and frequency at which the soil is free of water saturation. This factor strongly influences plant growth and various biophysical processes, such as nutrient cycling and greenhouse gas fluxes (Levine et al., 1994; Nuutinen et al., 2001; Smith et al., 2003). Soil drainage maps thus are crucial for crop, forest and environmental management purposes. As extensive field surveys are time- and resource-consuming, the application of spatial modelling techniques represents a great alternative for predicting soil drainage classes. In the past, numerous approaches have been assessed: multivariate discriminant analysis (Bell et al., 1994; Bell et al., 1992; Kravchenko et al., 2002; Liu et al., 2008; Niang et al., 2012), logistic modelling (Campling et al., 2002), indicator kriging and cokriging (Kravchenko et al., 2002), unsupervised classification (Peng et al., 2003) and linear transformation (Zhao et al., 2013). Moreover, Zhao et al. (2013) applied artificial neural networks (ANNs) to map seven drainage classes over a large area (55,000 km^2) in Nova Scotia, Canada. Cialella et al. (1997) and Niang et al. (2012) used decision tree classification (DTC) to map five drainage classes in relatively small areas: a 24-km^2 site in Maine, USA, and a 167-km^2 area in Quebec, Canada, respectively. Lemercier et al. (2012) also applied DTC to predict four soil drainage classes over a 4645-km^2 area in Brittany, France. Finally and most notably, Møller et al. (2017) used DTC for mapping soil drainage classes in Denmark (c. 43,000 km^2). ANNs and DTC both constitute machine-learning techniques. ANNs represent efficient pattern recognition and classification tools (Bonham-Carter, 1994). The ability to generalize from imprecise input data (Porwal et al., 2003) and to handle large datasets (Gershenfeld, 1999) constitutes the principal advantages of ANNs. Large input datasets (i.e. soil observations and environmental variables) were available for this study, which motivated the application of ANNs. Moreover, ANNs have been frequently assessed for predicting different soil attributes (Chang and Islam, 2000; Lentzsch et al., 2005; Minasny and McBratney, 2002; Viscarra Rossel and Behrens, 2010) or soil classes (Behrens et al., 2005; Boruvka and Penizek, 2007; Cavazzi et al., 2013; Chagas et al., 2013; Silveira et al., 2013; Zhu, 2000). DTC also constitutes an efficient classification tool, which presents several advantages: it is computationally inexpensive, makes no assumptions about the distribution of the environmental variables and is robust towards missing data and redundant environmental variables (Mitchell, 1997; Quinlan, 1996; Rokach and Maimon, 2005; Tan et al., 2014). In the past, DTC were successfully applied to map categorical soil variables in many different studies (Adhikari et al., 2014; Chaney et al., 2016; Giasson et al., 2011; Kheir et al., 2010a;

* Corresponding author.
E-mail addresses: amelie.beucher@agro.au.dk (A. Beucher), anbm@agro.au.dk (A.B. Møller), mogensh.greve@agro.au.dk (M.H. Greve).

https://doi.org/10.1016/j.geoderma.2017.11.004
Received 1 May 2017; Received in revised form 25 September 2017; Accepted 2 November 2017
Available online 13 November 2017
0016-7061/

Kheir et al., 2010b; Lagacherie and Holmes, 1997; McBratney et al., 2003; Moran and Bui, 2002; Odgers et al., 2014).

In Denmark, soil drainage conditions are divided into five classes: very well-drained soils (DC1), well-drained soils (DC2), moderately well-drained soils (DC3), poorly drained soils (DC4), and very poorly drained soils (DC5). The classes are mainly defined from morphological characteristics (e.g. the presence and depth of pseudogley, reduced horizons and histic epipedons; Madsen and Jensen, 1988) and have been used in the description of soil profiles.

The main objective of the present study is to assess the predictive classification abilities of an ANN approach and compare them to a DTC technique for mapping soil drainage classes in Denmark. Similar input datasets (i.e. soil observations and environmental variables) were used to enable comparison between the two techniques. Moreover, the present study utilized the best performing DTC model achieved in the study of Møller et al. (2017).

2. Study area

Denmark constitutes the study area and covers approximately 43,000 km^2. The climate is temperate with a mean annual temperature of 7.7 °C, ranging from 0 °C in January to 16 °C in July (measured in the years 1961–1990). The mean annual precipitation is 700 mm, ranging from 600 mm in the eastern part of Denmark to 800 mm in the western part (Danish Meteorological Institute, 1998). The landscape is generally flat with a mean elevation of 31 m above sea level. Geologically, the country can be divided in two main areas: the western part of Denmark mostly consists of sandy glacial outwash plains and Saalian moraine, and the eastern part of loamy Weichselian moraine. Luvisols and Podzols constitute the dominant soil types in the eastern and western parts of Denmark, respectively (Jacobsen, 1984). In the past, wetlands covered more than 20% of the country, but their extent was reduced during the 19th and 20th centuries, due to drainage activities. Wetlands have consisted of meadows in river valleys and along the coasts, marshes in the south-western part of the country, raised sea beds in the northern part, as well as sinks in kettled moraine landscapes (Madsen et al., 1992). Agriculture represents the main land use type and covers 66% of the country, followed by natural vegetation (16%) and urban areas (10%).

3. Materials and methods

3.1. Soil observations

Soil observations used in this study (Fig. 1) originate from two sources: 860 profiles were sampled along a main gas pipeline in the years 1981–1985 and 842 profiles were taken following a 7-km grid in the years 1987–1989 (Madsen et al., 1992). Each profile was dug down to 1.8 m depth or less depending on the water table. The soil profiles were described according to an adaptation of FAO's “Guidelines For Soil Profile Description” for Danish conditions (Madsen and Jensen, 1988). For each profile, samples were collected according to the genetic horizon sequence, which was described, as well as depth, colour and texture, among others. The samples were analyzed for various physical and chemical properties (e.g. pH, organic matter and calcium carbonate contents), as described in Madsen et al. (1992). For most of the profiles (n = 1697), the drainage classes assigned during the original surveys were used. The remaining five profiles had their drainage classes determined from the profile description and pictures. The original dataset (n = 1702) was divided into two datasets using a stratified random split to get equal proportions of drainage classes in the training data (two thirds) and validation data (one third; Table 1).

3.2. Environmental variables

In the present study, 31 environmental variables were utilized as predictors within the modelling (Table 2). Different categorical variables were used: geology map (Jakobsen et al., 2015), geo-region (Adhikari et al., 2013), landscape elements (Madsen et al., 1992), land use (European Environment Agency, 2014) and wetlands (Greve et al., 2014). Moreover, 16 terrain parameters were derived from a Light Detection and Ranging (LiDAR)-based Digital Elevation Model (DEM produced by the National Survey and Cadastre of the Danish Ministry of Environment), with a 30.4-m resolution aggregated from the original 1.6-m DEM (by calculating the mean value of the 19 × 19 cells covered by each of the cells in the new layer). The parameters (Table 2) were extracted in ArcGIS (ESRI, 2014) and SAGA GIS (SAGA GIS, S, n.d.). The depth to groundwater was calculated from a groundwater table modelled at a 500-m resolution (Henriksen et al., 2012). The groundwater table was first resampled to a 30.4-m resolution using bilinear interpolation and then subtracted from the DEM. Soil clay content layers for four depth intervals (i.e. 0–30, 30–60, 60–100 and 100–200 cm) were calculated by aggregating the original layers created by Adhikari et al. (2013) were also utilized as predictors within the modelling. Different spectral indices (i.e. Normalized Difference Vegetation Index or NDVI, Normalized Difference Water Index or NDWI, Normalized Difference Moisture Index or NDMI, and Soil Adjusted Vegetation Index or SAVI) were derived from Landsat 8 images in 30-m resolution recorded in March 2014, which was the only month for which cloud free images could be obtained for Denmark. The Landsat 8 images included Band 3 (Green 530–590 nm), Band 4 (Red 640–670 nm), Band 5 (Near Infrared, NIR, 850–880 nm) and Band 6 (Shortwave Infrared 1, SWIR1, 1570–1650 nm). The indices were calculated using the following equations:

$$\mathrm{NDMI} = \frac{\mathrm{NIR} - \mathrm{SWIR1}}{\mathrm{NIR} + \mathrm{SWIR1}} \quad (1)$$

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}} \quad (2)$$

$$\mathrm{NDWI} = \frac{\mathrm{NIR} - \mathrm{Green}}{\mathrm{NIR} + \mathrm{Green}} \quad (3)$$

$$\mathrm{SAVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red} + L}^{*}(1 + L) \quad (4)$$

With L corresponding to the soil brightness correction factor and set to 0.5.

3.3. Artificial neural networks

ANNs are supervised machine-learning techniques, which determine associations between known set of observations (i.e. training points) and different environmental variables in order to classify new, unknown data (Zell et al., 1998). The main advantages of ANNs are their ability to handle large datasets (Gershenfeld, 1999; Chagas et al., 2013), to approximate non-linear relationships (Viscarra Rossel and Behrens, 2010) and to generalize from relatively imprecise input data (Porwal et al., 2003). They also are robust to handle noise, outliers and overfitting (Gershenfeld, 1999; Viscarra Rossel and Behrens, 2010).

An ANN method based on Radial Basis Function (RBF) was applied in this study. The use of RBFs yields a simple and efficient interpolation approach (Porwal et al., 2003). The application of an ANN method includes two phases: training and classification. The architecture of the assessed neural networks comprises three interconnected layers. First, an input layer consists in several nodes (i.e. one for each tested environmental variable). Second, a hidden layer includes various artificial neurons (i.e. different numbers of neurons were tested), each representing a RBF (using a Gaussian activation function). Third, an output layer contains five artificial neurons (i.e. one for each predicted soil drainage class), transmitting predicted output values. The output values constitute probability values ranging between 0 and 1 (i.e. the lowest and highest probability to be part of a soil drainage class,

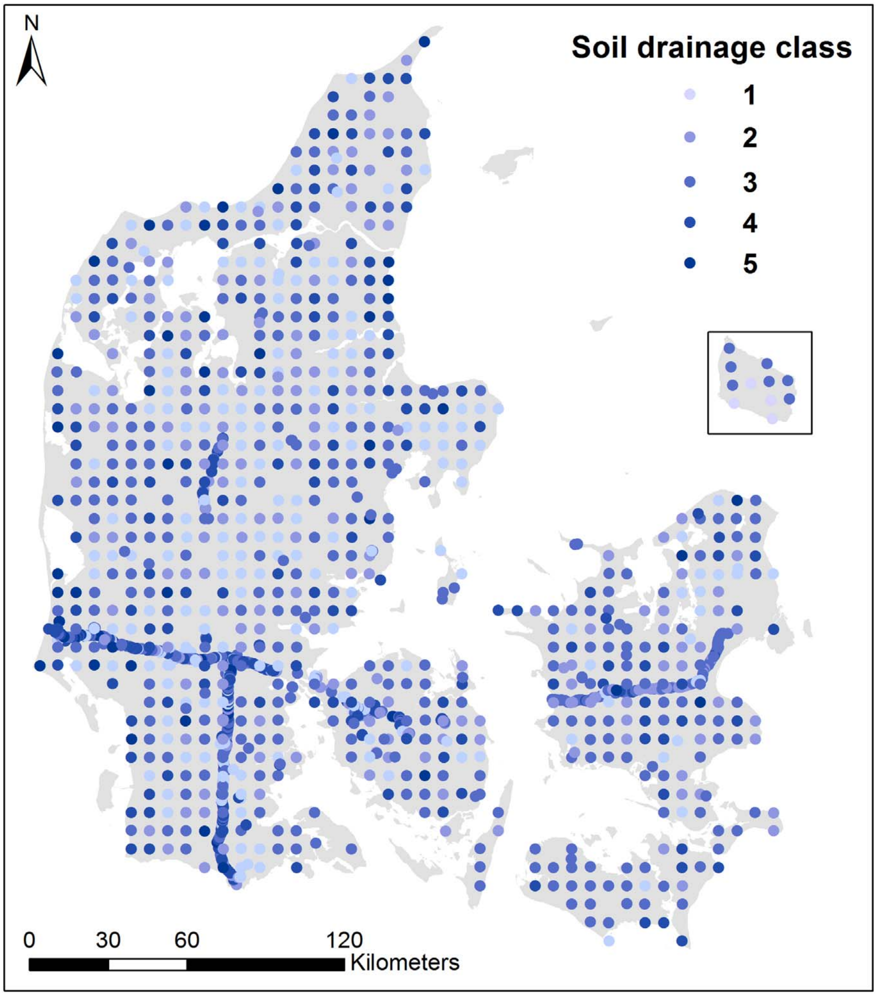

Fig. 1. Location of the soil profiles in the study area.

Table 1
Distribution of the soil observations in the different drainage classes (DC) within the training, validation and original datasets.

DC	Description	Training	Validation	Total
1	Very well-drained soils	221	110	331
2	Well-drained soils	191	95	286
3	Moderately well-drained soils	426	213	639
4	Poorly drained soils	248	125	373
5	Very poorly drained soils	49	24	73
	Total	1135	567	1702

respectively).

During the training, the network first receives the different input variables corresponding to known observations (i.e. training points representing the different drainage classes) through the input layer. Then, these input variables are transmitted to the hidden layer neurons, which compute and extract significant information from them in order to predict output values. Weights are defined for each connection through an initialization function (Zell et al., 1998). During the training, the network iteratively adjusts these weights using a learning function (Bergmeir and Benítez, 2012) so that the predicted output value matches as closely as possible the known target output value for each training point (Gershenfeld, 1999). During the classification, the network uses the unknown point data for the whole study area and classifies them using the calibrated weights (Beucher et al., 2015). Therefore, the network predicts output values for each pixel within the study area, enabling the creation of a predictive map.

The RBF-based ANN used in this study was implemented within R environment using the RSNNS (i.e. R Stuttgart Neural Network Simulator) package (Bergmeir and Benítez, 2012) for the ANN training and classification. The package constitutes an interface for the Stuttgart Neural Network Simulator (SNNS; Zell et al., 1998) used for ANN implementation. In-house routines were developed in R to automatically and systematically test a large number of ANN models, run validation, select and post-process the models achieving the highest performance results. By adjusting different parameters, several network topologies were assessed: the number of nodes in the input layer, the number of neurons in the hidden layer (from 10 to 100), the number of iterations (from 20 to 500), as well as the initialization and learning functions parameters. The amount of nodes in the input layer represents the number of environmental variables tested by the network. Therefore, variable selection was carried out by removing one variable at the time from the input data to evaluate the network performance.

The relative importance of the different environmental variables in the model was estimated using the Garson algorithm (Garson, 1991) modified according to Goh (1995). The algorithm determines the relative importance (i.e. strength of association to the output) of each environmental variable by partitioning the connection weights between the input node of interest and the hidden and output neurons. A single value, ranging between 0 and 1, describing the relationship of the input variable with the output in the model is obtained (Goh, 1995).

3.4. Decision tree classification

Decision tree classification (DTC) also constitutes a machine-

Table 2
Environmental variables used as predictors in the study (based on Møller et al., 2017).

Environmental variable	Mean (range)/number of classes	Data source/brief description
Slope aspect	181.1 (0–360)	DEM/direction of the steepest slope from the North
Reclassified slope aspect	8 classes	DEM/slope aspect reclassified into 8 general directions
Blue spot analysis	0.1 (0.0–92.5)	DEM/depth of sinks
Clay content (0–30 cm)	8.1 (0.0–51.2)	Clay content (%) for 0–30 cm soil depth
Clay content (30–60 cm)	10.1 (0.0–62.7)	Clay content (%) for 30–60 cm soil depth
Clay content (60–100 cm)	11.2 (0.0–59.1)	Clay content (%) for 60–100 cm soil depth
Clay content (100–200 cm)	10.9 (0.0–57.1)	Clay content (%) for 100–200 cm soil depth
Detrended elevation model	1.0 (− 57.9–105.4)	DEM/elevation minus the mean elevation within a 4 km radius
Direct insolation	1269 (122–1707)	DEM/potential incoming solar radiation calculated for a single year (kWh)
Cropping history	1.6 (− 4–4)	Digital Field Map/number of years with drainage dependent crops minus the number of years with drainage independent crops from 2011 to 2014
Elevation	30.9 (− 39.5–170.5)	DEM/LiDAR produced elevation of the land surface
Flow accumulation	60 (1–110,908)	DEM/number of upslope cells
Geology	10 classes	Scanned and registered geological map (Scale 1:25,000)
Geo-regions	7 classes	Scanned geographical regions map (Scale 1:100,000)
Horizontal distance to channel network	279 (0–4401)	Digital map of surface water/horizontal distance to the nearest waterbody
Landscape elements	12 classes	Landform types (Scale 1:100,000)
Land use	5 classes	CORINE land cover data (Scale 1:100,000)
Mid-slope position	0.27 (0–1)	DEM/covers the warmer zones of slopes
Multiresolution valley bottom flatness	4.3 (0.0–10.9)	DEM/calculated depositional areas
Normalized difference moisture index	0.08 (− 1–1)	Landsat 8/NDMI
Normalized difference vegetation index	0.52 (− 1–1)	Landsat 8/NDVI
Normalized difference water index	− 0.54 (− 0.99–1.00)	Landsat 8/NDWI
SAGA Wetness Index	14.5 (6.9–19.1)	DEM/Topographic wetness index with modified catchment area
Soil adjusted vegetation index	0.29 (− 0.29–0.72)	Landsat 8/SAVI
Slope gradient	1.6 (0.0–90)	DEM/local slope gradient (degrees)
Slope to channel network	1.0 (0.0–78.9)	DEM/slope (degrees) to the hydrologically nearest waterbody
Topographic wetness index	5.9 (− 15.8–63.3)	DEM/$TWI = \ln(a/\tan b)$, with flow accumulation a and local slope gradient b
Valley depth	7.5 (0.0–89.9)	DEM/extent of the valley depth
Vertical distance to channel network	1.4 (0.0–45.9)	DEM/calculated vertical distance (m) to waterbodies
Depth to groundwater	5.8 (0.0–126.0)	DEM/depth (cm) to upper groundwater table
Wetlands	4 classes	Presence of wetlands, central wetlands and peat (Scale 1:20,000)

Table 3
Matrix with differential costs for misclassification used for the DTC models in this study (from Møller et al., 2017).

		Original DC				
		1	2	3	4	5
Predicted DC	1	0	1	2	3	4
	2	1	0	1	2	3
	3	2	1	0	1	2
	4	3	2	1	0	1

Table 4
Validation results for three different ANN models, using all variables or excluding one or two variables (cropping history and reclassified slope aspect), and the best DTC model.

Validation criteria	Variable selection for ANN models			DTC model
	All	Cropping history out	Crop hist + recl slop asp out	
Accuracy	0.50	0.52	0.54	0.52
Kappa	0.30	0.33	0.35	0.32
MAE	0.76	0.74	0.72	0.74

Table 5
Network parameters and validation criteria for the ANN models yielding the highest performance results (in bold, the best performances).

Network parameters			Validation		
Number of			Criteria		
Input nodes	Hidden neurons	Iterations	OA	K	MAE
31	10	160	0.42	0.26	0.80
31	20	180	0.46	0.28	0.78
31	**40**	**200**	**0.50**	**0.30**	0.76
31	60	200	0.49	0.30	0.76
31	80	220	0.47	0.28	0.75
30	10	160	0.44	0.29	0.79
30	20	180	0.48	0.31	0.76
30	40	180	0.50	0.32	0.75
30	**50**	**200**	**0.52**	**0.33**	0.74
30	70	200	0.51	0.33	0.74
29	10	160	0.45	0.32	0.75
29	20	180	0.49	0.33	0.73
29	**40**	**180**	**0.54**	**0.35**	0.72
29	60	200	0.52	0.35	0.72
29	70	200	0.50	0.34	0.73

learning technique, working by recursive partitioning of a dataset to achieve a homogenous classification of a target variable. At each split the algorithm aims at reducing the entropy of the target variable in the resulting datasets by choosing the optimal split from a number of independent variables. The main advantages of this method are that it is computationally inexpensive, makes no assumptions about the distribution of the environmental variables and is robust towards missing data and redundant environmental variables (Mitchell, 1997; Quinlan, 1996; Rokach and Maimon, 2005; Tan et al., 2014). Since single decision trees usually are weak classifiers, the predictions of several decisions trees are commonly combined to obtain a better predictive performance by means of ensemble techniques such as bagging (Bauer and Kohavi, 1999; Dietterich, 2000). Bagging draws a number of bootstrap samples from the training data and builds a classifier from each bootstrap sample. The classifiers' predictions are then combined by simple voting (Breiman, 1996).

In particular, Møller et al. (2017) demonstrated that the best DTC model for predicting soil drainage classes in Denmark was achieved using bagging with all environmental variables and differential costs for misclassification. The DTC technique was implemented using the C5.0 R package (Kuhn et al., 2015) which is based on the decision tree

Table 6
Confusion matrices built with the validation dataset for the best performing ANN and DTC models.

	Observed DC					
	1	2	3	4	5	UA (%)
ANN predicted DC						
1	63	15	10	5	1	67.0
2	30	21	8	1	2	33.9
3	11	46	160	54	0	59.0
4	6	13	34	59	18	45.4
5	0	0	1	6	3	30.0
PA (%)	57.3	22.1	75.1	47.2	12.5	OA = 54%
DTC predicted DC						
1	65	24	19	13	3	52.4
2	13	16	13	4	0	34.8
3	28	49	154	45	0	55.8
4	4	6	27	59	20	50.9
5	0	0	0	4	1	20.0

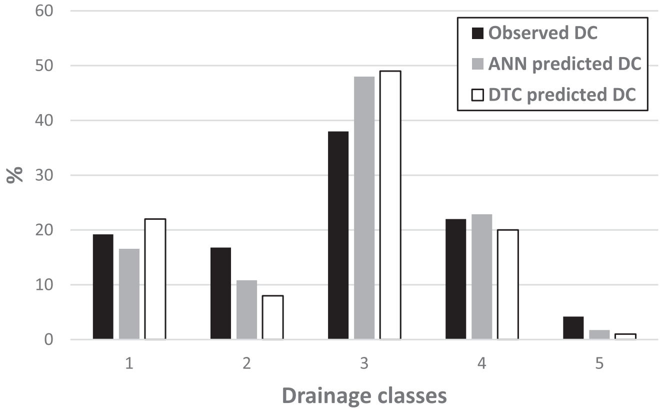

Fig. 2. Distribution of the different drainage classes in the observed data (black) and in the predictions of the best performing ANN (grey) and DTC (white) models.

algorithm by Quinlan (1993). Concerning the differential costs for misclassification, the instances in the training data are assigned weights, which are proportional to the cost of misclassifying the class to which the instance belongs. The algorithm then chooses the splits that minimize the costs associated with the resulting classification. In order to take into account the fact that drainage classes form an ordered series, costs equal to the absolute differences between the predicted and the original drainage classes were used (Table 3). The off-diagonal elements give the costs for misclassification and the diagonal elements are zero as they represent the cost for correct classification.

3.5. Validation

The performance of the different models was evaluated using a 30% holdback validation data. The overall accuracy (OA), kappa (K) coefficient and mean absolute error (MAE) were systematically calculated in order to compare models and select the ones yielding the best performances, which were then automatically recorded.

$$\mathrm{OA} = \frac{\sum_{i=1}^{m} E_{ii}}{n} \tag{5}$$

With m corresponding to the number of drainage classes, E_{ii} the sum of correctly classified observations and n the total number of observations.

$$\mathrm{K} = 1 - \frac{1 - \mathrm{OA}}{1 - p_e} \tag{6}$$

With p_e corresponding to the hypothetical probability of chance agreement.

Table 7
Relative importance of the environmental variables for the best performing ANN.

Environmental variable	Relative importance
Clay content (100–200 cm)	1
Wetlands	0.99
Slope to channel network	0.95
Clay content (30–60 cm)	0.93
Geology	0.92
Horizontal distance to channel network	0.86
Clay content (60–100 cm)	0.85
Vertical distance to channel network	0.82
SAGA Wetness Index	0.82
Depth to groundwater	0.80
Clay content (0–30 cm)	0.76
Direct insolation	0.73
Blue spot analysis	0.70
Land use	0.69
Geo-regions	0.68
Landscape elements	0.58
Multi-resolution valley bottom flatness	0.50
Flow accumulation	0.45
Slope gradient	0.41
NDMI	0.39
Topographic wetness index	0.38
Slope aspect	0.35
NDWI	0.33
SAVI	0.32
Valley depth	0.29
Elevation	0.28
NDVI	0.27
Mid-slope position	0.25
Detrended elevation model	0.24

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7}$$

With y_i and $\hat{y}_i$ corresponding to the observed and predicted value of the observation i, respectively, and n the total number of observations.

For the best performing models, confusion matrices were calculated in order to examine the relationship between observed and predicted drainage classes for the validation dataset. Two criteria were utilized to evaluate further the prediction accuracy. The User Accuracy (UA) indicates the probability that the predicted data matches the observed data within a class. The Producer Accuracy (PA) gives a measure of how well the model predicts the observed data within a class.

$$\mathrm{UA}_j = \frac{X_{jj}}{\sum_{j=1}^{m} X_{jk}} \tag{8}$$

$$\mathrm{PA}_k = \frac{X_{kk}}{\sum_{k=1}^{m} X_{jk}} \tag{9}$$

With X_{ii} and X_{kk} corresponding to the diagonal values for each class in one row and in one column, respectively, and X_{jk} being the sum of values in one row or column.

4. Results and discussion

4.1. Variable selection and model performance

Different ANN models were tested utilizing various selections of environmental variables (i.e. different numbers of input nodes in the network). When using all variables, the best performing ANN model displayed lower OA and K values than the best performing DTC model (Table 4). However, the removal of two variables, cropping history and reclassified slope aspect, resulted in higher ANN model performance (i.e. highest OA and K, as well as lowest MAE) than when using all variables (Table 4). Several reasons may explain this finding. First, the

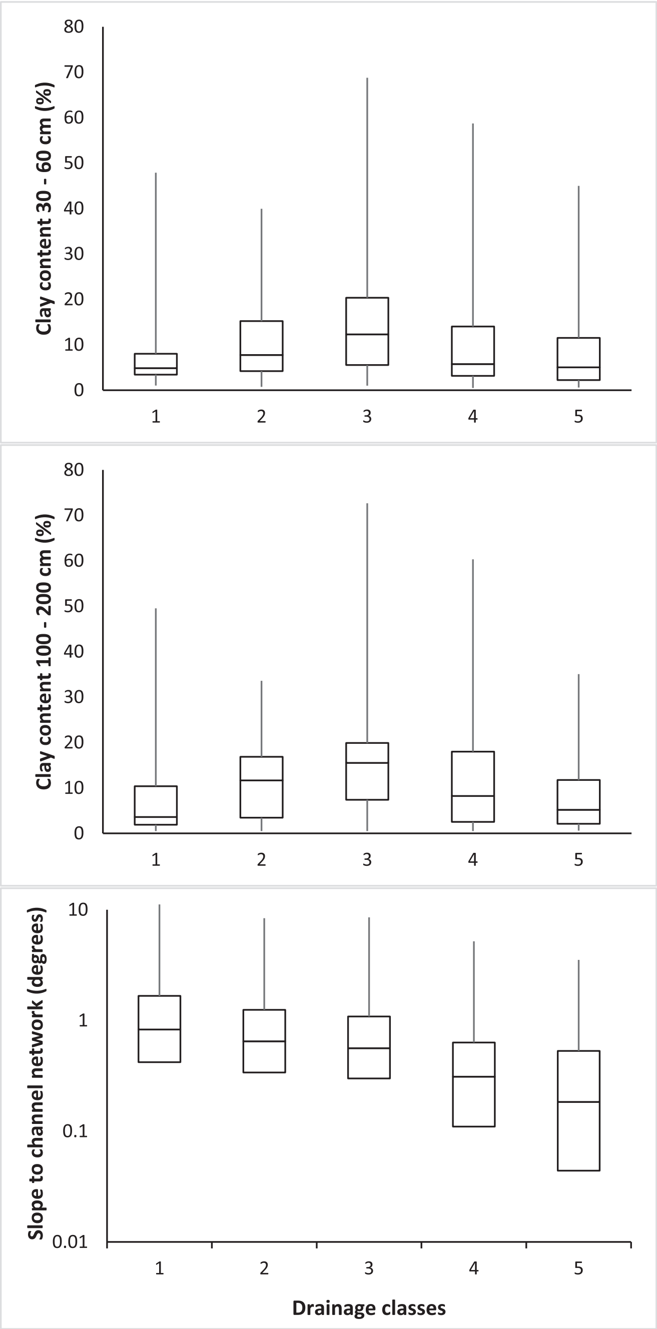

Fig. 3. Clay contents (from 30 to 60 cm and 100 to 200 cm depth) and slope to channel network values for the soil observations in the different drainage classes.

cropping history variable presented many missing values (i.e. NAs). This incomplete dataset might have hampered the predictive performance of the models. Secondly, the various drainage requirements of the crops might have not correlated with the characterization of soil drainage classes. Finally, the reclassified slope aspect may have appeared as a redundant variable since the slope aspect was also used as a predictor. The models performed better without using the variables that only added noise to the modelling. Even though ANNs have the ability to extract relevant information from large datasets (Chagas et al., 2013; Gershenfeld, 1999) and generalize from relatively imprecise information (Porwal et al., 2003), variable selection might still constitute an important factor to take into account while using ANNs. On the other hand, DTC appear as more resilient than ANNs, being able to deal with missing or redundant information within the variables (Quinlan, 1996).

Table 8
Distribution of the soil observations in the geological and wetland classes according to their drainage class (%).

%		Drainage class				
		1	2	3	4	5
Geology	Aeolian sand	38.0	14.1	16.9	23.9	7.0
	Freshwater clay	3.6	17.9	25.0	32.1	21.4
	Freshwater sand	19.2	4.1	16.4	47.9	12.3
	Freshwater peat	2.6	2.6	0.0	60.5	34.2
	Marine clay	0.0	0.0	30.0	50.0	20.0
	Marine sand	5.9	14.7	11.8	50.0	17.6
	Glacial clay	10.7	17.4	54.7	16.0	1.2
	Glacial sand	34.3	19.9	28.7	15.0	2.1
	Meltwater clay	12.5	29.2	29.2	25.0	4.2
	Meltwater sand	27.0	17.5	24.8	26.7	4.1
Wetlands	Non-wetlands	21.3	18.4	41.4	17.3	1.6
	Wetlands	13.2	11.8	19.1	41.4	14.5
	Central wetlands	9.6	7.2	22.9	43.4	16.9
	Peat	0.0	0.0	2.4	64.3	33.3

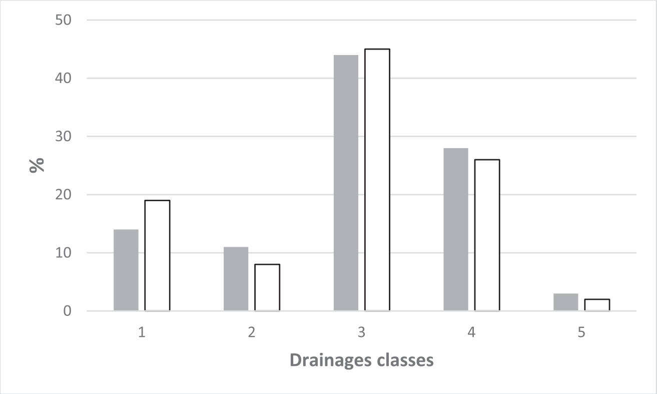

Fig. 4. Distribution of the drainage classes in the predictive maps produced by the ANN (grey) and DTC (white) models.

Other network parameters were also tested. For the best performing models, the number of hidden neurons ranged between 10 and 80, and the number of iterations between 160 and 220 (Table 5). The optimal number of hidden neurons was 40 for the models using all variables (n = 31) or 29 variables, and 50 for the model using 30 variables (Table 5). The optimal number of iterations was 180 for the model using 29 variables and 200 for the models using 30 or 31 variables (Table 5). The use of more hidden neurons or more iterations did not yield higher OA or K values. Moreover, the MAE values first stabilized before increasing with increasing numbers of iterations or hidden neurons (Table 5).

The best performing ANN model predicted DC1 and DC3 well (PA = 57.3 and 75.1%, respectively), DC4 moderately well (PA = 47.2%), and DC2 and DC5 poorly (PA = 22.1 and 12.5%, respectively; Table 6). Additionally, the predicted DC1 and DC3 matched the observed corresponding classes well (UA = 67.0 and 59.0%, respectively), the predicted DC4 moderately well (UA = 45.4%), and the predicted DC2 and DC5 poorly (UA = 33.9 and 30.0%, respectively; Table 6). The best DTC model predicted the different drainage classes in a similar way, except DC2 and DC5 which were more poorly predicted (PA = 16.8 and 4.2%, respectively; Table 6). The predicted DC5 also matched the observed drainage class poorly (UA = 20.0%; Table 6).

Considering each drainage class, most of the misclassifications were within +/− one drainage class of the correct observed class for both models (Table 6). In all, 90.2 and 86.4% of the validation data was predicted within +/− one drainage class of the observed data for the ANN and DTC models, respectively. In terms of frequency, DC3 was over-predicted while DC2 and DC5 were under-predicted by both

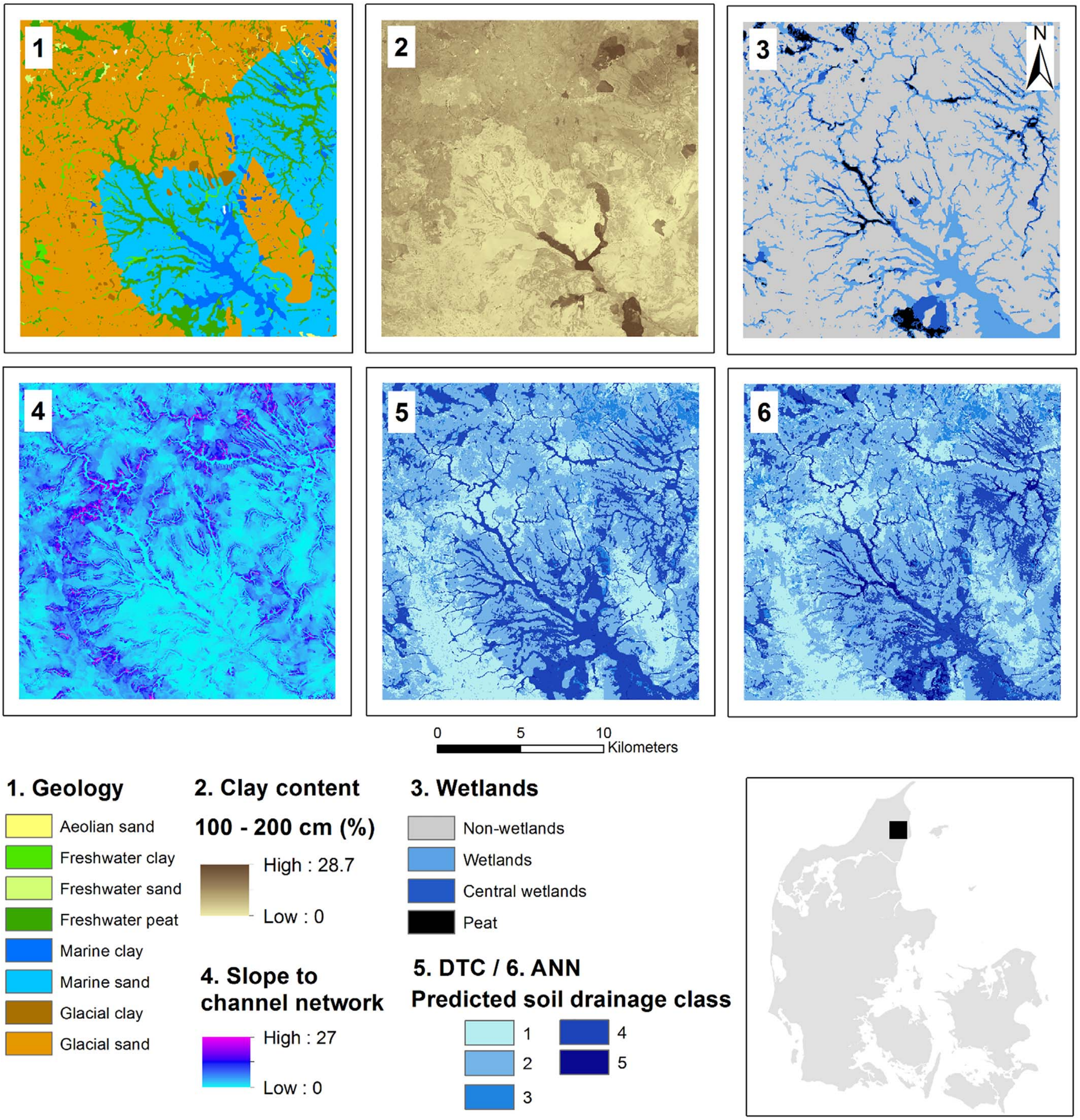

Fig. 5. Maps of the four most important environmental variables, as well as the best ANN and DTC models, for a small example area located in northern Denmark.

models (Fig. 2). However, both over- and under-predictions were less severe in the ANN model. The frequencies of ANN and DTC predicted DC1 and DC4 were rather similar to the frequencies of their observed classes (± 1 to 3%; Fig. 2).

4.2. Environmental variables

For the best ANN model, the five most important contributors were the clay content from 100 to 200 cm, wetlands, slope to channel network, clay content from 30 to 60 cm, and geology, with a relative importance higher than 0.9 (Table 7). The best contributors for the DTC model were similar, except land use replaced the clay content from 30 to 60 cm (Møller et al., 2017). Clay contents at depths 30–60 cm and 100–200 cm increased from DC1 to DC3 (Fig. 3) which is most probably due to the low hydraulic conductivity of clayey soils resulting in water stagnation. Clay contents then decreased from DC4 to DC5 (Fig. 3). Soil texture was considered as a major determining factor for drainage classes (Brady, 1990) and utilized in previous studies (Zhao et al., 2013; Zhao et al., 2008). The slope to channel network clearly decreased with the drainage class (Fig. 3), confirming the importance of this topographic feature for predicting soil drainage classes. Moreover, the horizontal and vertical distances to the channel network also had a relative importance higher than 0.82 (Table 7). This emphasizes the relevance of using topographic features associated with the position of water bodies for mapping drainage classes, which previous studies indicated as well (Bell et al., 1994, Bell et al., 1992; Kravchenko et al., 2002; Zhao et al., 2013, Zhao et al., 2008).

The geology was also one of the most important variables. Glacial and aeolian sands were the well-drained geological classes (more than 52% of the soil observations corresponding to DC1or DC2; Table 8) while freshwater peat and marine clay represented the most poorly drained classes (more than 70% of the soil observations corresponding to DC4 or DC5; Table 8). As expected, the non-wetland areas were mainly well drained (more than 81% of the soil observations corresponding to DC1, DC2 or DC3; Table 8). Wetlands and central wetlands were generally poorly drained (more than 56% of the soil observations corresponding to DC4 or DC5; Table 8) while peat areas were the most poorly drained areas (more than 97% of the soil observations corresponding to DC4 or DC5; Table 8). The wetland layer represented an important environmental variable as it enabled distinguishing the poorly drained classes (DC4 and DC5) from the well-drained classes (DC1 to DC3).

4.3. Mapping

The ANN and DTC models produced similar maps, the agreement and the mean absolute difference between the predicted drainage classes reaching 79% and 0.3, respectively. For both models, DC3 represented the dominant drainage class on the loamy till areas in eastern Denmark. Poorly drained soils, DC4 and DC5, were logically found in wetland areas, DC4 representing the dominant drainage class in wetlands. In western Denmark, well-drained soils, DC1 and DC2, mostly occurred in high-lying sandy areas while moderately well drained soils, DC3, were found in clayey areas and low-lying areas. The drainage class proportions in the maps produced by the two models were also similar (Fig. 4). The DTC model predicted slightly higher proportions of DC1 and DC3, and slightly lower proportions of DC2, DC4 and DC5 (Fig. 4). DC3 constituted the most common drainage class with 48% and 49% of the mapped areas in the ANN- and DTC-based maps, respectively. DC4 covered 23% and 20% of the mapped areas in the respective maps while DC5 was only found in 2% and 1% of the area (Fig. 4).

Fig. 5 displays the maps of the common four most important environmental variables for the best ANN and DTC models for an example area in northern Denmark, as well as the corresponding drainage class predictive maps. For both models, DC4 appear as the dominant drainage class in wetland areas. DC1 and DC2 were the dominant drainage classes in the glacial sand areas. In particular, DC1 correspond to areas with the lowest clay contents and the largest slope to the channel network. Both models predicted DC2 and DC4 in marine deposit areas. DC3 was seldom predicted in the example area even though it represented the most frequently predicted drainage class on a national scale.

Being the most and the least represented drainage classes in the original dataset (Table 1), DC3 and DC5 were clearly over- and under-predicted by the models, respectively, which constitutes a bias in the predictions. Both models predicted classes represented by few observations poorly as the performance of ANN models is positively correlated to the amount of observations used for the network training (Viscarra Rossel and Behrens, 2010) and DTC models induce general splitting rules (Holte et al., 1989). Furthermore, it is notable that the implementation of differential misclassification costs improved the predictive abilities of DTC models. These costs could not be implemented in ANN models, but it could be argued that the connecting weights, automatically and systematically adjusted by the networks, constitute an equivalent. However, the extent of the study area did constitute a challenge because of the long computation time associated with ANNs and the computation limits inherent to RSNNS. In particular, RSNNS cannot handle parallel computing which constitutes a strenuous limit for accelerating computation. DTC models thus represent the fastest option in terms of computation time.

5. Conclusion

Two machine-learning techniques were assessed for predicting soil drainage classes in Denmark: an Artificial Neural Network (ANN) method and a Decision Tree Classification (DTC) method. Both methods demonstrated promising predictive classification abilities over large areas, producing drainage maps mostly in accordance with one another. An ANN model using 29 environmental variables yielded the best performance with an overall accuracy of 54%, while the best DTC model using all variables ($n = 31$) and differential costs for misclassification reached an overall accuracy of 52%. While DTC models benefited from the use of all variables, ANN models performed better than DTC after variable selection. The ANN model still performed slightly better DTC model without implementing differential costs for misclassification. Moreover, computation time remains a clear limitation of ANNs in comparison with DTC. In the future, alternative machine-learning techniques such as Random Forest and Support Vector Machines should be evaluated for predicting soil drainage classes. Combining several models in an ensemble would also constitute a relevant method to evaluate in order to improve the prediction accuracy of soil drainage classes.

Acknowledgments

The authors thankfully acknowledge the financial support from Ingrid, Margit och Henrik Höijers Donationsfond II and Alfred Kordelinin Yleinen Edistys-ja Sivistysrahasto.

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/sssaj2012.027.

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. http://dx.doi.org/10.1016/j.geoderma.2013.09.023.

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

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldshmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168, 1–13. https://doi.org/10.1002/jpln.200421414.

Bell, J.C., Cunningham, R.L., Havens, M.W., 1992. Calibration and validation of a soil-landscape model for predicting soil drainage class. Soil Sci. Soc. Am. J. 56, 1860–1866.

Bell, J.C., Cunningham, R.L., Havens, M.W., 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Sci. Soc. Am. J. 58, 464–470.

Bergmeir, C., Benítez, J.M., 2012. Neural networks in R using the Stuttgart neural network simulator: RSNNS. J. Stat. Softw. 46 (7), 1–26.

Beucher, A., Siemssen, R., Fröjdö, S., Österholm, P., Martinkauppi, A., Edén, P., 2015. Artificial neural network for mapping and characterization of acid sulfate soil: application to Sirppujoki River catchment, southwestern Finland. Geoderma 247-248, 38–50. https://doi.org/10.1016/j.geoderma.2014.11.031.

Bonham-Carter, G.F., 1994. Geographic information systems for geoscientists – modeling with GIS. In: Computer Methods in the Geosciences. 13 Pergamon, Oxford (398 p).

Boruvka, L., Penizek, V., 2007. A test of an artificial neural network allocation procedure using the Czech soil survey of agricultural land data. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science 31. Elsevier, Amsterdam, pp. 415–424.

Brady, N., 1990. The Nature and Properties of Soils, 10th ed. Macmillan Co., New York, NY.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Campling, P., Gobin, A., Feyen, J., 2002. Logistic modeling to spatially predict the probability of soil drainage classes. Soil Sci. Soc. Am. J. 66 (4), 1390–1401. http://dx.doi.org/10.2136/sssaj2002.1390.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma 195-196, 111–121. http://dx.doi.org/10.1016/j.geoderma.2012.11.020.

Chagas, C.d.S., Vieira, C.A.O., Filho, E.I.F., 2013. Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. Rev. Bras. Ciênc. Solo 37 (2), 339–351. http://dx.doi.org/10.1590/S0100-06832013000200005.

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. http://dx.doi.org/10.1016/j.geoderma.2016.03.025.

Chang, D.H., Islam, S., 2000. Estimation of soil physical properties using remote sensing and artificial neural network. Remote Sens. Environ. 74, 534–544. https://doi.org/10.1109/tgrs.2003.809935.

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

Danish Meteorological Institute, 1998. Danmarks Klima 1997. Danmarks Meteorologiske Institut, Copenhagen.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40 (2), 139–157.

ESRI, 2014. ArcGIS Desktop: Release 10.3. Environmental Systems Research Institute, Redlands, CA.

European Environment Agency, 2014. Corine Land Cover (CLC) 2012 - Denmark, Version 1, Oct. 2014 [dataset]. http://download.kortforsyningen.dk/content/corine-land-cover.

Garson, G.D., 1991. Interpreting neural network connection weights. Artif. Intell. Expert. 6 (4), 46–51.

Gershenfeld, N., 1999. The Nature of Mathematical Modelling. Cambridge University Press, Cambridge, pp. 356.

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.

Goh, A.T.C., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9 (3), 143–151.

Greve, M.H., Christensen, O.F., Greve, M.B., Bou Kheir, R., 2014. Change in peat coverage in Danish cultivated soils during the past 35 years. Soil Sci. 179 (5), 250–257. https://doi.org/10.1097/SS.0000000000000066.

Henriksen, H.J., Højberg, A.L., Olsen, M., Seaby, L.P., van der Keur, P., Stisen, S., Troldborg, L., Sonnenborg, T.O., Refsgaard, J.C., 2012. Klimaeffekter på hydrologi og grundvand - Klimagrundvandskort. Aarhus University.

Holte, R.C., Acker, L., Porter, B.W., 1989. Concept learning and the problem of small disjuncts. IJCAI 89, 813–818.

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

Jakobsen, P.R., Hermansen, B., Tougaard, L., 2015. Danmarks digitale jordartskort 1:25000 version 4.0. GEUS.

Kheir, R.B., Bøcher, P.K., Greve, M.B., Greve, M.H., 2010a. 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.

Kheir, R.B., Greve, M.H., Bocher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010b. 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. https://doi.org/10.1016/j.jenvman.2010.01.001.

Kravchenko, A.N., Bollero, G.A., Omonode, R.A., Bullock, D.G., 2002. Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. Soil Sci. Soc. Am. J. 66 (1), 235–243. https://doi.org/10.2136/sssaj2002.2350.

Kuhn, M., Weston, S., Coulter, N., Culp, M., Quinlan, J.R., 2015. Package 'C50'. https://cran.r-project.org/web/packages/C50/C50.pdf, Accessed date: 20 April 2017.

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.

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.

Lentzsch, P., Wieland, R., Wirth, S., 2005. Application of multiple regression and neural network approaches for landscape-scale assessment of soil microbial biomass. Soil Biol. Biochem. 37, 1577–1580. https://doi.org/10.1016/j.soilbio.2005.01.017.

Levine, E., Knox, R., Lawrence, W., 1994. Relationships between soil properties and vegetation at the northern experimental Forest, Howland, Maine. Remote Sens. Environ. 47 (2), 231–241.

Liu, J., Pattey, E., Nolin, M.C., Miller, J.R., Ka, O., 2008. Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. Geoderma 143 (3–4), 261–272. https://doi.org/10.1016/j.geoderma.2007.11.011.

Madsen, H.B., Jensen, N.H., 1988. Potentially acid sulfate soils in relation to landforms and geology. Catena 15, 137–145.

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

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Minasny, B., McBratney, A.B., 2002. The neuro-m method for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Am. J. 66, 352–361. https://doi.org/10.2136/sssaj2002.0352.

Mitchell, T., 1997. Decision tree learning. In: Machine Learning. McGraw Hill, New York, pp. 52–80.

Møller, A.B., Iversen, B.V., Beucher, A., Greve, M.H., 2017. Prediction of Soil Drainage Classes in Denmark by Means of Decision Tree Classification. (submitted for publication).

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. Int. J. Geogr. Inf. Sci. 16 (6), 533–549. https://doi.org/10.1080/13658810210138715.

Niang, M.A., Nolin, M., Bernier, M., Perron, I., 2012. Digital mapping of soil drainage classes using multitemporal RADARSAT-1 and ASTER images and soil survey data. Appl. Environ. Soil Sci. 2012, 1–17. https://doi.org/10.1155/2012/430347.

Nuutinen, V., Pöyhönen, S., Ketoja, E., Pitkänen, J., 2001. Abundance of the earthworm *Lumbricus terrestris* in relation to subsurface drainage pattern on a sandy clay field. Eur. J. Soil Biol. 37 (4), 301–304.

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.

Peng, W., Wheeler, D.B., Bell, J.C., Krusemark, M.G., 2003. Delineating patterns of soil drainage class on bare soils using remote sensing analyses. Geoderma 115 (3–4), 261–279. https://doi.org/10.1016/S0016-7061(03)00066-1.

Porwal, A., Carranza, E.J.M., Hale, M., 2003. Artificial neural networks for mineral potential mapping; a case study from Aravalli Province, Western India. Nat. Resour. Res. 12 (3), 155–171. https://doi.org/10.1023/A:1025171803637.

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

Quinlan, J.R., 1996. Learning decision tree classifiers. ACM Comput. Surv. 28 (1), 71–72.

Rokach, L., Maimon, O., 2005. Decision trees. In: Data Mining and Knowledge Discovery Handbook. Springer, pp. 165–192.

SAGA GIS, S. System for Automated Geoscientific Analyses http://www.saga-gis.org.

Silveira, C.T., Oka-Fiori, C., Santos, L.J.S., Sirtoli, A.E., Silva, C.R., Botelho, M.F., 2013. Soil prediction using artificial neural networks and topographic attributes. Geoderma 195-196, 165–172. https://doi.org/10.1016/j.geoderma.2012.11.016.

Smith, K.A., Ball, T., Conen, F., Dobbie, K.E., Massheder, J., Rey, A., 2003. Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes. Eur. J. Soil Sci. 54 (4), 779–791. https://doi.org/10.1046/j.1351-0754.2003.0567.x.

Tan, P.-N., Steinbach, M., Kumar, V., 2014. Classification: basic concepts, decision trees, and model evaluation. In: Introduction to Data Mining. 2014. Pearson Education, Limited, pp. 145–205.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Hermann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., Wieland, J., 1998. SNNS Stuttgart Neural Network Simulator User Manual, Version 4.2. IPVR, University of Stuttgart and WSI, University of Tübingen. http://www.ra.cs.uni-tuebingen.de/SNNS/.

Zhao, Z.Y., Chow, T.L., Yang, Q., Rees, H.W., Benoy, G., Xing, Z.S., Meng, F.R., 2008. Model prediction of soil drainage classes based on digital elevation model parameters and soil attributes from coarse resolution soil maps. Can. J. Soil Sci. 88 (5), 787–799. https://doi.org/10.4141/CJSS08012.

Zhao, Z.Y., Ashraf, M.I., Meng, F.-R., 2013. Model prediction of soil drainage classes over a large area using a limited number of field samples: a case study in the province of Nova Scotia, Canada. Can. J. Soil Sci. 93 (1), 73–83. https://doi.org/10.4141/cjss2011-095.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resour. Res. 36, 663–677. https://doi.org/10.1029/1999WR900315.