Machine learning techniques for acid sulfate soil mapping in southeastern Finland

Estévez, Virginia; Beucher, Amélie; Mattbäck, Stefan; Boman, Anton; Auri, Jaakko; Björk, Kaj Mikael; Österholm, Peter

Published in:
Geoderma

DOI:
10.1016/j.geoderma.2021.115446

Published: 15/01/2022

Document Version
Final published version

Document License
CC BY

Geoderma 406 (2022) 115446

Contents lists available at ScienceDirect

Geoderma

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

Machine learning techniques for acid sulfate soil mapping in southeastern Finland

Virginia Estévez [a,*], Amélie Beucher [b], Stefan Mattbäck [c,d], Anton Boman [d], Jaakko Auri [e], Kaj-Mikael Björk [a], Peter Österholm [c,*]

[a] Arcada University of Applied Sciences, Jan-Magnus Janssonin aukio 1, 00550 Helsinki, Finland
[b] Department of Agroecology, Aarhus University, Blichers Allé 20, PO Box 50, 8830 Tjele, Denmark
[c] Geology and Mineralogy, Åbo Akademi University, Domkyrkotorget 1, 20500, Åbo, Finland
[d] Geological Survey of Finland, PO Box 97, 67101 Kokkola, Finland
[e] Geological Survey of Finland, PO Box 96, 02151 Espoo, Finland

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Acid sulfate soils
Soil probability mapping
Machine learning
Random forest
Gradient boosting
Support vector machine

ABSTRACT

Acid sulfate soils are one of the most environmentally harmful soils existing in nature. This is because they produce sulfuric acid and release metals, which may cause several ecological damages. In Finland, the occurrence of this type of soil in the coastal areas constitutes one of the major environmental problems of the country. To address this problem, it is essential to precisely locate acid sulfate soils. Thus, the creation of occurrence maps for these soils is required. Nowadays, different machine learning methods can be used following the digital soil mapping approach. The main goal of this study is the evaluation of different supervised machine learning techniques for acid sulfate soil mapping. The methods analyzed are Random Forest, Gradient Boosting and Support Vector Machine. We show that Gradient Boosting and Random Forest are suitable methods for the classification of acid sulfate soils, the resulting probability maps have high precision. However, the accuracy of the probability map created with Support Vector Machine is lower because this method overestimates the non-AS soils occurrences. We also compare these modeled probability maps with the conventionally produced occurrence map. In general, the modeled maps are more objective and accurate than the conventional maps. Moreover, the mapping process using machine learning techniques is faster and less expensive.

1. Introduction

Acid sulfate (AS) soils are one of the most environmentally harmful soils existing in nature. These soils contain sulfidic materials, which under oxidizing conditions produce sulfuric acid and relase metals. This can lead to several ecological damages. Generally, AS soils are defined as soils where the soil pH has dropped or may drop below 4 due to oxidation of sulfidic material (Pons, 1973). The lowered soil-pH leads to acidification of the soil and mobilization of metals (Åström and Björklund, 1997; Roos and Åström, 2006; Österholm and Åström, 2002). This potentially toxic combination often ends up in watercourses causing serious damages to the living water organisms (e.g. fish kills Hudd, 2000; Urho, 2002). This type of soil usually appears in coastal regions and in freshwater wetlands. The estimate of AS soils worldwide is approximately 500,000 km^2 (Michael et al., 2017). The highest incidences of AS soils are located in Australia, Asia, Latin America and Africa. In Europe, the largest occurrences of AS soils are in Finland (Andriesse and van Mensvoort, 2006). So far, the extent of AS soils in Finland is unclear. A first study shows up to 3,360 km^2 located on the coastal plains (Palko, 1994), whereas a later work estimates an extent ranging from 480–1,300 km^2 (Yli-Halla et al., 1999). Since 2009, the Geological Survey of Finland (GTK) has made a great effort in localizing AS soils areas along the coastal plains of Finland, with extensive sampling and the consideration of new criteria for the classification of this type of soil (Boman et al., 2019). As a result, a substantial increase of the AS soil covered areas is expected.

In Finland, the soil materials that form AS soils have a sedimentary origin, and consist of mineral and organic soil materials that contain sulfide minerals. Although nowadays the sulfidization continues in the sediments of the current Baltic Sea (Jokinen et al., 2018; Jokinen et al., 2020), most of the sulfidic sediments were formed in the preceding Littorina Sea (Fig. 1) (Palko, 1994; Yli-Halla et al., 1999; Kivinen, 1950;

* Corresponding authors.
E-mail address: estevezv@arcada.fi (V. Estévez).

https://doi.org/10.1016/j.geoderma.2021.115446
Received 31 March 2021; Received in revised form 27 August 2021; Accepted 28 August 2021
Available online 24 September 2021

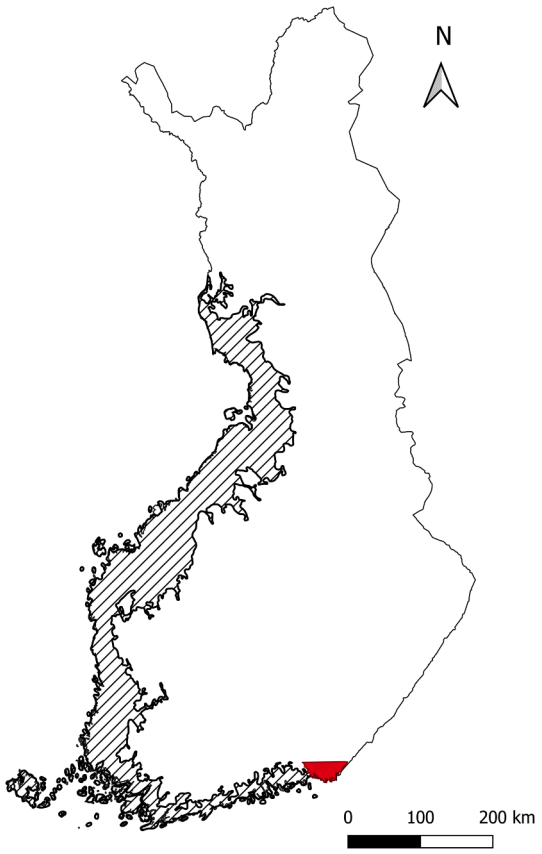

Fig. 1. Location of the study area (red color) and the maximal extent of the Littorina Sea (diagonal lines), where acid sulfate (AS) soils are usually encountered in Finland.

Purokoski, 1959; Erviö, 1975; Puustinen et al., 1994). After the glaciation, the isostatic land uplift has led to these sediments emerging above the current sea level. Thus, the sediments are subjected to oxidizing conditions when drained by anthropogenic actions (e.g. ditching in agriculture and forestry). The largest areas of AS soils in Finland are consequently postglacial fine-grained sediments located in coastal areas (Fig. 1). This so called Littorina Sea area is about 51,000 km^2.

AS soils strongly affect the contamination of watercourses, agriculture and its productivity, or the construction of infrastructures (Michael, 2013). Moreover, some studies have shown that crop products (Palko, 1986) and cow milk (Alhonen et al., 1997) in AS soil areas have high concentrations of some metals. This could potentially pose a high risk to human health. Thus, AS soils constitute one of the major environmental problems in some countries. In order to mitigate the possible ecological damages created by this type of soil, it is necessary to map the area covered by AS soils. At present, there are few AS soil maps as well as many possible AS soil areas unknown worldwide. In Finland, although the general characteristics of AS soils have been widely studied (e.g. Yli-Halla et al., 1999; Yli-Halla, 1997; Österholm and Åström, 2004; Roos and Åström, 2005; Edén et al., 2012a; Fältmarsch et al., 2008; Toivonen et al., 2013), there are less studies in AS soil mapping (Palko, 1994; Edén et al., 2012b). This is mainly due to the fact that traditional or conventional methods, based on soil sampling and subsequent laboratory analysis, are very laborious and time-consuming. Nowadays, machine learning can be used to create maps of AS soil occurrence. Machine learning techniques can streamline the mapping process as well as improve its accuracy. Furthermore, less sampling is needed for the mapping (Brus et al., 2011). The use of these techniques allows the combination of soil observations and environmental data for the creation of soil maps. So far, several machine learning techniques have been applied in digital soil mapping (DSM) (McBratney et al., 2003), among the most used are Random Forest (RF) and Artificial Neural Network (ANN). In Finland, machine learning techniques such as ANN (Beucher et al., 2013; Beucher et al., 2015) and Fuzzy Logic (Beucher et al., 2014) have been used for AS soil mapping. ANN has displayed substantial predictive classification abilities for AS soil mapping at catchment scale (Beucher et al., 2013; Beucher et al., 2015), as well as promising results for characterizing different soil properties (Beucher et al., 2015). Although ANN is a suitable method, there are several machine learning techniques that have never been applied in AS soil mapping that could be more appropriate for this purpose. Thus, more advanced studies are required for the mapping of AS soils which represents a crucial subject in many parts of the world. At the moment, very few studies exist and they were only presenting one method. In this work, we have compared three different methods which constitutes a step forward on the topic. This comparison will enable the selection of the most suitable method for a precise classification of AS soils and the creation of accurate maps.

The main goal of this work is to evaluate different supervised machine learning techniques for the classification of AS soils and prediction of their spatial occurrence in a catchment area located in southeastern of Finland (Fig. 1). The study is a binary classification of the soils into AS and non-AS soils. The methods analyzed are Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM). To our knowledge, GB and SVM have never been used for AS soil mapping. Furthermore, the probability AS soil maps for the study area are created from these methods. Finally, we compare the AS soil probability maps generated by means of machine learning techniques with the conventionally produced probability map in the study area.

2. Material and methods

2.1. Study area

The study site is located in southeastern Finland and corresponds to the coastal area of Virolahti and its surroundings (Fig. 1). This area is part of the boreal ecosystem, and land use corresponds predominantly to forestry, agricultural lands and some urban areas. The total study area is approximately 1,091 km^2, out of which ~ 39 km^2 is water (3.19% of the area). The Littorina Sea maximum extent is 83% of the study area (~ 905 km^2). The geological basement consists almost entirely of 1.66 – 1.60 Ga Rapakivi granite (Geological Survey of Finland, 2021a; Lehtinen et al., 1998). The basement is covered by glacial till and alluvial deposits (Haavisto-Hyvärinen and Kutvonen, 2007). The main part of the uppermost meter in the area consists of bedrock, outcrops and block fields (57.66 %). The soil types are clay (16.91%), fine sand to gravel (7.21%), till (5.85%), thick peat deposits (4.63%), gyttja (2.01%), fine-grained sediment or fine silt low humus content 2–6 % (1.21%), fine silt (1.12%), and man made soils. A small fraction of the area is unmapped (0.22%).

2.2. Soil Samples

The soil samples or cores were collected for the national AS soil mapping by the Geological Survey of Finland (GTK). In the study area, the soil cores down to 2–3 m depth were collected using a gouge auger. For the locations of the soil cores a non-statistical sampling design was used. The sampling plan was designed to cover the landscape as evenly as possible with a density about 1 probe/km^2, and in the way that all classes of soils and materials of the study area were part of the set of samples. For this, all sediment texture classes in quaternary sediment maps, anomalies and non-anomalies (EC) in airborne geophysical data and different positions in the topography were considered. The sampling density is less dense in some areas due to the restricted road network, and the exclusion of glacial till, bedrock, outcrops, water and man made soils from the sampling. Moreover, sampling was restricted to the Littorina Sea maximal extent since the majority of the Finnish AS soils are

considered to occur here (Geological Survey of Finland, 2021b; Yli-Halla et al., 1999; Palko, 1994). Depending on the value of soil-pH, these soil cores were classified into AS and non-AS soils.

The classification of AS soils is based on the presence of specific soil materials such as sulfuric or hypersulfidic materials. The presence of these materials was indicated by pH-measurements. The field-pH was measured at the sampling site by inserting a pH-electrode (Hamilton Flatrode) directly into the soil core. The pH measurements were made every 20 cm for two purposes: identification of the lowest pH values in the soil core and the oxidation depth. A second measurement of the pH was made in the laboratory after 8–19-weeks of incubation of collected soil samples. During this oxidation period, the soil materials have the possibility to react under atmospheric oxygen at room temperature (Creeper et al., 2012). The soil cores were classified as AS soils according to international classification systems (IUSS Working Group WRB, 2015; Sullivan et al., 2010), i.e. if the field-pH was lower than 4 (indicating sulfuric material), and/or if the incubation-pH was lower than 4 (indicating hypersulfidic material) and the pH-drop during incubation was at least 0.5 pH-units.

Soil samples play a fundamental role in the creation of AS soils maps, both for the conventional method as well as for machine learning techniques. In the conventional case, the proportion of AS soil samples compared to the total number of soil samples for a given area determines the AS soil probability of that area. However, since the sampling locations were not selected through a probability sampling design, some bias may occur when determining the AS soil probability in the conventional map. In the case of machine learning methods, the soil samples are used for training and validating the models.

2.3. Environmental data

The environmental or raster data have been created from remote sensing data using Qgis software (QGIS Development Team, 2019). For this study, different types of environmental covariates have been used: Quaternary geology, digital elevation model, terrain data and aero-geophysical layers. These environmental covariates are critical to localize AS soils, as these soils are present in some quaternary geology classes, they occur in flat and low-lying areas, or they can show strong

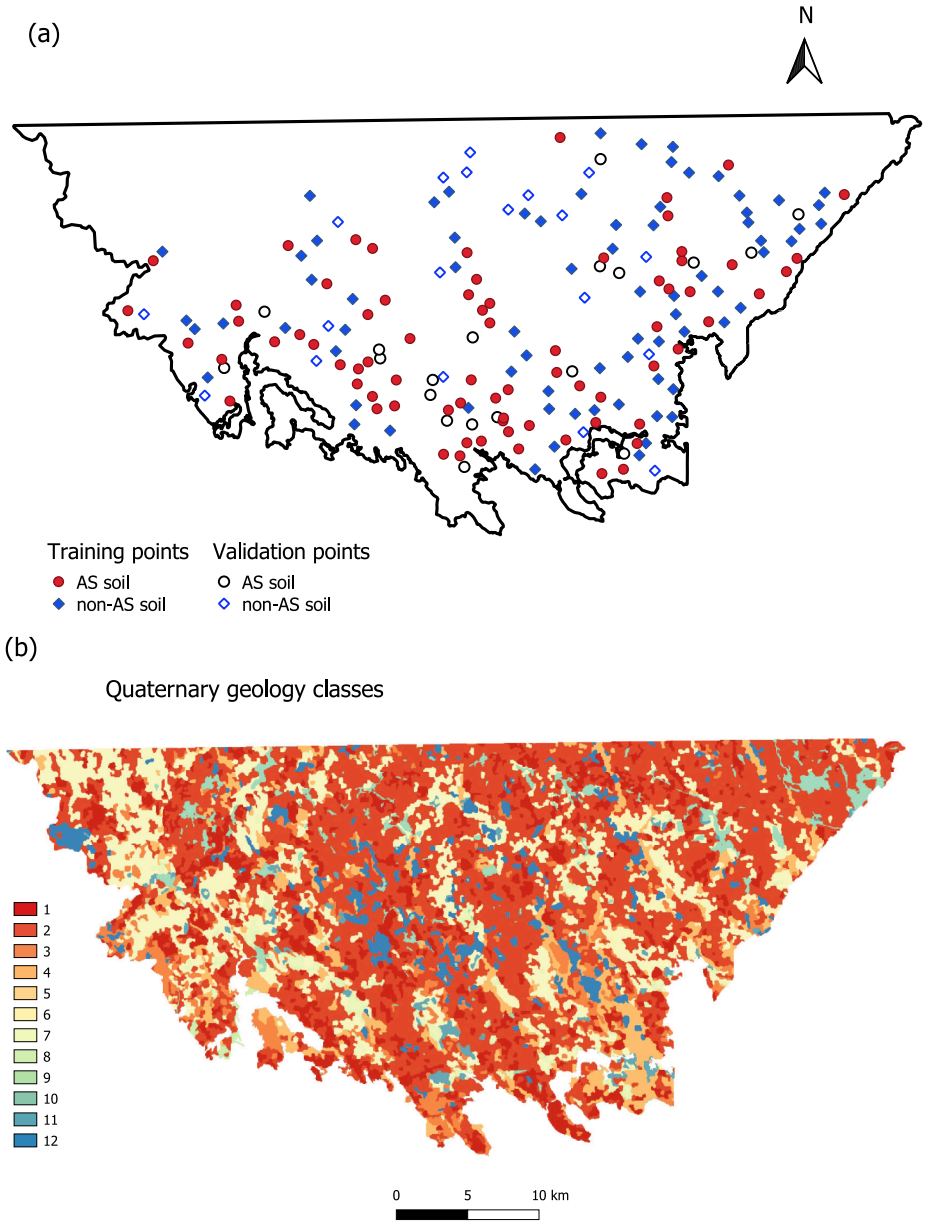

Fig. 2. (a) Training and validation points in the study area. (b) Quaternary geology classes.

electromagnetic anomalies.

2.3.1. Quaternary geology

The quaternary geology layer is composed of 12 classes (Fig. 2 (b)). The quaternary geology classes can be seen in Table 1. Finnish AS soils are typically fine-grained (i.e., containing clay, gyttja and fine silt), but may in certain settings be composed of coarse-grained soil materials such as sand (Mattbäck et al., 2017). The original Quaternary geology map (1:200 000) has been produced by GTK (Korpela and Niemelä, 1985).

2.3.2. Digital elevation model

The digital elevation model (DEM) has been created from airborne laser scanning data (i.e. Light Detection And Ranging, LiDAR data) produced by the National Land Survey of Finland (NLS). This layer has a cell size of 2 m x 2 m. Fig. 3 (a) shows the DEM for the study area. The DEM layer is fundamental in the detection of AS soils. In Finland, this type of soil typically appears in areas with an elevation below 50 m (Palko, 1994).

2.3.3. Slope

Slope is a terrain attribute derived from the DEM. This layer is in degrees, and has been calculated following the method of Horn (Horn, 1981) and a vertical/horizontal ratio equal to one. The slope layer is represented in Fig. 3 (b), and enables locating very low-relief areas, such as plains, swamps and river valleys where AS soils typically occur.

2.3.4. Aerogeophysics

In this work, aerogeophysics covariates refer to airborne electromagnetic data (Airo et al., 2014). These data are very useful for the classification of soil materials. Aerogeophysical data were provided by GTK: low electromagnetic frequency (3 kHz) imaginary and real components which were derived from high resolution low altitude airborne geophysics (flight altitude from 30 to 40 m and line spacing mainly 200 m). Figs. 4 (a) and 4 (b) show the real and imaginary components of the electromagnetic induction, respectively. The imaginary component enables detecting shallow weak anomalies (mainly related to variations in topsoil thickness and/or electric conductivity). Whereas the real component indicates anomalies originating deep in the bedrock (e.g. from black schists which are associated with sulfide deposits and might cause high metal contents in soil or groundwater Airo and Loukola-Ruskeeniemi, 2004). Sulfide bearing sediments yield strong electromagnetic anomalies (i.e., high electric conductivity values) because of their high contents of soluble salts (Suppala et al., 2005; Vanhala et al., 2004), and appear as diffuse and round-shaped anomalies. Black schists appear as thin and elongated, high electric conductivity anomalies, however, they do not occur in the study area.

2.4. Conventional AS soil probability map

The conventional AS soil probability map of the study area was drawn by hand in a GIS software using all the sampled soil cores together with environmental and legacy data. The source materials are: 1) Classified soil cores (primary material), 2) Soil maps (1:20 000 and 1:200 000; GTK), 3) Peat survey data (GTK), 4) Low-altitude airborne geophysics (GTK) including a) Imaginary component of electrical conductivity (3 kHz), b) Real component of electrical conductivity (3 kHz) and c) Apparent characteristic resistivity of electrical conductivity (3 kHz), 5) Terrain database of the NLS and 6) Laser scanning data of the NLS. The minimum extent of a drawn area was approximately 6 hectares.

Table 1
Quaternary geology classes.

Nº	Quaternary geology classes
1	Bedrock
2	Outcrops
3	Till
4	Fine sand to gravel
5	Cobbles and boulders
6	Fine silt
7	Clay
8	Gyttja
9	Unmapped area
10	Water
11	Fine silt with low humus
12	Thick peat deposits

The conventional AS soil map is represented in Fig. 5, which shows four different probability classes: high (98.5%), moderate (52.5%), low (1.7%) and very low (0%) probability of encountering AS soils. The calculation of the probability of encountering AS soils in a given area was based on the proportion of sampling points classified as AS soils compared to the total number of sampling points located in that area according to the methods used in Denmark (Madsen and Jensen, 1988). The extent of the conventional probability map is $\sim$ 905 km^2, where 2.7% of the total area corresponds to the high class, 8.7% to the moderate class, 23% to the low class and 65.6% to the very low class.

2.5. Machine learning methods

2.5.1. Random forest

Random Forest (Breiman, 2001) is a supervised machine learning technique widely used in soil science, both in prediction of soil properties (Behrens et al., 2010; Grimm et al., 2008; Lieβ et al., 2012; Schmidt et al., 2014; Wiesmeier et al., 2011) and classification of soils (Gambill et al., 2016; Teng et al., 2018; Brungard et al., 2015; Heung et al., 2014; Heung et al., 2016). This method is effective and gives predictions with high accuracy. This ensemble method is based on decision trees, and makes the prediction taking into account the results of many decision trees, which have been created on data randomly selected. In this way, the overfitting is reduced.

2.5.2. Gradient boosting machines

Gradient Boosting Machines is one of the most powerful machine learning techniques for classification and regression problems. Similarly to RF, GB is also an ensemble method based on decision trees (Friedman, 2001). Contrary to RF, this method generates the trees serially. In this way, each tree tries to improve the prediction by correcting the errors of the previous one. As a result, the prediction of the model improves. This method is beginning to be used more frequently also in soil science, showing promising results. So far, GB has been used for the prediction of soil classes (Lemercier et al., 2012) and properties (Sindayiheburaa et al., 2017; Tziachrisa et al., 2019; Hengl et al., 2017). In this work, we evaluate the abilities of GB to classify AS soils.

2.5.3. Support vector machine

As in the case of previous methods, Support Vector Machine is a supervised machine learning technique used for classification and regression predictions (Cortes and Vapnik, 1995; Vapnik, 1995). This method is widely used for binary classification. In the case of classification, this technique tries to separate the two classes by a line, a plane or a hyperplane. SVM has been used in soil science for predicting soil classes and properties (Brungard et al., 2015; Heung et al., 2016; Kovačevic et al., 2010; Gill et al., 2006) and developing pedotransfer functions (Lamorski et al., 2008). So far, this method has never been used for the classification of AS soils.

2.5.4. Model and tuning parameters

In this study, the library used for the machine learning methods is Scikit-learn (Pedregosa et al., 2011) in Python (Van Rossum and Drake, 2009). For the modeling, it is fundamental to take into account the parameters of the machine learning methods (Müller et al., 2016). The consideration of these parameters can improve the performance of the

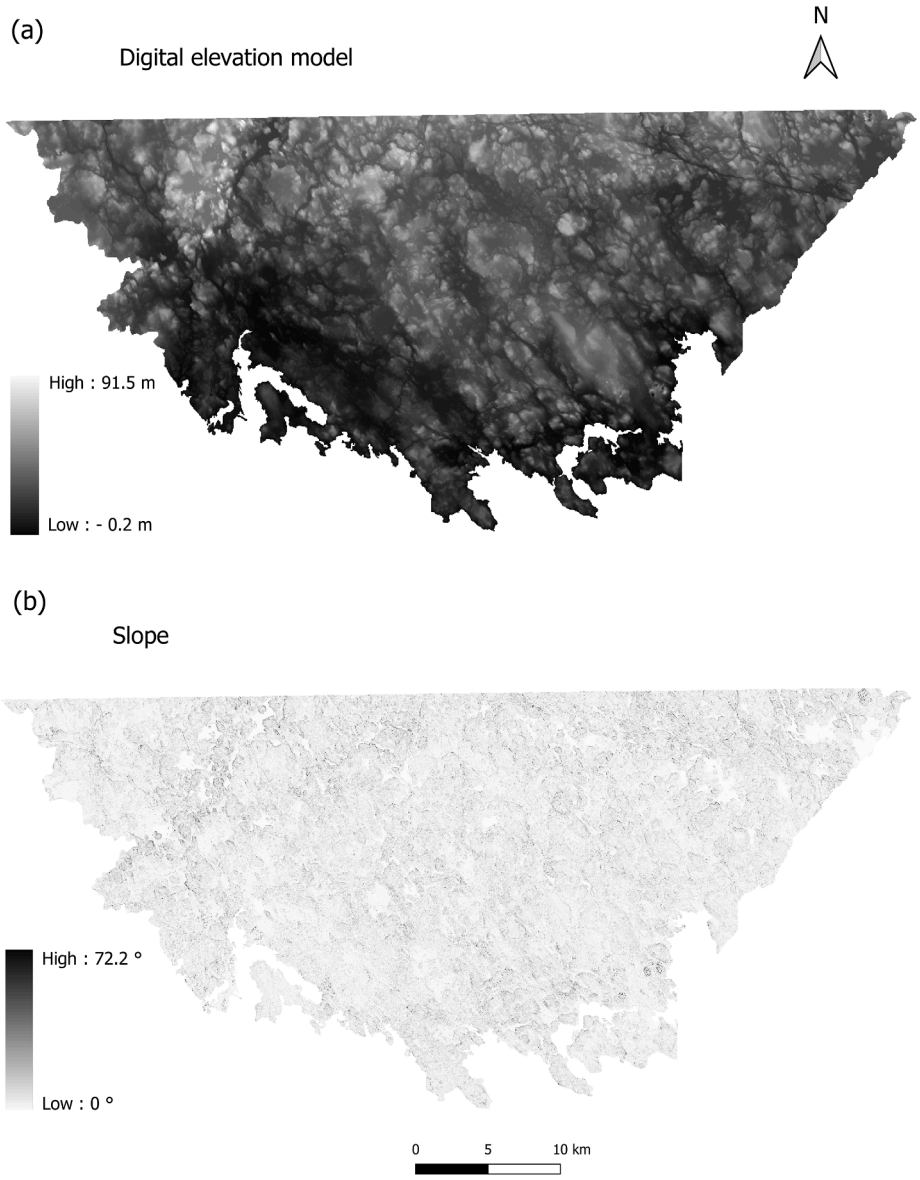

Fig. 3. (a) Digital elevation model. (b) Slope.

models. As all the codes for the models have been written in Python, the parameters for each machine learning method are in the corresponding nomenclature. In the case of RF, the most important parameters are *n_estimators*, *max_features* and *max_depth*. The *n_estimators* is the number of trees of the model. In general, a larger number of trees will contribute to reduce the overfitting. Typical values are in the range from 10 to 10,000, although it will depend on the dataset. The *max_features* is the number of features taken into account for each split in the trees. A small value reduces overfitting. A typical value for classification is the square root of the number of features. The maximum depth of each tree is controlled by the *max_depth*. This parameter can help to reduce the complexity of the trees. For GB, the important parameters are *n_estimators* and *max_depth*. Unlike RF, in GB a large number of trees can lead to overfitting. The values of *max_depth* are quite low for GB, normally between 1 and 5. In the case of SVM, as this method is based on a very different algorithm, the parameters are also very different to the previous parameters. In this model the important parameters will depend on the kernel function, which can also be chosen. Most typical kernels are *linear* and *radial basis function(rbf)*. For the *linear* kernel, the important parameter is *C*, whereas for *rbf* there are two, *C* and *gamma* parameters. *C* is a regularization parameter, which controls the misclassification of the samples with respect to the hyperplane that separates the two classes. For high values of *C*, the model tries to correctly classify all samples, while for low values, the model is more tolerant to the misclassification of some samples. Typical parameters for *C* are from 0.001 to 100. The *gamma* parameter controls the width of the *rbf* kernel, i.e., it determines the clustering of the samples. For low values of *gamma*, the radius of the kernel is large and many points are clustered. On the contrary, high *gamma* values give rise to groups where only the very close points are included. This can lead to overfitting. Typical values for *gamma* are from 0.0001 to 10. In the case in which both parameters have to be taken into account, a high value of *gamma* will reduce the importance of *C* in the model.

For all machine learning methods, depending on the selection of the parameters, the results obtained with a model for a given dataset can be very different. For this reason, it is very important to find the parameters that lead to the best performance of the model. However, finding these parameters is a difficult task. In this work, the parameters for the three machine learning techniques have been selected using grid search with cross-validation (GridSearchCV). This method analyzes all possible combinations of the parameters, selecting the one which performs better. The selection is based on the best cross-validation score.

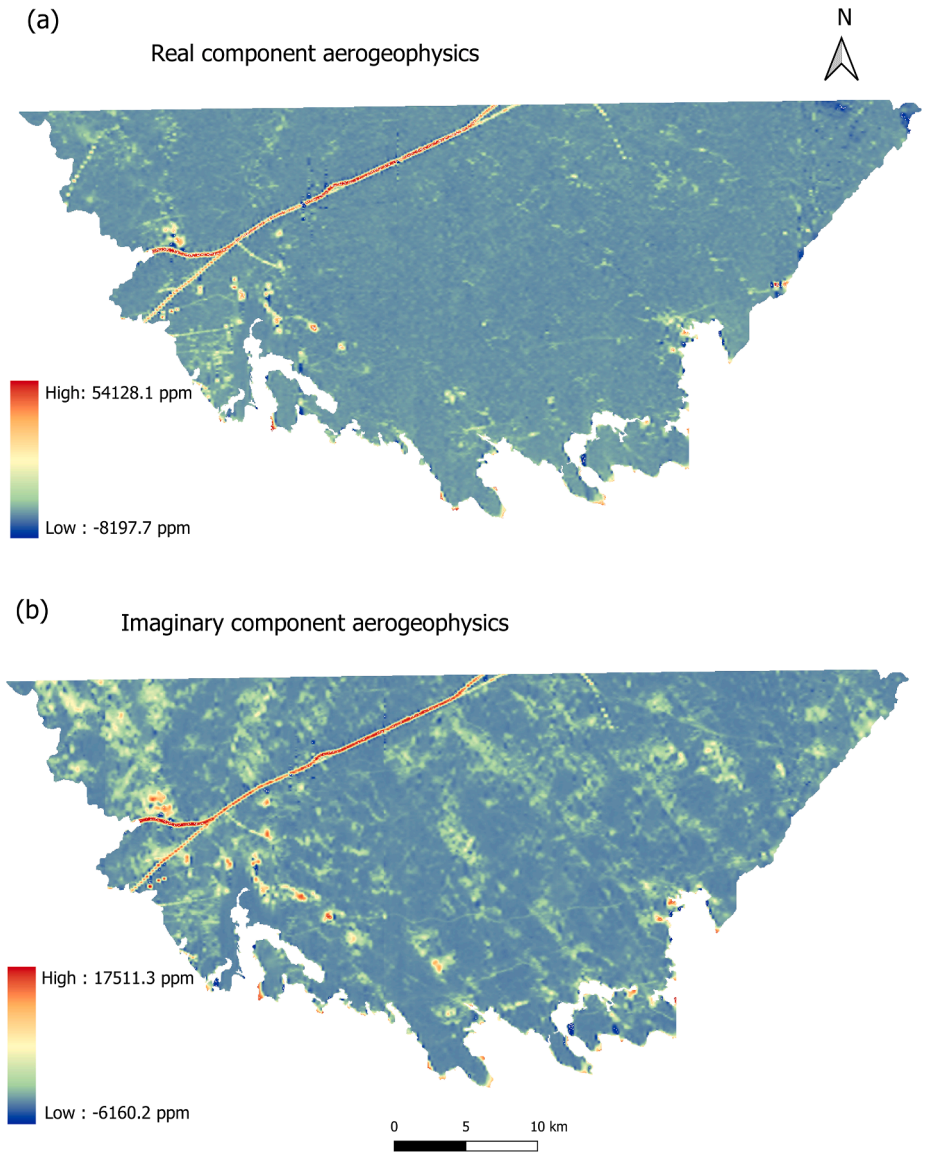

Fig. 4. Aerogeophysical covariates of the study area. (a) Real component aerogeophysics. (b) Imaginary component aerogeophysics.

2.6. Data pre-processing

2.6.1. Training and validation points

Machine learning methods perform better when the data are balanced (Weiss and Provost, 2001; Porwal et al., 2003; Wei and Dunbrack, 2013). In our case, this means equal number of AS and non-AS soil samples in the training set. As in the dataset the number of non-AS soil samples is two times larger than the one of AS soils, the non-AS soil samples have been randomly selected. As a result, 186 soil samples have been considered in this study, 93 for each class. For the modeling, the soil samples have been randomly split into two parts, one for training the model (80%) and the other for its validation (20%). Then, the training set consists of 148 soil samples, 74 for each class, and the validation set of 38 points, 19 for each class, see Fig. 2 (a) where the points are represented. The same soil cores have been used for the evaluation and comparison of the three machine learning methods as well as for the validation of the probability maps created from the models.

2.6.2. Environmental data

For the modeling, the five covariate layers have been pre-processed to the same resolution, a grid size of 50 m x 50 m. Moreover, all covariate layers have the same coordinate reference system, which is the one corresponding to Finland (ETRS89/TM35FIN(E,N)).

2.7. Modeled AS soil probability maps

The modeled AS soil probability maps are the maps created from the machine learning methods. Once the machine learning models have been trained with the soil cores and their corresponding values of the environmental covariates, they are able to predict and classify the AS soils. The probability maps for AS soil occurrence are generated from the predictions made by the models. The predictions have been calculated taking into account the values of the covariates for each cell of the study area, which has been modeled by 50 m x 50 m cells (in total 434,145 cells). The model predicts the probability of encountering AS soils in each cell. The probability values have been classified in four classes, [0–0.25), [0.25–0.5), [0.5–0.75) and [0.75–1], which correspond to very low, low, high and very high probability, respectively.

2.8. Metrics for the evaluation of the models

In order to know if a method is appropriate for the classification of AS soils, different metrics can be used. We have considered the metrics related to the confusion matrix, which are typical metrics used in binary

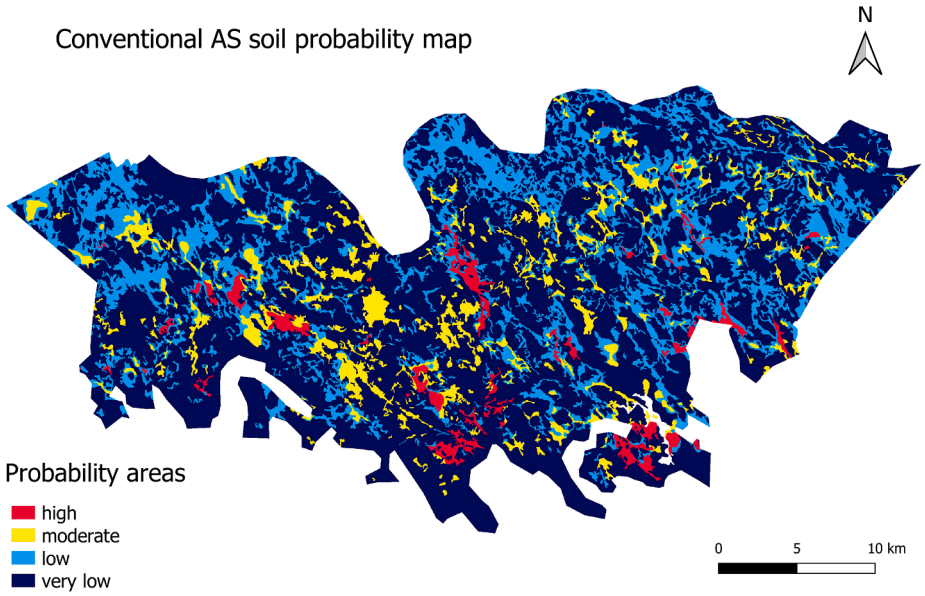

Fig. 5. Conventionally produced probability map of acid sulfate (AS) soils. A high probability of encountering AS soils is indicated by a red color, whereas moderate, low and very low probabilities are colored yellow, blue and dark blue, respectively.

classification (Powers, 2011). These metrics are precision, recall and F1-score. The precision is the percentage of samples correctly predicted for a given class with respect to the total number of samples predicted for that class. Whereas the recall or sensitivity is the proportion of samples correctly predicted for a given class. This metric is also known as True Positive Rate. For a correct interpretation of the suitability of a model, both precision and recall have to be taken into account. Only the consideration of the precision can lead to misinterpretations. The F1-score is a combination of the precision and the recall, and it is given by the following formula

$$F1 - score = 2\left(\frac{precision * recall}{precision + recall}\right) \quad (1)$$

This metric shows how the model makes the prediction for each class. The higher the value of F1-score, the better the model will work for a given class. These metrics provide enough information to determine the suitability of the methods.

3. Results and discussion

3.1. Evaluation of the machine learning methods

In this study, three machine learning methods have been analyzed for AS soil mapping. The results obtained for the three methods can be seen in Table 2, where the metrics related to the confusion matrix are represented.

Table 2
Metrics related to the confusion matrix for Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM). The classes are acid sulfate (AS) and non acid sulfate (non-AS) soils.

Method	Class	Precision	Recall	F1-score
RF	non-AS	0.72	0.68	0.70
	AS	0.70	0.74	0.72
GB	non-AS	0.78	0.74	0.76
	AS	0.75	0.79	0.77
SVM	non-AS	0.59	0.84	0.70
	AS	0.73	0.42	0.53

In a binary classification, the suitability of a method will be determined by its ability to properly classify the two classes. As it was already mentioned in subSection 2.8, the precision and the recall have to be considered at the same time for a good interpretation of the results. High values of precision and recall for a given class mean that the model is able to predict and classify this class adequately. For RF and GB, the values of the precision and recall are high for both classes. This means that RF and GB can successfully distinguish both classes, leading to good results for this study. A more accurate classification is obtained with the GB method. The results improve between 5% and 6% with respect to RF.

On the contrary, the results for the SVM method are very different. The first thing that can be observed is the difference between the metrics for both classes. This indicates that the model is not working properly for one of the classes, the AS soils. The model is able to correctly classify 84% of the non-AS soil samples, but only 42% of the AS soils. Looking at the values of the precision it can be seen that only 59% of the non-AS soil samples predicted are actually non-AS soils. Then, the remaining 41% of the predicted non-AS soils are wrong. This means that although this method is very good for classifying non-AS soils, at the same time is considering a large number of AS soil samples as non-AS soils. This explains the low proportion of AS soils properly classified. As a result, this method overestimates the non-AS soil class. On the other hand, the high precision and low recall for AS soils confirm that the model hardly predicts this class, but is highly reliable when it does. Thus, SVM does not work well for the classification of AS soil samples.

A better idea about how the model works for each class is given by the F1-score. For the RF and GB models, the F1-score is quite similar for each class. There is a difference of 2% between the classes for RF, and 1% for GB. This balance between the classes, as well as the high values of F1-score, show that these models work very well for both classes. However, the results obtained with SVM confirm that this method does not perform properly for both classes. As it can be seen in Table 2, the values of the F1-score are very different for each class, with a difference of 17%. Although the F1-score for the non-AS class is good and similar to the value obtained with RF, the F1-score for the AS class is much too low. Moreover, a value of 53% for AS soils indicates that this model is predicting this class almost randomly. As a result, the predictions and classifications made by this model will have a considerable error, which will lead to AS soil maps with low accuracy. Thus, the results display that RF and GB are good methods for the classification of AS soils,

whereas SVM is not valid because it is not able to classify correctly one of the classes.

Previous works, where different machine learning methods have been compared for predicting soil classes (Brungard et al., 2015; Heung et al., 2016) and soil textures (Lieβ et al., 2012), have also shown that RF generally gives the most accurate results. In the case of SVM, there are some works that have obtained satisfactory accuracy for classification of soils using this method (Heung et al., 2016; Kovačevic et al., 2010). However, it should be noted that in the case of a binary classification, SVM separates the two classes using a hyperplane. This type of separation can be inappropriate for certain types of datasets. In our case, the poor results obtained with SVM for the classification of AS soils could be related with the type of environmental covariates used in the study. In general, the performance of a machine learning model depends on the data. Depending on the relationship between the features and the label or output response some machine learning models will perform better than others (James et al., 2013; Kuhn and Johnson, 2013). For example, if there is a linear relationship, linear models will fit well. However, if the relationship is non-linear and complex, methods based on decision trees such as RF or GB may perform much better than the linear models. It has been shown that RF performs better than the linear models when the relationship is non-linear (Hengl et al., 2015) or that SVM may outperform RF when the relationship is linear (Statnikov et al., 2008). As in our case RF and GB models perform better than SVM, this indicates that the relationship between the environmental covariates (features) and the AS soils (label) is complex and non-linear.

3.2. *Probability maps created from the machine learning methods*

Once the suitability of the machine learning methods for AS soil classification has been evaluated, we have created the modeled AS soil probability maps. Fig. 6 shows the probability maps for the three models, whereas their corresponding % of the study area for each probability class is represented in Table 3. As it can be seen, the probability maps created from RF and GB show greater heterogeneity than the map generated from SVM, where the most of the study area is predicted as low probability for AS soils. As it was already shown in the previous subsection, this model has problems distinguishing the AS soils, and tends to consider most of the samples as non-AS soils. This is reflected in the probability map with the overestimation of areas with low probability for AS soils. Other surprising thing of this method is that 84% of the study area is predicted as low probability and only 1% as very low. As this model classifies very well the non-AS soils, it was expected that the very low probability area was much larger. Although to a lesser extent, something similar occurs with the high and very high probability areas (Table 3). In general, the model is classifying the most of the cells between only two areas, those with low and high probability. This is another weak point of this method in the classification of AS soils. Unlike SVM, RF and GB predict the study area for the four different probabilities areas, leading to more realistic probability maps. And as we will see with greater accuracy.

Other thing that attracts attention is the linear feature showed up in the three maps (Fig. 6). This line appears in the aerogeophysics layers (Figs. 4)) and is related to power lines. The probability of encountering AS soils in this linear feature is very high in the map created by SVM, but low in the cases of RF and GB. This difference may be due to the importance of the each environmental covariate in the models. In the aerogeophysics layers, the line corresponding to the power line is giving information about the power line, but not of the soils in that area. This can lead to wrong predictions in that area if the importance of these covariates is very high for the model. However, the other covariates considered (Quaternary, DEM and slope) provide information about the soils in the line. Thus, if one or several of these covariates are the most relevant for the model, the prediction for the line will be more accurate.

The validation of the probability maps has been done with the same validation points used in the evaluation of the models. These validation

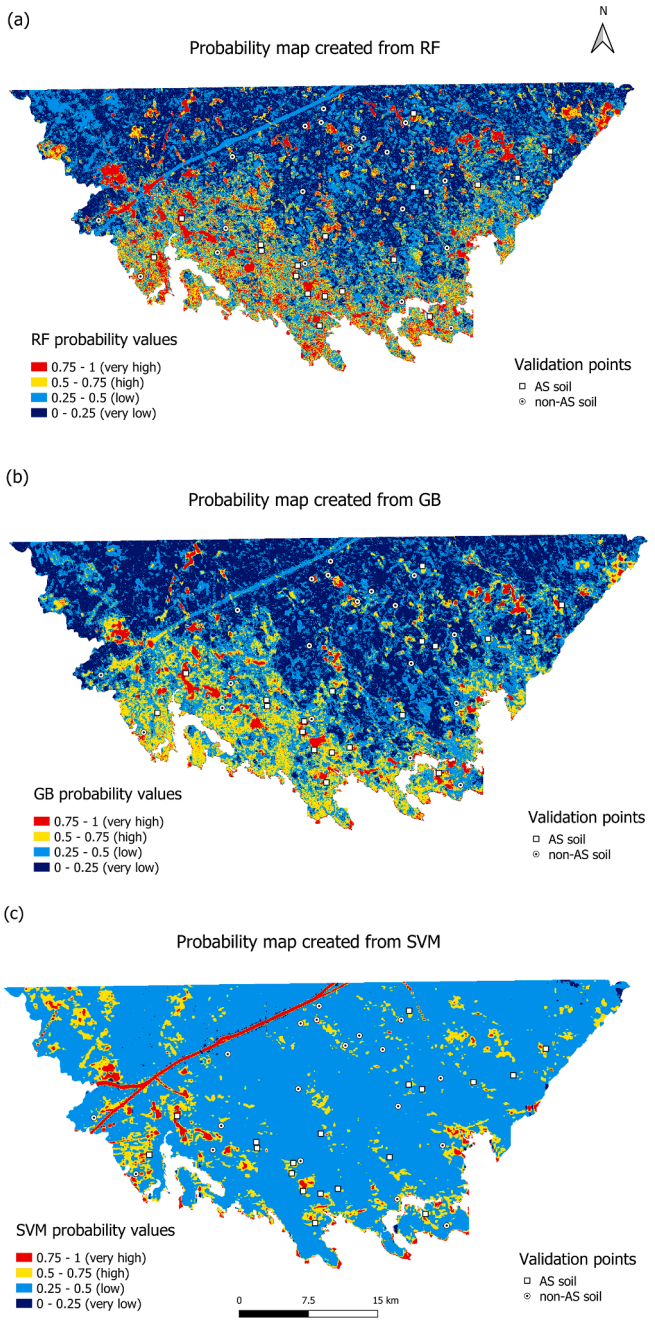

Fig. 6. Probability maps created from the different machine learning models. (a) Random Forest (RF). (b) Gradient Boosting (GB). (c) Support Vector Machine (SVM).

points, AS and non-AS soils, are represented in the probability maps (Fig. 6). The validation of the maps consists of checking if the prediction made by the model for the cells coincides with the validation points. For RF and GB maps, the percentage of validation points correctly classified is equal to the obtained in the evaluation of the methods (Table 2). This means equal recall values. However, in the case of SVM the percentage of validation points corresponding to AS soils correctly classified in the map is 37%, slightly smaller than the one obtained in the evaluation of the model (42%) (Tables 2 and 3). This is because the predictions of the model for one point and for the cell where this point is located on the probability map are different. Note that the prediction values for the probability maps are calculated for the 50 m x 50 m cells. Then, there is a small possibility that the prediction for a single point is different from the prediction made for the cell where this point is located. This

Table 3
Validation of the probability maps created from the machine learning methods. Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM). Validation points are acid sulfate (AS) soils and non-acid sulfate (non-AS) soils.

Model	Probability zone	% of study area	Validation points	
			AS	non-AS
RF	Very high [0.75 −1]	10	7	2
	High [0.5–0.75)	24	7	4
	Low [0.25–0.5)	28	2	5
	Very low [0–0.25)	38	3	8
GB	Very high [0.75 −1]	4	6	2
	High [0.5–0.75)	17	9	3
	Low [0.25–0.5)	36	3	8
	Very low [0–0.25)	43	1	6
SVM	Very high [0.75 −1]	3	0	0
	High [0.5–0.75)	12	7	3
	Low [0.25–0.5)	84	11	16
	Very low [0–0.25)	1	1	0

situation has only been observed for this model and with one of the points.

On the other hand, it is interesting to see which probability area has been predicted for the cells where the validation points are located. This can give additional information about the suitability of the models. Table 3 shows these results for the three methods. Although GB correctly predicts a greater number of validation points, RF is able to classify properly more points in the very high and very low probability areas. This can be related with the weight or importance of the environmental covariates in each model. For the case of SVM, the most of the validation points are located in the low and high probabilities areas. Curiously, there are no AS soil validation points classified in the very high probability area, just as there are no non-AS soil validation points in the very low probability area. Moreover, the number of AS soil samples incorrectly classified is larger than the correct ones (Table 3). It should also be noted that this model only predicts 1% of the studied area for the case of very low probability for AS soils, and however, one of the AS soil validation points is located in this area. This confirms that this method has problems to classify AS soils.

3.3. Comparison between AS soil probability maps

Once the AS soil probability maps for the three machine learning methods have been created, it is important to see the reliability of the maps. For this purpose it is convenient to analyze the predictions of the different probability areas for each map, and see if the modeled probability maps improve the accuracy with respect to the conventional probability map produced by the Geological Survey of Finland (Fig. 5). In order to evaluate the differences and similarities between all AS soil probability maps, a comparison between them has been done. It should be noted that the modeled probability maps have a larger size (∼ 1,091 km^2) than the conventional map (∼ 905 km^2). This is due to the conventional map is restricted to the Littorina Sea maximum extent. In the comparison only the common area between the modeled and conventional probability maps is considered. Table 4 shows the distribution of the predicted probability areas for the common area of each probability map. As the four probability areas or classes are different for the conventional and the modeled maps, each map is represented with its corresponding nomenclature. The different criteria in the modeled and conventional maps for the probability areas is due to the probability being calculated differently in both cases (SubSections 2.4 and 2.7). In the modeled probability map, the very low, low, high and very high probability classes correspond to [0–0.25), [0.25–0.5), [0.5–0.75) and [0.75–1], respectively. In the conventional map the proportion of AS soils for the high probability is larger than 75%, whereas for moderate, low and very low it is around 50%, 10% and 0%, respectively. From the results shown in Table 4, one thing that attracts attention is the large extent of the area predicted as very low probability in the conventional map. This is a consequence of the method used for the calculation of the probability. In the conventional case, it is based on the ratio of soil cores classified as AS soils with respect to the total number of soil cores in a given area. Thus, the probability depends on the number of soil cores and also on the extent of the area taken into account. In the uppermost meter of the study area, around 58% corresponds to bedrock, outcrops and block fields, which was not sampled. A large part of the area classified as very low probability is located on the bedrock outcrops. Thus, the absence of soil cores in this area can result in an overestimation of the very low probability area.

On the other hand, the area with a probability of encountering AS soils larger than 75% is quite similar for all the methods except for the case of RF, which is three or four times larger. As it was already shown in the previous subsection, RF is able to correctly classify more AS soils in the very high probability area than the other models. In general, if there is not a validation point, it is difficult to determinate which method classifies a given area correctly when the predictions are different. Moreover, if two methods predict a similar percentage for a probability area, it does not mean that these predictions match for the same areas of the maps. Thus, in order to obtain more information from the probability maps, we have to compare the predictions of each cell or pixel of the maps. However, it can only be done for the modeled probability maps, where the probability classes are the same. For the conventional map, the areas with very low and low probability correspond to the very low probability in the modeled maps, whereas the moderate area is between the low and the high probability areas in the modeled maps. Only the high probability area in the conventional map and the very high in the modeled maps can be compared as both represent a probability larger than 75%. Thus, the comparison between the conventional and the modeled maps is restricted to this last case. While the comparison between the modeled maps has been done for the four probability areas. Table 5 shows the percentage of equal predicted areas by the machine learning methods and their corresponding percentage for the probability areas. The predictions for the three machine learning methods (3 ML in Table 5) match for the 22% of the total common area. Most of this area corresponds to the low probability. Similar results in the distribution of the probability areas are found when SVM is compared to RF or GB. This is due to SVM overestimates this probability class.

A better idea about the correctly predicted areas can be obtained

Table 4
Distribution of the probability areas of the probability maps for the common area between the different maps. Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM).

Methods	Probability areas (%)			
	very low	low	high	very high
RF	33	36	19	12
GB	36	39	21	4
SVM	0	86	11	3
	very low	low	moderate	high
Conventional	65	23	9	3

Table 5
Comparison of the predicted areas of the modeled probability maps created from the machine learning models. 3 ML (three machine learning methods), Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM).

Methods	% of equal predicted areas	Probability areas (%)			
		very low	low	high	very high
3 ML	22	0	86	9	5
RF & GB	59	43	35	16	6
RF & SVM	37	0	86	9	5
GB & SVM	40	0	85	12	3

from the comparison between the maps created from RF and GB, as both methods are very good in the classification of AS soils (Table 2). In almost 60% of the study area, the prediction of both methods matches (Table 5). Moreover, the distribution of the probability areas are quite similar to the distribution of the probability areas for RF and GB (Table 4).

One of the main goals in the AS soil probability maps is the correct localization of the areas with the highest probability for AS soils. Table 6 shows the percentage of equal predicted areas by different methods when the probability of encountering AS soils is larger than 75%. The comparison has been done for all probability maps, including the conventional map. All methods predict a quite similar percentage (~3) for this probability area except RF (Table 4). However, when each modeled map is compared to the conventional map, the area that matches decreases considerably (Table 6). For RF or GB, only around the 1% of the study area matches with the conventional map for this probability area. The worst result is obtained for SVM with only 0.5%. Although this result is not surprising since the SVM model performs poorly in the classification of AS soils. Furthermore, all modeled maps have been compared between them and to the conventional map (Table 6). As it can be seen, in all the cases analyzed the predicted area that matches with the conventional map is at least reduced to one third. A clear example is the comparison between the RF and GB maps, where the predictions for this probability class match for 3.5% of the total area, but only 1% with the conventional map.

These differences between the predictions in the conventional and modeled maps are related to the way of calculating the probability. In the conventional case, it depends on the number of soil samples and on the extent of area considered. This can lead to relative results, which may not be realistic in some cases. Contrary, in the case of machine learning methods, the predictions of the probability are made considering only the values of the environmental layers for each cell (50 m x 50 m). The probability is completely independent of the number of soil cores in the cell. Soil cores are only used to train and validate the models. On the other hand, the small size of the cells allows the creation of maps with larger accuracy than the conventional map, where the predictions are made for areas with a minimum size of 0.06 km^2.

Furthermore, one should take into account that the conventional map is highly subjective and strongly dependent on the person creating the map, while the modeled maps are objective and more easily reproducible. Thus, future studies should focus on improving the machine learning models in order to get more accurate maps. It is expected that the use of more input data, both soil samples and environmental covariates, will contribute to a better training of the models, leading to more accurate predictions. In the case of the environmental covariates, a variable selection of the most relevant layers for the classification of AS soils will be fundamental for this purpose.

4. Conclusions

In the present study, we have analyzed in detail the predictive abilities of three supervised machine learning techniques for mapping acid sulfate (AS) soils. The methods evaluated are Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM). Our results show that both RF and GB have high predictive abilities for mapping AS soils. GB yields accuracies 5 to 6% larger than RF. SVM is not able to correctly distinguish AS soils, which makes it an unsuitable method for this case study.

The AS soil probability maps created from the machine learning methods also show the predictive abilities of the models. The probability map created from SVM clearly displays the overestimation of the areas with low probability for AS soils. While the probability maps created from GB and RF are more accurate. The predictions of these two models match for 60% of the study area, where 3.5% corresponds to areas with a probability of encountering AS soils larger than 75%. Only 1% of this probability area matches with the conventional map. In general, the mapping process using machine learning methods is faster, more objective and accurate, and less expensive. Future studies should evaluate the use of machine learning methods for AS soil mapping on larger extents. A crucial development would also be the assessment of uncertainty, for example through the use of quantile regression forest.

Table 6
Comparison of the predicted area with a probability of encountering acid sulfate (AS) soils larger than 75% for all probability maps: the modeled and the conventional (CM). 3ML (three machine learning methods), Random Forest (RF), Gradient Boosting (GB) and Support Vector Machine (SVM).

Methods	% of area with a probability ⩾75%
RF & CM	1.4
GB & CM	1.3
SVM & CM	0.5
3ML	1.2
3ML & CM	0.4
RF & SVM	1.8
RF & SVM & CM	0.5
GB & SVM	1.3
GB & SVM & CM	0.4
RF & GB	3.5
RF & GB & CM	1

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Anton Akusok for useful discussions. This work has been financial supported by Stiftelsen Arcada foundation (Finland).

References

Airo, M.-L., Loukola-Ruskeeniemi, K., 2004. Characterization of sulfide deposits by airborne magnetic and gamma-ray responses in eastern Finland. Ore Geol. Rev. 24, 67–84.
Airo, M.-L., Hyvönen, E., Lerssi, J., Leväniemi, H., Ruotsalainen, A., 2014. Tips and tools for the application of GTK's airborne geophysical data. Geological Survey of Finland. Report of Investigation 215.
Alhonen, P., Mantere-Alhonen, S., Vuorinen, A., 1997. Preliminary observations on the metal content in some milk samples from an acid geoenvironment. Bull. Geol. Soc. Finland 69, 31–41.
Andriesse, W., van Mensvoort, M.E.F., 2006. Acid sulfate soils: Distribution and extent. In: Lal, R. (Ed.), Encyclopedia of soil science (electronic version) (2nd edition). Taylor & Francis Group, LLC., New York, pp. 14–19.
Åström, M., Björklund, A., 1997. Geochemistry and acidity of sulphide-bearing postglacial sediments of western Finland. Environ. Geochem. Health 19, 155–164.
Behrens, T., Schmidt, K., Zhu, A.X., Scholten, T., 2010. The ConMap approach for terrain-based digital soil mapping. Eur. J. Soil Sci. 61, 133–143.
Beucher, A., Österholm, P., Martinkauppi, A., Edén, P., Fröjö, S., 2013. Artificial neural network for acid sulfate soil mapping: Application to the Sirppujoki River cathment area, south-western Finland. J. Geochem Explor 125, 46–55.
Beucher, A., Fröjö, S., Österholm, P., Martinkauppi, A., Edén, P., 2014. Fuzzy logic for acid sulfate soil mapping: Application to the southern part of the finnish coastal areas. Geoderma 226–227, 21–30.
Beucher, A., Siemssen, R., Fröjö, S., Österholm, P., Martinkauppi, A., Edén, P., 2015. Artificial neural network for mapping and characterization of acid sulfate soils: Application to the Sirppujoki River catchment, southwestern Finland. Geoderma 247–248, 38–50.
Boman, A., Sohlenius, G., Mattbäck, S., Becher, M., Liwata-Kenttälä, P., Öhrling, C., Auri, J., Edén, P., 2019. Classification of Finnish and Swedish acid sulfate soil materials. Geophys. Res. Abstracts 21. EGU2019-6597-1.
Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83.
Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62, 394–407.
Cortes, C., Vapnik, V., 1995. Support vector networks. Mach. Learn. 20, 273–297.
Creeper, N., Fitzpatrick, R., Shand, P., 2012. A simplified incubation method using chip-trays as incubation vessels to identify sulphidic materials in acid sulphate soils. Soil Use Manag. 28, 401–408.

Edén, P., Rankonen, E., Auri, J., Yli-Halla, M., Österholm, P., Beucher, A., Rosendahl, R., 2012a. Definition and classification of Finnish Acid Sulfate Soils. 7th IASSC abstract, Vaasa, Finland.

Edén, P., Auri, J., Rankonen, E., Martinkauppi, A., Österholm, P., Beucher, A., Yli-Halla, M., 2012b. Mapping acid sulfate soils in Finland – methods and results. 7th IASSC abstract, Vaasa, Finland.

Erviö, R., 1975. Cultivated sulphate soils in the drainage basin of river Kyrönjoki. J. Sci. Agric. Society Finland 47, 550–561.

Fältmarsch, R.M., Åström, M.E., Vuori, K.-M., 2008. Environmental risk of metals mobilized from acid sulphate soils in Finland: a literature review. Boreal Environ. Res. 13, 444–456.

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

Gambill, D.R., Wall, W.A., Fulton, A.J., Howard, H.R., 2016. Predicting USCS soil classification from soil property variables using Random Forest. J. Terrramech. 65, 85–92.

Geological Survey of Finland, 2021a. Maankamara – map services. http://gtkdata.gtk.fi/Maankamara/index.html.

Geological Survey of Finland, 2021b. Acid Sulfate Soils – map services. http://gtkdata.gtk.fi/hasu/index.html.

Gill, M.K., Mariush, T.A., Kemblowski, W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42 (4), 1033–1046.

Grimm, R., Behrens, T., Maerker, M., Elsenbeer, A., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island – Digital soil mapping using Random Forests analysis. Geoderma 146, 102–103.

Haavisto-Hyvärinen, M., Kutvonen, H., 2007. Maaperäkartan käyttöopas. Geological Survey of Finland. 66, pp.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., et al., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE. 10 (e0125814).

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.V., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748.

Heung, B., Bulmer, C.E., Schimdt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214–215, 141–154.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schimdt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.

Horn, B.K.P., 1981. Hill shading and the reflectance map. Proc. IEEE 69 (1), 14–47.

Hudd, R., 2000. Springtime episodic acidification as a regulatory factor of estuary spawing fish recruitment. PhD Thesis, Helsinki University, Finland. 42 p.

IUSS Working Group WRB. 2015. World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013) An Introduction to Statistical Learning, V 103 ISBN: 978-1-4614-7137-0.

Jokinen, S.A., Virtasalo, J.J., Jilbert, T., Kaiser, J., Dellwig, O., Arz, H.W., Hänninen, J., Arppe, L., Collander, M., Saarinen, T., 2018. A 1500-year multiproxy record of coastal hypoxia from the northern Baltic Sea indicates unprecedented deoxygenation over the 20th century. Biogeosciences 15, 3975–4001. https://doi.org/10.5194/bg-15-3975-2018.

Jokinen, S.A., Jilbert, T., Tiihonen-Filppula, R., Koho, K., 2020. Terrestrial organic matter input drives sedimentary trace metal sequestration in a human-impacted boreal estuary. Sci. Total Environ. 717, 137047.

Kivinen, E., 1950. Sulphate soils and their management in Finland. In: Transactions of the Fourth International Congress of Soil Science. Amsterdam, pp. 259–262.

Korpela, K., Niemelä, O., 1985. Maaperäkartat 1:20 000 ja 1:50 000. Maankäyttö 2. Geological Survey of Finland.

Kovačevic, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154, 340–347.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer.

Lamorski, K., Pachepsky, Y., Stawiński, C., Walczak, R.T., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72, 1243–1247.

Lehtinen, M., Nurmi, P., Rämö, T., 1998. Suomen kallioperä 3000 vuosimiljoonaa. Suomen Geologinen Seura ry, Helsinki, p. 375.

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.

Lieβ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture Comparison of regression tree and Random Forest models. Geoderma 170, 70–79.

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

Mattbäck, S., Boman, A., Österholm, P., 2017. Hydrogeochemical impact of coarse-grained post-glacial acid sulfate soil materials. Geoderma 308, 291–301.

McBratney, A., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Michael, P.S., 2013. Ecological Impacts and Management of Acid Sulphate Soil: A Review. Asian Journal of Water, Environment and Pollution, Vol. 10, No. 4 (2013), pp. 13–24.

Michael, P.S., Fitzpatrick, R.W., Reid, R.J., 2017. Effects of live wetland plant macrophytes on acidification, redox potential and sulphate content in acid sulphate soils. Soil Use Manage. 33, 471–481.

Müller, A.C., Guido, S., 2016. An Introduction to Machine Learning with Python, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Österholm, P., Åström, M., 2002. Spatial trends and losses of major and trace elements in agricultural acid sulphate soils distributed in the artificially drained Rintala area. W. Finland. Appl. Geochem. 17 (9), 1209–1218.

Österholm, P., Åström, M., 2004. Quantification of current and future leaching of sulfur and metals from Boreal acid sulfate soils, western Finland. Aust. J. Soil Res. 42, 547–551.

Palko, J., 1986. Mineral element content of timothy (Phleum pretense L.) in an acid sulphate soils and their agricultural and environmental problems in Finland. Acta Agric. Scand. 36, 399–409.

Palko, J., 1994. Acid sulphate soils and their agricultural and environmental problems in Finland. Acta University Oulu, C75. University Oulu (PhD thesis).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pons, L.J., 1973. Outline of the genesis,characteristics, classification and improvement of acid sulfate soils. In: Dost, H. (Ed.), Acid sulphate soils, Introductory papers and bibliography, ILRI Publication 18 Proceedings of the International Symposium 13–20 August 1972, Wageningen, Vol. 1, pp. 3–27.

Porwal, A., Carranza, E.J.M., Hale, M., 2003. Artificial neural networks for mineral potential mapping: a case study from Aravalli Province, Western India. Natural Resour. Res. 12 (3), 155–171.

Powers, D.M.W., 2011. Evaluation: from precision, recall, and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. V 2, 37–63.

Purokoski, P., 1959. Rannikkoseudun rikkipitoisista maista. Referat: Über die schwefelhaltigen Böden an der Küste Finnlands. Agrogeologische Publicationen 74.

Puustinen, M., Merilä, E., Palko, J., Seuna, P., 1994. Kuivatustila, viljelyköytäntö ja vesistökuormitukseen vaikuttavat ominaisuudet Suomen pelloilla. Summary: Drainage level, cultivation practices and factors affecting load on waterways in Finnish farmland. National Board of Waters and Environment, research report A198. 323 pp.

QGIS Development Team, 2019. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

Roos, M., Åström, M., 2005. Hydrochemistry of rivers in an acid sulphate soil hotspot area in western Finland. Agric. Food Sci. 14, 24–33.

Roos, M., Åström, M., 2006. Gulf of Bothnia receives high concentrations of potentially toxic metals from acid sulphate soils. Boreal Environ. Res. 11, 383–388.

Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T., 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232–234, 243–256.

Sindayiheburaa, A., Ottoyb, S., Dondeyneb, S., Van Meirvennec, M., Van Orshovenb, J., 2017. Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus. Catena 156, 161–175.

Sullivan, L.A., Fitzpatrick, R.W., Bush, R.T., Burton, W.D., Shand, P., Ward, N.J., 2010. The classification of acid sulfate soil materials: further modifications. Southern Cross GeoScience Technical Report No. 310. Southern Cross University, Lismore, NSW, Australia 12 pp.

Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9 (1), 319.

Suppala, I., Lintinen, P., Vanhala, H., 2005. Geophysical characterising of sulphide rich finegrained sediments in Seinäjoki area, western Finland. Geol. Surv. Finland Spec. Pap. 38, 61–71.

Teng, H.T., Viscarra Rossel, R.A., Shi, Z., Behrens, T., 2018. Updating a national soil classification with spectroscopic predictions and digital soil mapping. Catena 164, 125–134.

Toivonen, J., Österholm, P., Fröjdö, S., 2013. Hydrological processes behind annual and decadal-scale variations in the water quality of runoff in Finnish catchments with acid sulfate soils. J. Hydrol. 487, 60–69.

Tziachrisa, P., Aschonitisa, V., Chatzistathisa, T., Papadopoulou, M., 2019. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. Catena 174, 206–216.

Urho, L., 2002. The importance of larvae and nursery areas for fish production. PhD Thesis, Helsinki University, Finland. 135 p.

Vanhala, H., Suppala, I., Lintinen, P., 2004. Integrated geophysical study of acid sulphate soil area near Seinäjoki, Southern Finland. Sharing the Earch: EAGE 66th Conference & Exhibition, Paris, France, 7–10 June 2004: Extended Abstracts. EAGE, Houten (4 pp. Optical disc (CD-ROM)).

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Wei, Q., Dunbrack Jr., R.L., 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLOS ONE 8 (7), e67863.

Weiss, G., Provost, F., 2001. The Effect of Class Distribution on Classifier Learning: An Empirical Study. Tech Rep.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24.

Yli-Halla, M., 1997. Classification of acid sulphate soils of Finland according to Soil Taxonomy and the FAO/Unesco legend. agric. Food Sci. 6, 247–258.

Yli-Halla, M., Puustinen, M., Koskiaho, J., 1999. Area of cultivated acid sulphate soils in Finland. Soil Use Manag. 15, 62–67.