environmental sciences proceedings

MDPI

Proceeding Paper

A First Approximation for Acid Sulfate Soil Mapping in Areas with Few Soil Samples †

Virginia Estévez [1,*], Stefan Mattbäck [2,3] and Anton Boman [3]

[1] Graduate School and Research, Arcada University of Applied Sciences, 00560 Helsinki, Finland
[2] Department of Geology and Mineralogy, Åbo Akademi University, 20500 Åbo, Finland; stefan.mattback@gtk.fi
[3] Geological Survey of Finland, 67101 Kokkola, Finland; anton.boman@gtk.fi
* Correspondence: estevezv@arcada.fi
† Presented at the 5th International Electronic Conference on Remote Sensing, 7–21 November 2023; Available online: https://ecrs2023.sciforum.net/.

Abstract: Acid sulfate soil mapping is the first step to avoid possible environmental damages created by one of the most problematic soils existing in nature. One of the problems in acid-sulfate soil mapping is the lack of soil samples in some regions. This prevents the creation of occurrence maps. For the first recognition of these regions, a possible solution could be the use of soil samples from other areas with similar characteristics. In this study, we analyze if a machine learning method is able to correctly classify the soil samples in an area where it has not been trained. For this, Random Forest and two different regions located in southern Finland with a similar composition of soils are considered.

Keywords: acid sulfate soils; digital soil mapping; random forest; remote sensing

Citation: Estévez, V.; Mattbäck, S.; Boman, A. A First Approximation for Acid Sulfate Soil Mapping in Areas with Few Soil Samples. *Environ. Sci. Proc.* **2024**, *29*, 4. https://doi.org/10.3390/ECRS2023-15831

Academic Editor: Riccardo Buccolieri

Published: 15 January 2024

1. Introduction

Nowadays, artificial intelligence shows great potential to solve different problems in various research fields related to the economy and society. One of these research fields is soil science, which has a direct impact on agriculture, energy, or climate change. Among the different types of soils, acid sulfate (AS) soils are considered the most problematic from an environmental point of view [1]. This is because this type of soil, when oxidized, can acidify the soil and mobilize toxic metals that may lead to serious environmental damage. To minimize possible environmental problems, it is essential to avoid the oxidation and drainage of these soils. A fundamental part of this is to locate the areas where these soils appear. The application of machine learning techniques in digital soil mapping has meant a great advance in this field [2]. One of the advantages is that these techniques allow the creation of more precise maps with a lower number of samples than the traditional methods [3]. In the case of AS soils, not all machine learning techniques are successful in their predictions [4]. This is mainly due to the complex relationship between the AS soils and the environmental covariates [4]. For example, Support Vector Machine (SVM) is a method unable to correctly classify this type of soil [4], while Random Forest (RF) and Gradient Boosting (GB) are very good methods for the accurate classification and prediction of AS soils [4,5]. Furthermore, the AS soil probability maps created with these two models have high precision [4,5]. Another method able to classify AS soils under some conditions is the Extreme Learning Machine (ELM) [6]. Artificial Neural Network [7], Fuzzy Logic [8], or Fuzzy K-means [9] have also been used for AS soil mapping.

The use of a supervised machine learning technique in AS soil mapping requires two different types of data: the soil samples and the environmental covariates generally created from remote sensing data. The soil samples and their corresponding environmental covariate values are used to train and validate the model. For the prediction, the model

only considers the environmental covariates. One of the problems in AS soil mapping is the lack of soil samples in some regions. This hinders the creation of AS soil occurrence maps by means of supervised machine learning techniques. In such regions, the presence of AS soils may be overlooked using a traditional mapping approach. Consequently, there is a risk that AS soils are unknowingly drained, leading to environmental damage as a result of agricultural, forestry, or urban activities. For the first recognition of such regions, in addition to using the remote sensing data of the area, a possible solution could be the use of soil samples from other areas with similar characteristics for training the model. The question is whether a machine learning model could correctly classify AS soils in an area where it has not been trained. So far, there are no previous works where this study has been carried out. However, if this were possible, this first prediction could be used to design an efficient sampling plan for the region. In previous works, Random Forest has shown high abilities for the correct prediction of AS soils [4,5]. In this work, we analyze if RF is able to correctly classify the soil samples in an area where it has not been trained. For this, two different regions located on the coast of the Gulf of Finland, with a similar composition of their soils, are considered.

2. Materials and Methods

2.1. Study Areas

For this study, two different study areas in the south of Finland were considered: Virolahti and its surroundings, and a coastal area situated between the cities of Helsinki and Loviisa (Hel-Lov); see Figure 1. Both areas are located in the Littorina Sea, where AS soils are usually found. The land cover in these two regions mainly consists of bedrock, outcrops, and blockfields. Furthermore, the soils in both areas are similar. This has allowed us to carry out the study of this work.

Figure 1. Location of the study areas (pink color) and the maximal extent of the Littorina Sea (diagonal lines). Both study areas are located in the Littorina Sea.

2.2. Datasets

The datasets of the two study areas include two different types of data: the soil samples and the environmental covariates. The soil samples correspond to AS soils and non-AS soils. The classification of the samples was made following the method described in [10] by the Geological Survey of Finland (GTK). In the area of Virolahti and its surroundings, there are 186 soil samples: 93 for AS soils and 93 for non-AS soils. In the Hel-Lov area, there are a total of 458 soil samples, 229 for each class.

The environmental covariates (raster data) considered in this study are: quaternary geology, digital elevation model (DEM) derived from LiDAR data, terrain layers derived from DEM, and aerogeophysics layers. LiDAR and aerogeophysics are remote sensing data from airborne surveys. All these covariates can provide useful information for the characterization of AS soils. Each dataset consists of 17 environmental covariates with a resolution of 50 m and the Finnish coordinate reference system (ETRS89/TM35FIN(E,N)). The environmental covariates of Virolahti and its surroundings were created in a previous work [11], whereas the environmental covariates of the Hel-Lov area have been created for this study with Qgis.

2.3. Method

In this study, we have analyzed the ability of a supervised machine learning method for the classification of AS soils in an area different from the one in which the model was trained. The method considered is Random Forest (RF), which is a method based on decision trees [12]. This technique has been used for the classification and prediction of AS soils in previous studies [4,5,11]. We have used grid search and cross-validation (GridSearchCV) for the selection of the best tuning parameters for the performance of the model; for more information, see [4]. The metrics used for the determination of the suitability of the model are those related to the confusion matrix: precision, recall, and F1-score [13].

3. Results and Discussion

A supervised machine learning method can make predictions if it has been previously trained and validated with the dataset. In the present study, the only way to know if a model is capable of correctly predicting the AS soils of an area where it has not been trained is by validating the model with the soil samples of that area. For this reason, an area with a large dataset available has been selected for the study. This area is the region between Helsinki and Loviisa (Hel-Lov). The Virolahti area dataset was already used in a previous study, where it was analyzed how variable selection improves the prediction accuracy of the AS soil mapping [5]. In principle, it cannot be guaranteed that the importance of environmental covariates is the same in the two regions. For this reason, a dataset with 17 layers has been considered for this study. As already mentioned, RF is a method with high predictive capacity for the prediction of AS soils [4,5,11]. Furthermore, it is one of the machine learning techniques that is least affected by irrelevant covariates or redundant information [5,14]. This makes it an ideal method for the present study.

The process of this study has two steps. First, the RF model is trained and validated with the dataset of Virolahti and its surroundings. 80% of the soil samples are used for training the model and the remaining 20% for its validation (Figure 2). The next step is the validation of the model with the dataset from the Hel-Lov area. All the soil samples of this region are used for validation.

The results are shown in Table 1, where the metrics considered for the evaluation of the model are represented. As it can be seen, the model is able to distinguish the AS soils and non-AS soils of the Hel-Lov area when it is trained using the data from Virolahti and its surroundings. The similar F1-score values for both classes indicate that the performance of the model is good for both classes. This is also confirmed by the balance between precision and recall for both classes. These results, where all the metrics are above 60%, are very good for a model that has not been trained in the area of the prediction. On the other hand, we

have also analyzed the performance of the model when it is trained with soil samples from the own area. For this, the proportion of points for training the model and its validation was 80 and 20%, respectively. Training the model in the same area improves the results by up to 13%; see Table 1. Therefore, RF is a model that is able to make a prediction in an area where it has not been trained. However, the prediction accuracy of the model will not be as good as the prediction made when the model is trained with soil samples from the area itseft. Thus, training the RF model in a different area can be used for the first recognition of areas with limited soil samples as well as for the creation of an efficient sampling plan design in those areas.

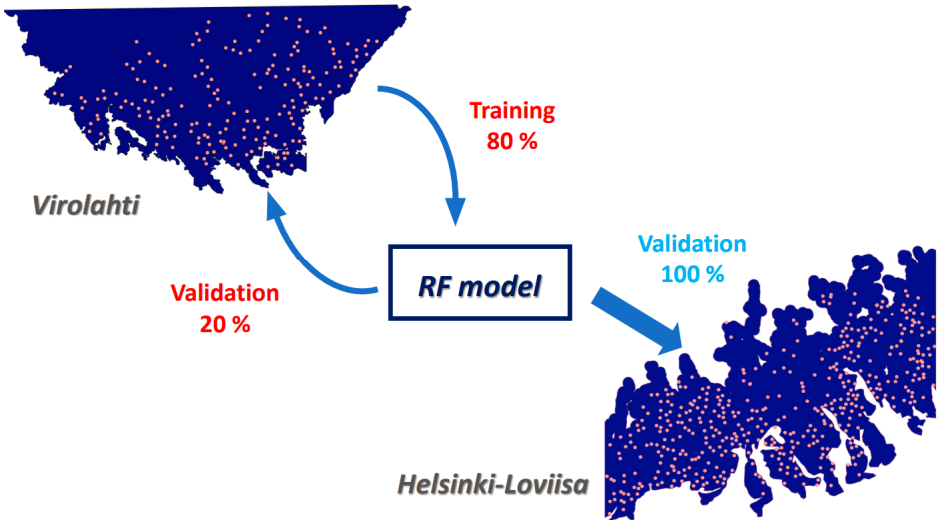

Figure 2. Schematic diagram of the validation process.

Table 1. Metrics related to the confusion matrix for the Random Forest (RF) prediction in the area between Helsinki and Loviisa (Hel-Lov) when the model is trained using soil samples from Virolahti and its surroundings and when the model is trained with the soil samples from the own area (Hel-Lov). The two classes are acid sulfate (AS) and non-acid sulfate (non-AS) soils.

Trained with the Soil Samples from:	Class	Precision	Recall	F1-Score
Virolahti	non-AS	0.64	0.62	0.63
	AS	0.63	0.65	0.64
Hel-Lov	non-AS	0.77	0.72	0.74
	AS	0.73	0.78	0.76

It must be taken into account that training the model with soil samples from other areas should only be carried out if there are at least some points in the area itself that allow the validation of the model and if the geology of both areas can be considered relatively similar. Future work should analyze the ability of the model to predict AS soils in areas where the soil composition is different from the one in the training area.

4. Conclusions

In this study, we have analyzed the ability of a machine learning method such as Random Forest to make a correct prediction of the acid sulfate soils in an area where it has not been previously trained. The training of the model has been carried out with soil samples from another region with similar characteristics in the composition of the soils. Our results show that the method performs well in classifying both classes, with all metrics evaluated above 60%. This is an advance in the field since it allows the first recognition of the regions with a limited number of soil samples. Furthermore, this first prediction will permit the development of a more efficient sampling plan design in these regions that

is also oriented towards the application of machine learning techniques for acid sulfate soil mapping.

Author Contributions: Conceptualization, V.E.; methodology, software and validation, V.E.; formal analysis, V.E.; resources and data curation, V.E. and S.M.; writing—original draft preparation, V.E.; writing—review and editing, V.E., A.B. and S.M.; visualization, V.E.; supervision, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by HaSuRiski project and Stiftelsen foundation (Finland).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Raw data will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Michael, P.S. Ecological Impacts and Management of Acid Sulphate Soil: A Review. *Asian J. Water Environ. Pollut.* **2013**, *10*, 13–24.
2. McBradney, A.B.; Mendonça Santos, M.L.; Minasmy, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
3. Brus, D.J.; Kempen, B.; Heuvelink, G.B.M. Sampling for validation of digital soil maps. *Eur. J. Soil Sci.* **2011**, *62*, 394–404. [CrossRef]
4. Estévez, V.; Beucher, A.; Mattbäck, S.; Boman, A.; Auri, J.; Björk, K.-M.; Osterhölm, P. Machine learning techniques for acid sulfate soil mapping in southeastern Finland. *Geoderma* **2022**, *406*, 115446. [CrossRef]
5. Estévez, V.; Mattbäck, S.; Boman, A.; Beucher, A.; Björk, K.-M.; Osterhölm, P. Improving prediction accuracy for acid sulfate soil mapping by means of variable selection. *Front. Environ. Sci.* **2023**, *11*, 1213069. [CrossRef]
6. Estévez, V.; Mattbäck, S.; Björk, K.-M. Importance of the activation function in extreme learning machine for acid sulfate soil classification. In Proceedings of the ELM 2022, Helsinki, Finland, 8–9 December 2022.
7. Beucher, A.; Österholm, P.; Martinkauppi, A.; Edén, P.; Fröjdö, S. Artificial neural network for acid sulfate soil mapping: Application to the Sirppujoki River cathment area, south-western Finland. *J. Geochem. Explor.* **2013**, *125*, 46–55. [CrossRef]
8. Beucher, A.; Fröjdö, S.; Österholm, P.; Martinkauppi, A.; Edén, P. Fuzzy logic for acid sulfate soil mapping: Application to the southern part of the finnish coastal areas. *Geoderma* **2014**, *226–227*, 21–30. [CrossRef]
9. Huang, J.; Nhan, T.; Wong, V.N.L.; Nohaton, S.G.; Lark, R.M.; Triantafilis, J. Digital Soil Mapping of a Coastal Acid Sulfate Soil Landscape. *Soil Res.* **2014**, *52*, 327–339. [CrossRef]
10. Boman, A.; Mattbäck, S.; Becher, M.; Sohlenius, G.; Auri, J.; Öhrling, C.; Liwata-Kenttälä, P.; Edén, P. Classification of Acid Sulfate Soils and Materials in Finland and Sweden: Re-Introduction of Pseudoacid Sulfate Soil Materials. In *Abstract Book, Proceedings of the 9th International Acid Sulfate Soils Conference, Adelaide, Australia, 26–31 March 2023*; University of Adelaide: Adelaide, Australia, 2023. Available online: https://set.adelaide.edu.au/acid-sulfate-soils-centre/ua/media/50/9th-iassc-abstract-book.pdf (accessed on 31 January 2023).
11. Estévez, V. Machine Learning Methods for Classification of Acid Sulfate Soils in Virolahti. Master's Thesis, Arcada University of Applied Sciences, Helsinki, Finland, June 2020.
12. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
13. Powers, D.M.W. Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
14. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013.