Geoderma 320 (2018) 30–42

Contents lists available at ScienceDirect

Geoderma

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

Predicting artificially drained areas by means of a selective model ensemble

Anders Bjørn Møller*, Amélie Beucher, Bo V. Iversen, Mogens H. Greve

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

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Denmark
Agriculture
Soil management
Water management
Geostatistics
Models

ABSTRACT

Farmers often install subsurface drainage systems to improve yields on wet soils, which has large impacts on the hydrological system. The present study uses an ensemble of machine learning models to map the extent of artificially drained areas in Denmark. The prediction is based on 745 field observations, of which one third is held out for evaluation, and 46 covariate layers. A library of 308 models is trained using 77 machine learning methods and four datasets containing either a combination of topographic variables, satellite imagery, soil properties and land use information or principal components based on these variables.

A stepwise algorithm then selects models from the library, based on their predictions on a hillclimb dataset. The results show that models trained using principal components generally yielded a better performance than the models trained with the raw covariates. Furthermore, the best results were obtained when only a random fraction of the models was available for selection at each step. The covariates that were most important for the prediction of artificially drained areas mostly related to soil properties and topography. Overall, the ensemble predicted artificially drained areas with an accuracy of 76.5%. The study demonstrates machine learning as an accurate method for mapping artificially drained areas, which is likely to benefit both farmers and decision makers.

1. Introduction

Soil drainage is a major agricultural concern, as insufficient drainage can greatly reduce crop yields (Collaku and Harrison, 2002; Ren et al., 2014; Watson et al., 1976). Farmers often respond to poor drainage conditions by installing subsurface drainage pipes in the soil. Information on the location of the pipes is often important to both farmers and environmentalists as it is necessary when new pipes are installed (Allred et al., 2004; Allred and Redman, 2010), and because the systems influence the hydrological cycle (Boland-Brien et al., 2014) and the leaching of nutrients (Ernstsen et al., 2015). The individual contractors that conduct the drainage work rarely keep shared records of the drainage systems, leading to a loss of information. Consequently, there is a strong interest in methods, which can map artificially drained areas.

Studies have shown that ground-penetrating radar can reliably locate drainage pipes at the field level, unless the soil is water saturated or has a high clay content (Allred et al., 2004; Allred and Redman, 2010).

In larger areas, studies have mapped drainage systems based on aerial photography (Northcott et al., 2000; Verma et al., 1996), an approach which later studies automated by means of image processing techniques (Naz et al., 2009; Naz and Bowling, 2008). The two later studies masked out areas without potential for artificial drainage using a simple decision tree model based on land cover, slope and soil drainage class. Thayn et al. (2011) further developed the use of aerial photography by using photographs taken before and after a rainfall event.

Other studies have favored statistical approaches, usually covering large areas at a coarse resolution. These studies have usually mapped the probability of artificial drainage, rather than the individual drainage pipes. Behrendt et al. (2003) combined information from statistical yearbooks with a survey of expert authorities, while Hirt et al. (2005) combined maps of artificially drained areas with a map of soil classes in order to extrapolate the results.

Feick et al. (2005) mapped the percentages of artificially drained areas of the World on a 5 × 5 minute grid based mostly on international datasets. The final map contained 167 million hectares of artificially drained land globally. Sugg (2007) estimated the percentages of artificially drained areas at county level for 18 states in the USA, using maps of soil drainage classes and the extent of row crops.

Tetzlaff et al. (2009) combined the identification of artificially drained areas from aerial photography with a statistical approach. The authors first identified 2734 artificially drained fields in 231 aerial photographs from northern Germany. The authors then split the study area into 51 parts based on a number of geographic variables and calculated the percentage of artificially drained areas for each of the

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

https://doi.org/10.1016/j.geoderma.2018.01.018
Received 27 October 2017; Received in revised form 12 January 2018; Accepted 16 January 2018
Available online 04 February 2018

parts.

Machine learning present a possible further development of the statistical approaches previously applied. Numerous studies have shown that machine learning models, combining soil observations with maps of known variables, can successfully predict soil properties (McBratney et al., 2003; Scull et al., 2003). The decision to install subsurface drainage systems depends largely on the natural conditions of the soil, and in turn, they affect the surrounding soil and vegetation. It is therefore likely that an approach based on machine learning can predict the extent of artificially drained areas.

Researchers have used a large number of approaches in order to map soil properties. These include discriminant analysis (Bell et al., 1992, 1994; Kravchenko et al., 2002), artificial neural networks (Zhao et al., 2013; Zhao et al., 2008), logistic modelling (Campling et al., 2002) and decision tree analysis (Adhikari et al., 2014; Giasson et al., 2011; Henderson et al., 2005) amongst others.

Some studies have compared methods or variations on methods in order to optimize the predictions (Giasson et al., 2011; Knotters et al., 1995; Zhao et al., 2013). However, ensembles of diverse models can often achieve a better performance than individual, optimal models (Breiman, 1996; Dietterich, 2000). Despite this finding, only few studies have combined predictions from several methods for mapping soil properties. Malone et al. (2014) tested several methods of model averaging for combining the predictions from a disaggregated conventional soil map and a regression tree model. Later, while mapping soil properties globally, Hengl et al. (2017) averaged the predictions of two machine learning models for each soil property in order to avoid overshooting effects of the individual models.

To our knowledge, no studies have combined predictions of soil properties from more than two models. There are several algorithms for creating ensembles with one type of model. Boosting (Freund and Schapire, 1996) successively adds weight to instances with incorrect predictions and builds new models using the weights, while bagging (Breiman, 1996) trains a number of independent models by drawing bootstrap samples from the training data. However, a different approach is necessary in order to combine models of different types.

Caruana et al. (2004) presented a solution in the form of the selective ensemble technique. The technique first trains a 'library' containing a large number of models based on various machine learning algorithms. An algorithm then builds an ensemble by forward stepwise selection of models from the library.

The present study uses the selective ensemble technique for the prediction of artificially drained areas in Denmark. Firstly, we test two ways of avoiding overfitting while selecting models for the ensemble. Secondly, we test an approach for reducing the prediction time of the ensemble by taking the prediction times of the models into account in the selection process.

We base our study on 745 field observations on the presence of artificial drainage and 46 environmental covariates, including soil properties, topographic variables, satellite imagery, land use information and climatic data.

2. Materials and methods

2.1. Study area

Denmark is a country in northern Europe at 54.56–57.75°N and 8.08–15.20°E (Fig. 1) with a total area of 43,000 km^2. The terrain is mostly weakly undulating and flat with a mean elevation of 31 m and a maximum elevation of 171 m. The dominant parent material in the eastern part of the country is loamy Weichselian moraine, while sandy deposits dominate in the western part of the country in the form of Weichselian outwash plains and Saalian moraine. The climate is temperate coastal with temperatures ranging from 1.5 °C in January to 16.3 °C in July in the period 2001–2010. The mean annual precipitation ranges from about 650 mm in the eastern part of the country to about 875 mm in the western part with a mean value of 770 mm (Wang, 2013). The main land use is agriculture, accounting for 66% of the area, while natural vegetation and urban areas make up 16% and 10% of the area, respectively (Statistics Denmark, n.d.).

Olesen (2009) estimated that approximately 50% of the agricultural area of Denmark is artificially drained. Drainage work started in 1848 and occurred mainly during two periods when the government subsidized the work. In the second half of the 19th century, drainage work focused on the loamy moraines of eastern Denmark. However, from the 1930s to the 1970s, most drainage work took place in wetland areas in the western part of the country (Hansen et al., 2004; Madsen, 2010).

2.2. Input data

In this study, the full dataset consisted of 745 point observations of the presence or absence of artificial drainage systems (Fig. 1). The observations were collected for a previous study mapping the probability of artificial drainage (Olesen, 2009). 571 observations were collected from the locations of soil profiles situated in a 7 km grid, while 174 observations were collected from additional sites without soil observations. The dataset contained 401 artificially drained locations and 344 locations without artificial drainage. We used two thirds of the observations for the training dataset and one third for evaluation. The full dataset comprised 46 covariates extracted from map layers (Table 1).

The covariate layers comprised topographic variables, data on soil texture and parent material, satellite imagery and data on land use, cropping history and climate (Table 1). We calculated the topographic variables from a digital elevation model (DEM) with a 30.4-meter grid size aggregated from a model with a 1.6-meter resolution as the mean value of the underlying 19 × 19 grid cells. The topographic variables included the depth to the groundwater interpolated from point observations and extracted from a hydrological model (Henriksen et al., 2012). We transformed the groundwater levels from the original 500-meter resolution of the hydrological model to the 30.4-meter resolution of the DEM by means of bilinear interpolation with the tool *Resample* in ArcMap.

The satellite imagery was a mosaic of Landsat 8 scenes from March 2014, which was the only month with cloud-free images in the entire study area, resampled to the 30.4-meter resolution of the DEM (NASA Landsat Program, 2014). The imagery comprised the surface reflectance from the raw bands as well as normalized indices for vegetation, soil-adjusted vegetation, moisture and water.

The study used maps of clay contents produced by Adhikari et al. (2013), aggregating the original depth intervals into four new intervals as weighed means, and a map of soil drainage classes produced by Møller et al. (2017). Soil data also included rasterized choropleth maps of geology (Jakobsen et al., 2015), landscape elements (Madsen et al., 1992) and wetland areas (Greve et al., 2014).

The land use map consisted of CORINE 2012 data (European Environment Agency, 2014). We based the cropping history on a digital field map with the farmers' registration for the common agricultural policy (The Danish Agricultural Agency, 2014). We divided the crops into categories depending on their drainage dependence, understood as the potential negative effect of water saturation on crop yields. We then counted the number of years with each category in the period 2011–2014. Drainage-dependent crops comprised mostly winter cereals while possibly drainage-dependent crops were in most cases spring cereals. Most of the drainage-independent crops were grasses for various uses. We distinguish between drainage-dependent and possibly drainage-dependent crops because Denmark has a precipitation surplus during the winter. Poor drainage conditions in the soil are therefore more likely to affect winter cereals than spring cereals.

We interpolated the precipitation data from point data by means of kriging.

For the point observations with soil profiles, we used the clay