Geoderma 171-172 (2012) 75–84

Contents lists available at ScienceDirect

Geoderma

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

Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach

Blandine Lemercier [a,b,*], Marine Lacoste [a,b], Macoumba Loum [a,b,c], Christian Walter [a,b]

[a] AGROCAMPUS OUEST, UMR 1069 Sol Agro et hydrosystème Spatialisation, F-35000 Rennes, France
[b] INRA, UMR 1069 Sol – Agro et hydrosystème – Spatialisation, 65 rue de Saint-Brieuc, F-35000 Rennes, France
[c] Gaston Berger University, Geography department, BP 234, Saint-Louis, Sénégal

ARTICLE INFO

Article history:
Received 30 April 2010
Received in revised form 3 December 2010
Accepted 4 March 2011
Available online 22 April 2011

Keywords:
Predictive soil mapping
Boosted classification tree
Parent material
Soil drainage class
Regional scale

ABSTRACT

Digital soil mapping can be helpful in providing pedological information over wide areas where existing soil information is limited. The aim of this study was to predict soil properties at a regional scale by parametrizing soil-landscape models using a machine-learning method recently applied to soil science concerns: boosted classification and regression trees. To examine soil properties interdependence, a two-step approach was tested: first soil parent material (PM), including bedrock formations and superficial deposits, was predicted; then, predicted PM was included as a predictive variable to estimate natural soil drainage (SD). Others predictive variables included environmental data representing known soil-forming factors: terrain attributes (elevation, slope, profile and plan curvatures, sub-watershed hillslope length, hydrological distance from the nearest stream, aspect, relative elevation above the nearest stream and a Compound Index initially proposed by Beven and Kirkby (1979) and modified by Mérot et al. (1995)), geological data, airborne gamma-ray spectrometry (K:Th ratio, deviation from mean K emissions of the related lithological unit) and landscape data (derived from remotely sensed data). The study area is located in Brittany (northwestern France) and covers 4645 km². The training dataset was constructed from existing detailed soil maps (scale 1:25,000) available for 11% of the study area. An additional set of 1148 punctual soil observations spread over the study area represented an independent validation dataset. Based on 20,000 randomly selected pixels from the training area, PM and SD were predicted with overall accuracies of 73 and 70% respectively. While calculated on punctual observations, correct agreement between prediction and observation decreased to 49% for PM and 52% for SD. Predicted PM was the most influential variable for SD prediction, illustrating the relevance of the two-step approach tested. Boosted classification tree appeared to be a particularly adequate and robust procedure for predicting soil properties. Probability of occurrence of the predicted PM was demonstrated to be a relevant indication of prediction quality, allowing distinction between well-predicted and poorly-predicted situations.

1. Introduction

Spatial data, quantitative methods to describe soil patterns and processes, and tools for data management and visualization now are widely available, providing new opportunities to predict soil properties and processes (Grunwald, 2009). Detailed soil maps generally have a limited spatial extension; for example, in Brittany (northwestern France), existing soil maps at 1:25,000 scale cover only 10% of the area. Nevertheless, detailed soil maps constitute a consistent information source on soil spatial distribution. Implicit models of soil spatial organization established by soil surveyors can be retrieved and expressed using environmental attributes (Bui et al., 1999; Bui and Moran, 2003; Lagacherie, 1992; Lagacherie et al., 1995; McKenzie and Ryan, 1999; Moran and Bui, 2002; Walter et al., 2007). Relationships between soil and environmental attributes have been conceptualized and formalized by McBratney et al. (2003) through the model *scorpan*, which referred to seven soil-forming factors: *s*: soil, previously measured or other properties of the soil at a point; *c*: climate, climatic properties of the environment at a point; *o*: organisms, vegetation or fauna or human activity; *r*: topography, landscape attributes; *p*: parent material, lithology; *a*: age, the time factor; and *n*: space, spatial position. This model is used as a framework for predicting soil classes, attributes or functions, taking into account soil forming factors and their interactions.

Techniques for the spatial extrapolation of soil properties or classes include spatial interpolation methods such as geostatistical approaches, or machine-learning methods. Spatial interpolation techniques are useful in soil science studies, but they require fairly dense sampling datasets. Machine-learning uses an algorithm to determine or make explicit the relationship between a response and its predictors. These

* Corresponding author at: AGROCAMPUS OUEST, UMR 1069 Sol Agro et hydrosystème Spatialisation, F-35000 Rennes, France. Tel.: +33 2 23 48 52 29; fax: +33 2 23 48 54 30.
E-mail address: Blandine.Lemercier@agrocampus-ouest.fr (B. Lemercier).

0016-7061/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2011.03.010

approaches can be unsupervised (Odeh et al., 1992) or supervised. Supervised classification methods aim to extract a soil spatial organization model from a learning dataset corresponding to known situations. Extrapolation of this model to unknown situations will provide a soil prediction. Neural networks belonging to supervised classification techniques are useful methods because they are effective, but the resulting models are not explicit and are difficult to interpret (McKenzie and Ryan, 1999). Another supervised approach is decision tree-based models. Decision trees are increasingly used in environmental data analysis and modeling. In this context, decision trees first were applied in ecology using remote sensing data (Lawrence et al., 2004; Lees and Ritman, 1991; Michaelsen et al., 1994; Moisen et al., 2006; Pal and Mather, 2003). Because of their advantages and potential for spatial pattern recognition and modeling, decision trees increasingly are employed to predict soil classes or units (Grinand et al., 2008; Minasny and McBratney, 2007; Scull et al., 2005) or soil properties such as soil drainage class (Cialella et al., 1997), soil carbon (Bou Kheir et al., 2010; Brown, 2007; Brown et al., 2006; Henderson et al., 2005; McKenzie and Ryan, 1999; Vasques et al., 2008), pH, total phosphorus and nitrogen, texture (Henderson et al., 2005), cation exchange capacity (Brown et al., 2006), clay content (Brown et al., 2006; Henderson et al., 2005), soil thickness (Henderson et al., 2005; McKenzie and Ryan, 1999), occurrence or absence of erosion types (Geissen et al., 2007) and soil bulk density (Martin et al., 2009).

Initially developed by Breiman et al. (1984), tree-based models involve a binary recursive partition of the predictor's space into a number of disjoint regions (Elith et al., 2008). In the case of classification trees, the model returns constants fitted to each region corresponding to the most probable class and probabilities of occurrence of each response variable modality. Predictors and split points are chosen to minimize prediction errors. According to Friedman and Meulman (2003), regression and classification trees have some interesting features for predictive learning. They can handle input variables of all types (numeric, ordinal, binary and categorical) equally well. Classification and regression trees are non-parametric; so, no hypothesis is required regarding variable distribution, and no data transformation is needed. Trees are not susceptible to missing values, outlier values or to the presence of data without relation with the response variable, even if they are numerous. Interactions between predictors are taken into account without a priori knowledge. In addition, model outcomes are not susceptible to differing scales of measurement among predictors. Classification and regression trees can be handled with a limited calibration dataset, and results are more explicit and easier to interpret than those of neural networks. Therefore, classification- and regression-tree modeling is a relevant approach for predicting spatial distribution of soil properties or classes. More recently, a boosting procedure (Freund and Schapire, 1996; Friedman, 2001) was added to regression and classification trees to improve accuracy. In boosting, large numbers of relatively simple tree models are adaptively combined to optimize predictive performance. Individual trees gradually are fitted iteratively to the training data to increase the emphasis on observations modeled poorly by the existing collection of trees (Elith et al., 2008). Boosted models can be expressed in the general form (Friedman, 2001):

$$F\left(x;\{\beta_m,a_m\}_1^M\right)=\sum_{m=1}^{M}\beta_m h(x;a_m) \quad (1)$$

where F is a function mapping the predictive variables x to a response variable, that minimizes the expected value of some specified loss functions, $h(x;a)$ represents the "base learner" (i.e. a simple regression or classification tree of input variables with parameters a being the splitting variables, split locations and terminal-node means of the individual trees), M represents the number of iterations (individual trees), and β_m is a weighted coefficient for step m. It has been shown that boosting significantly increases the predictive performance of tree-based models for soil properties (Moran and Bui, 2002; Vasques et al., 2008) or land cover (Lawrence et al., 2004; Pal and Mather, 2003) estimation. Moreover, introduction of stochasticity through the bagging procedure (Breiman, 1996) can improve the accuracy of gradient boosting substantially and also reduces computation time. Stochastic Gradient Boosting is a hybrid of the boosting and bagging approaches (Friedman, 2002). A random sample of the data is selected at each step of the boosting process. Trees constructed during the process are summed, and each observation is classified according to the most common classification among the trees. According to our knowledge, no other method, such as classification trees, can include nonlinear interactions between predictors without knowing interactions a priori, mix qualitative and quantitative predictors, and predict categorical variables.

In this study, a stochastic gradient boosting algorithm fitted to existing detailed soil maps was used to predict categorical soil properties. Indeed, it is expected to be less complex to model and interpret relationships between environmental factors and a single soil property than between environmental factors and soil units that combine several individual soil properties. Moreover, soil map users usually need information about soil properties more than about soil taxonomic units.

Parent material (PM) is a major soil-forming factor and a key soil property influencing most soil properties and potentials, though obtaining information about PM is essential to understand soil behavior. PM is inherited from geological bedrock or superficial deposits covering large areas. Superficial deposits most often are represented poorly in existing geological maps, which provide an erroneous view of PM. Recent and available information sources, such as Digital Elevation Models (DEM) or airborne gamma-ray spectrometry may allow a substantial enhancement of soil-PM knowledge.

All soil textural and structural characteristics, and hence soil permeability, are influenced by the PM. Natural soil drainage (SD), the second soil property of interest in this study, is therefore largely dependent on PM. Following Liu et al. (2008), drainage here refers to the natural ability of soil to allow water to infiltrate and percolate. SD is important, as it directly affects soil agronomic potential, water fluxes in landscapes, genesis of runoff and erosion processes, and nutrients cycling and transport.

The objective of this study was to assess a two-step regional prediction approach of soil properties based on soil-landscape models derived from boosted classification trees fitted to data extracted from detailed maps. First pedological PM was predicted; next, SD was predicted using the previously predicted PM as a predictive variable.

2. Materials and methods

2.1. Study area

The study area is located in France, in northwestern Brittany (Fig. 1). The area covers 4645 km². The climate is oceanic with mean annual rainfall of 650 mm and mean annual temperature of 11.2 °C. This area belongs to the Armorican Massif, a well-marked geological entity characterized by a complex geological history. The diversity of bedrock formations attests to a succession of long sedimentary periods, punctuated by tectonic, metamorphic and plutonic movements. In the north of the study area, loess deposition occurred during the Quaternary, resulting in more or less thick and scattered current patches of silty materials (Haase et al., 2007). The distribution and the thickness of these Quaternary loess (aeolian silt) deposits strongly influence soil properties (Hughes et al., 2009), but their spatial distribution is highly complex; thus, they are of major interest for the prediction of soil spatial distribution. Other superficial formations of interest are alluvial (terraces and modern deposits) and colluvial deposits.

The topography is generally smooth and largely shaped by geological context. The range in elevation is 258 m over the entire study area, but