Available online at www.sciencedirect.com
ScienceDirect
Geoderma 143 (2008) 180–190
GEODERMA
www.elsevier.com/locate/geoderma

Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context

Clovis Grinand, Dominique Arrouays *, Bertrand Laroche, Manuel Pascal Martin

INRA, InfoSol, US 1106, Avenue de la Pomme de Pin, BP 20619, Ardon, F-45166 Olivet Cedex, France

Received 18 December 2006; received in revised form 22 July 2007; accepted 3 November 2007
Available online 3 December 2007

Abstract

This paper aims to investigate the potential of using soil-landscape pattern extracted from a soil map to predict soil distribution at unvisited location. Recent machine learning advances used in previous studies showed that the knowledge embedded within soil units delineated by experts can be retrieved and explicitly formulated from environmental data layers However, the extent to which the models can yield valid prediction has been little studied. Our approach is based on a classification tree analysis which has underwent a recent statistics advance, namely, stochastic gradient boosting. We used an existing soil-landscape map to test our methodology. Explanatory variables included classical terrain factors (elevation, slope, curvature plan and profile, wetness index, etc.), various channels and combinations of channels from LANDSAT ETM imagery, land cover and lithology maps. Overall classification accuracy indexes were calculated under two validation schemes, either taken within the training area or from a separated validation area. We focused our study on the accuracy assessment and testing of two modelling parameters: sampling intensity and spatial context integration. First, we observed strong differences in accuracy between the training area and the extrapolated area. Second, sampling intensity, in proportion to the class extent, did not largely influence the classification accuracy. Spatial context integration by the use of a mean filtering algorithm on explanatory variables increased the Kappa index on the extrapolated area by more than ten points. The best accuracy measurements were obtained for a combination of the raw explanatory dataset with the filtered dataset representing regional trend. However, the predictive capacity of models remained quite low when extrapolated to an independent validation area. Nevertheless, this study offers encouragement for the success of extrapolating soil patterns from existing soil maps to fill the gaps in present soil map coverage and to increase efficiency of ongoing soil survey.

Keywords: Digital soil mapping; Boosted classification tree; Extrapolation; Sampling intensity; Validation; France

1. Introduction

The increasing amount of numerical data combined with fast development of new information processing tools change significantly the way in which information on soils is acquired and managed. The use of Digital Soil Mapping (DSM), based on geographical information science, statistics and pedology (McBratney et al., 2003), is continuously increasing. The generic framework of Digital Soil Mapping has been defined by McBratney et al. (2003) as scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method. It is based on the seven predictive scorpan factors, a generalisation of Jenny's five factors, namely: (1) *s*: soil, other or previously measured attributes of the soil at a point; (2) *c*: climate, climatic properties of the environment at a point; (3) *o*: organisms, including land cover and natural vegetation; (4) *r*: topography, including terrain attributes and classes; (5) *p*: parent material, including lithology; (6) *a*: age, the time factor; (7) *n*: space, spatial or geographic position. Interactions between these factors are also considered. Digital Soil Mapping has been tested in a wide range of soil and scale mapping contexts throughout the world (McBratney et al., 2003; Grunwald, 2006; Dobos et al., 2006). It has been used to understand and quantify the relationships between soils and their environmental attributes, mostly derived from exhaustive and easy-to-access datasets such as Digital Elevation Models (DEM) and remote sensing imagery. Recent soil landscape predictive algorithms such as neural networks, fuzzy logic or tree model tools arose mainly from data-mining and

* Corresponding author. Tel.: +33 238 41 48 02.
E-mail address: dominique.arrouays@orleans.inra.fr (D. Arrouays).

0016-7061/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2007.11.004

machine learning fields, also referred to as knowledge discovery in a database in its overall process (Fayyad et al., 1996).

Digital soil mapping is computer-assisted production of a digital map of soil type and soil properties (Dobos et al., 2006). Predictive soil-landscape mapping usually involves two modelling approaches: (i) grouping observations (pixels) that present homogenous attributes without prior knowledge (unsupervised classification), or (ii) training the model on known soil type observations (supervised classification). Irvin et al. (1997) compared those techniques for soil-mapping purposes. Supervised approaches usually produce more suitable results as they use available knowledge (training data) to fit the model and thus lead to more interpretable maps. As related by Lagacherie (2002), in a supervised classification process one makes the implicit hypothesis that a soil-landscape is structured in such way that it can be repeatedly predicted from a specific combination of soil-forming factors. Therefore, a forthcoming question is to what extent this hypothesis is true? An attempt to digitally identify a relevant area for extrapolation was carried out by thresholding a measure of distance to delineate a representative area where the model could be applied, based on elevation, slope and geology layers (Lagacherie et al., 2001). Another approach involved the rule induction process to detect a relevant physiographic region using entire reference areas as single separate target classes (Bui and Moran, 2003). A key issue is to have an independent and representative test sample to get relevant estimates of the extrapolation accuracy (Dobos et al., 2006). Land cover mapping by remote sensing faces similar issues (Foody, 2002). More generally, soil landscape prediction from existing maps involves recovering the mental model used by the soil surveyor to set up the map (Lagacherie et al., 1995; Bui, 2004). This is a reverse soil mapping process and has broad relevance to any other application of knowledge discovery from natural resource maps (Qi and Zhu, 2003).

Classification tree analysis (CTA) is a modelling technique that is being used increasingly (Lawrence et al., 2004). CTA has several advantages that seem to suit well soil-landscape modelling applications. One of the most interesting features is that they are non-parametric, which means that no assumption is made regarding variable distribution. Thus, it avoids variable transformation caused, for instance, by bi-modal or skewed histograms, which are frequent in soil class signatures. They are non-sensitive to missing data, perform automatic variable subset selection, are not sensitive to the inclusion of a large number of irrelevant variables, and finally, they can handle quantitative and categorical data, making it possible to integrate DEM-derived variables, remote sensing bands or indexes together with geology or land cover categorical layers. Efficiency of using CTA for predictive soil landscape mapping was demonstrated in a few studies at regional and subregional scale (Moran and Bui, 2002; Scull et al., 2005). Recent studies showed their potential for land cover mapping from remote sensing images analysis (Lawrence et al., 2004; Friedl and Brodley, 1997) and geomorphological mapping (Luoto and Hjort, 2005).

As mentioned by Luoto and Hjort (2005), CTA was practically used in two linked but distinct purposes: induction and prediction. Induction-oriented studies used CTA to uncover the relationship between soil units or properties and environmental attributes, to

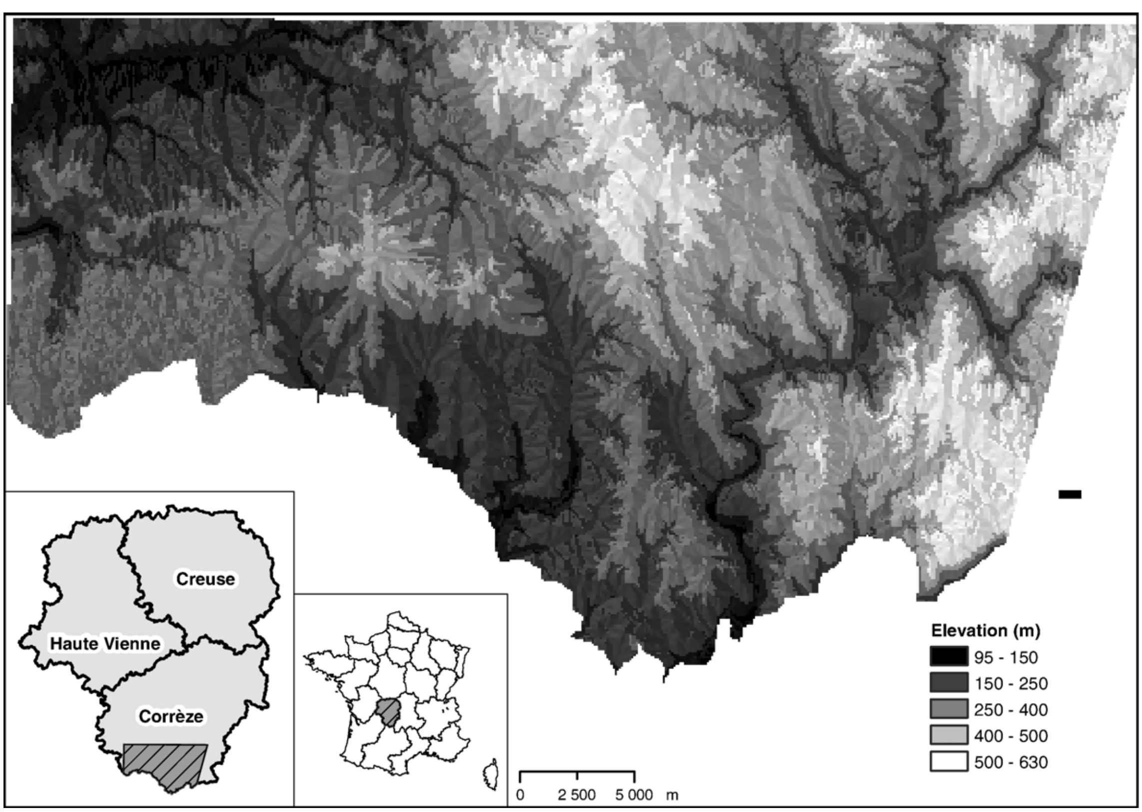

Fig. 1. Location of the study area.