Geoderma Regional 1 (2014) 21–30

Contents lists available at ScienceDirect

Geoderma Regional

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

Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France)

Fanny Collard [a], Bas Kempen [b], Gerard B.M. Heuvelink [b], Nicolas P.A. Saby [a], Anne C. Richer de Forges [a], Sébastien Lehmann [a], Pierre Nehlig [c], Dominique Arrouays [a,*]

[a] INRA, InfoSol Unit, US 1106, 45075 Orléans, France
[b] ISRIC-World Soil Information, PO Box 353, 6700 AJ Wageningen, The Netherlands
[c] BRGM, Direction des Géoressources, UMR 7327, BRGM-CNRS-Université d'Orléans, F-45060 Orléans, France

ARTICLE INFO

Article history:
Received 6 May 2014
Received in revised form 16 July 2014
Accepted 21 July 2014
Available online 23 July 2014

Keywords:
Digital soil mapping
Validation
Classification trees
Random forest
Multinomial logistic regression
Probability sampling
France

ABSTRACT

Reconnaissance soil maps at 1:250,000 scale are the most detailed source of soil information for large parts of France. For many environmental applications, however, the level of detail and accuracy of these maps is insufficient. Funds are lacking to refine and update these maps by traditional soil survey. In this study we investigated the merit of digital soil mapping to refine and improve the 1:250,000 reconnaissance soil map of a 1580 km^2 area in Haute-Normandie, France. The soil map was produced in 1988 and distinguishes nine soil class units. The approach taken was to predict soil class from a large number of environmental covariates using regression techniques. The covariates used include DEM derivatives, geology and land cover maps. Because very few soil point observations were available within the area, we calibrated the regression model by sampling the soil map on a grid. We calibrated three models: classification tree (CT), multinomial logistic regression (MLR) and random forests (RF), and used these models to predict the nine soil classes across the study area. The new and original maps were validated with field data from 123 locations selected with a stratified simple random sampling design. For MLR, the estimate of the overall purity was 65.9%, while that of the reconnaissance map was 55.5%. The difference between the purity estimates of these maps was statistically significant ($p = 0.014$). The significant improvement over the existing soil map is remarkable because the regression model was calibrated with the existing soil map and uses no additional soil observations.

1. Introduction

Reconnaissance soil maps at 1:250,000 scale are the most detailed source of soil information for large parts of France. The geographical coverage of 1:250,000 soil maps in mainland France is about 75% of the territory, while more detailed soil maps only cover about 35% of the country. For many environmental applications (e.g., threats to water quality, pollution of soils, soil erosion by water or wind, loss of, or damage to, rare soils, loss of terrestrial carbon store, loss of soil biodiversity; see a list of applications in France in Richer de Forges and Arrouays (2010)), however, the level of detail and accuracy of 1:250,000 maps is insufficient. Funds are lacking to refine and update these maps by traditional soil survey. This lack of detailed soil data and funds to increase resolution and accuracy through conventional soil survey is widely spread over the world (Hartemink, 2008).

Digital soil mapping (DSM) techniques (McBratney et al., 2003) have been proposed as a tool to update (Kempen et al., 2009) or disaggregate soil class maps (Häring et al., 2012; Nauman and Thompson, 2014; Subburayalu et al., 2014; Odgers et al., 2014), or to create new maps (Adhikari et al., 2014). Kempen et al. (2012a) show that DSM can be an efficient alternative to traditional soil survey for updating soil class maps. Various methods for calibration and mapping using DSM have been used, including expert based rules (Lagacherie et al., 1995; van Zijl et al., 2014), fuzzy logic systems (MacMillan et al., 2007; Zhu et al., 2001; Yang et al., 2011), neural networks (Behrens et al., 2005) and various methods of classification and regression (Carré and Girard, 2002; Grinand et al., 2008; Kempen et al., 2009; Häring et al., 2012; Adhikari et al., 2014; Nauman and Thompson, 2014; Subburayalu et al., 2014; Odgers et al., 2014).

DSM models are typically calibrated with observed point data (e.g., Häring et al., 2012; Kempen et al., 2012b; Adhikari et al., 2014). However, when resources for collecting new field point data are limited, obtaining a calibration dataset by sampling an existing soil map might be an attractive alternative, even though mapped soil properties and soil types are no substitute for real observations. This approach is taken by, for example, Lagacherie et al. (1995), Grinand et al. (2008), Debella-Gilo and Etzelmüller (2009), and more recently by Nauman and Thompson (2014), Subburayalu et al. (2014) and Odgers et al. (2014). However, some of these studies did not validate the resulting maps with independent field data (Debella-Gilo and Etzelmüller,

* Corresponding author.

http://dx.doi.org/10.1016/j.geodrs.2014.07.001
2352-0094/

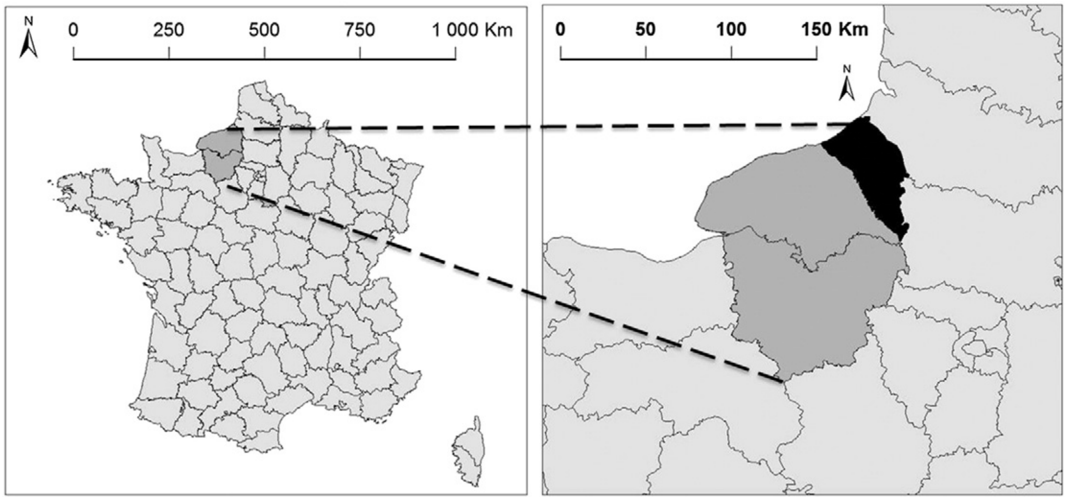

Fig. 1. Location of the study area.

2009; Grinand et al., 2008; Lagacherie et al., 1995), which makes it difficult or impossible to assess their accuracy. Others focused on a single prediction method (Grinand et al., 2008; Häring et al., 2012; Adhikari et al., 2014; Subburayalu et al., 2014; Odgers et al., 2014), and only Nauman and Thompson (2014) compared the accuracy of the digital, disaggregated soil class map with the legacy soil map. And none of these studies used independent validation collected with a probability sampling design that allows for statistically valid and unbiased accuracy assessment and model comparison.

In this paper we use multinomial logistic regression and two tree-based methods (classification trees and random forests) to investigate the merit of DSM to refine and improve the 1:250,000 reconnaissance soil map of a 1580 km^2 area in Haute-Normandie, France. We sampled the reconnaissance map and used this sample to calibrate the prediction models. Ground truth validation data were collected using probability sampling to evaluate whether i) the pedometric soil maps are more accurate than the original map, and ii) there are differences in accuracy between the three pedometric methods. This will provide insight if this method is an attractive alternative to traditional soil survey for updating and upgrading soil class maps in France.

2. Materials and methods

2.1. Study area

The study area is located in North-West France, along the Channel coast (Fig. 1). In this region, the parent materials are mainly loess deposits, chalk, sands and clays and more locally sand and gravel from alluvial deposits. Two main loess plateaus are located in the east and the north-west with elevations ranging from 200 m to 250 m and from 80 m to 180 m, respectively. Their land use is mainly intensive agriculture. Chalk soils occur mainly on steep slopes surrounding the plateau and are mostly occupied by forest. The south-eastern part is characterized by gently undulating relief. The soils are developed on sands and clays and land use is mainly permanent grassland. The climate is oceanic. The mean annual temperature is about 9 °C and the total annual precipitation is about 800 mm. A description of the nine soil classes of the 1:250,000 reconnaissance soil map of the area (Wolf et al., 1998) is given in Table 1.

2.2. Environmental ancillary data

Classical relief attributes were derived from the SRTM 90 m DEM.[1] Parent material was represented by a harmonized 1:50,000 lithological map that was synthesized from all geological surveys available for the region (Quesnel et al., 2007; Van Lint et al., 2003). Land use information was provided by the Corine Land Cover 2006 European database (Commission of the European Community, 1993) and climate information by the Ecoclimap database, a global database of land surface parameters at 1 km resolution (Masson et al., 2003). An exhaustive list of the 19 covariates used and their resolution or map scale is given in Table 2. Several of the DEM-derived covariates are (strongly) mutually correlated. Furthermore, cross-tabulating the categorical covariates with the reconnaissance soil map (from which the calibration points are derived) shows presence of zero-cell counts (Hosmer and Lemeshow, 2000). This means that the frequency distributions in the cross-table contain one or more zeros, i.e. not all combinations of predictor categories and soil classes occur. The presence of zero-cell counts causes numerical instabilities during modeling and should be avoided (Hosmer and Lemeshow, 2000). Hosmer and Lemeshow (2000) suggest combining classes of the categorical predictors in a sensible way to handle the zero-cell problem. However, this does not solve the issue about the correlated covariates. We, therefore, decided to convert the 19 covariate layers to 59 principal components (each class of the categorical covariates becomes one component after transformation), which are candidate predictors for the models.

The 1:250,000 reconnaissance soil map and the geological map were rasterized to 90 m resolution grids, corresponding to the resolution of the SRTM-derived terrain parameter grids. The Ecoclimap was resampled from 1 km to 90 m resolution.

2.3. Soil point observations

The point dataset for model calibration was obtained by sampling the reconnaissance soil map using a systematic, square grid with a random origin and 500 m grid spacing. The soil class was extracted at the grid nodes, which resulted in a sample of 6323 points.

2.4. Models

Three different methods were applied: multinomial logistic regression (MLR), classification tree modeling (CT), and random forests (RF).

2.4.1. Multinomial logistic regression

The logistic model belongs to the family of generalized linear models and is used when the response variable is categorical (Hosmer and Lemeshow, 2000). Suppose that variable Y represents the observed soil class at a sampling location, which can assume any of K categories, where K is the number of soil classes. In case K equals 2, Y has a

[1] http://srtm.csi.cgiar.org/.