Geoderma 223-225 (2014) 97–107

Contents lists available at ScienceDirect

Geoderma

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

Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale

M.P. Martin [a,*], T.G. Orton [b], E. Lacarce [a], J. Meersmans [c], N.P.A. Saby [a], J.B. Paroissien [a], C. Jolivet [a], L. Boulonne [a], D. Arrouays [a]

[a] INRA, US1106 Unité Infosol, F-45000 Orléans, France
[b] Faculty of Agriculture and Environment, The University of Sydney, 1 Central Avenue, Australia Technology Park, Eveleigh, NSW 2015, Australia
[c] Department of Geography, College of Life and Environmental Sciences, University of Exeter, Amory building — room 431, Rennes Drive, EX4 4RJ Exeter, UK

ARTICLE INFO

Article history:
Received 24 December 2012
Received in revised form 20 December 2013
Accepted 2 January 2014
Available online 28 February 2014

Keywords:
Soil organic carbon
Spatial distributions
Boosted regression trees
Geostatistics
National accounting

ABSTRACT

Soil organic carbon (SOC) plays a major role in the global carbon budget. It can act as a source or a sink of atmospheric carbon, thereby possibly influencing the course of climate change. Improving the tools that model the spatial distributions of SOC stocks at national scales is a priority, both for monitoring changes in SOC and as an input for global carbon cycles studies. In this paper, we compare and evaluate two recent and promising modelling approaches. First, we considered several increasingly complex boosted regression trees (BRT), a convenient and efficient multiple regression model from the statistical learning field. Further, we considered a robust geostatistical approach coupled to the BRT models. Testing the different approaches was performed on the dataset from the French Soil Monitoring Network, with a consistent cross-validation procedure. We showed that when a limited number of predictors were included in the BRT model, the standalone BRT predictions were significantly improved by robust geostatistical modelling of the residuals. However, when data for several SOC drivers were included, the standalone BRT model predictions were not significantly improved by geostatistical modelling. Therefore, in this latter situation, the BRT predictions might be considered adequate without the need for geostatistical modelling, provided that i) care is exercised in model fitting and validating, and ii) the dataset does not allow for modelling of local spatial autocorrelations, as is the case for many national systematic sampling schemes.

1. Introduction

Soils are the second biggest carbon pool of the planet, containing about 1500 Pg C (Batjes, 1996; Eswaran et al., 1993; Post et al., 1982). As such, their behaviour as a greenhouse gas source and sink needs to be quantified, when facing climate change induced by increasing atmospheric greenhouse gases concentrations (Batjes, 1996; Lal, 2004). Quantifying temporal changes of this pool requires estimating its spatial distribution at different dates and at various scales, with the national scale being of particular importance for international negotiations. The reliability of such estimates depends upon suitable data in terms of organic carbon content and soil bulk density and on the methods used to upscale point data to comprehensive spatial estimates. These estimates may also be used for defining the baseline state for soil organic carbon (SOC) change simulations (van Wesemael et al., 2010), or setting some of the parameters for models of SOC dynamics (Tornquist et al., 2009).

Interestingly, there is quite a diversity regarding the nature of the models used for upscaling SOC point measurements to the national level. The validity of each method depends on the datasets and on the scale (defined by its grain or precision and extent, Turner et al., 1989). The mapping approaches range from simple statistics or pedotransfer rules, relating SOC contents or stocks to soil type (Yu et al., 2007) or soil type and land use (Arrouays et al., 2001; Tomlinson and Milne, 2006), to multivariate regression models (Meersmans et al., 2008, with multiple linear models and Yang et al., 2008, with generalized linear models or Suuster et al., 2012, with mixed models). Recent studies have used techniques adapted from the data mining and machine learning literature, with piecewise linear tree models (Bui et al., 2009) or multiple regression trees for regional studies (Grimm et al., 2008; Lo Seen et al., 2010; Suuster et al., 2012). Among the studies considering small extent (<50 km^2), many have considered the use of geostatistics, some including SOC predictors *via* cokriging (CK) or regression kriging (RK) (Don et al., 2007; Mabit and Bernard, 2010; Rossi et al., 2009; Spielvogel et al., 2009; Yun-Qiang et al., 2009). As the extent increases, the use of geostatistics becomes less common and despite the spatial dimension of such studies, few geostatistical approaches for SOC mapping have been proposed for use at the national scale (but see Chaplot et al., 2009; Kerry et al., 2012; Rawlins et al., 2009).

SOC mapping for France has been performed, during the last decade, by using class specific SOC means (Arrouays et al., 2001) or regression models (Martin et al., 2011; Meersmans et al., 2012). The most recently

* Corresponding author.
E-mail address: manuel.martin@orleans.inra.fr (M.P. Martin).

0016-7061/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2014.01.005

proposed models are still not able to fully satisfactorily predict SOC stocks or contents on independent locations: R^2 reached 0.50 and 0.49 and root mean squared prediction errors (RMSPE) 2.27 kg/m^2 and 1.45%, for Martin et al. (2011) on SOC stocks and for Meersmans et al. (2012) on SOC contents, respectively. Martin et al. (2011) obtained unbiased predictions (the bias was estimated to be -0.002 kg/m^2 by cross-validation), which might ensure unbiased mapping of the stock at the national level. Nevertheless, these R^2 and RMSPE results showed that there is potentially room for improvement, especially if one is willing to use such models for regional assessments. Adding spatial autocorrelation terms in these models might be a way to improve their performance.

Recently, new approaches have been proposed for coupling regression models, relating environmental factors to the studied property, with geostatistical models, representing the spatial autocorrelation among the observations (Marchant et al., 2010). Such methods were also designed to handle local anomalies (*i.e.* outliers). Nevertheless, these methods do not currently include some features that other statistical models, such as boosted regression trees (BRT) used by Martin et al. (2011), have (*i.e.* handling nonlinear relationships between qualitative and quantitative predictors and the independent variable, nonlinear interactions between the predictors, in an automated manner). Both approaches share the robustness to the presence of outliers in the dataset. As they are tackling different problems, the spatial autocorrelation for the geostatistical approaches, and the modelling of the complex interactions between SOC stocks and their drivers for the regression methods, both might be considered as complementary.

The aim of this paper is to combine these recent robust geostatistical approaches with the BRT models currently applied to map SOC stocks at the national scale for France. We apply the methods to a dataset of 2166 paired observations of SOC and bulk densities from the French soil quality monitoring network (RMQS). We use this study to assess the modelling methods in order to determine i) how useful it is to combine BRT and geostatistical modelling, and ii) if any advantages are dependent on the number of ancillary variables included as predictors in the BRT models. The aim is not specifically to study the relative importance of SOC stock drivers for France (Martin et al., 2011; Meersmans et al., 2012), nor to produce a new map of SOC stocks in France.

2. Materials and methods

2.1. Data

Soil organic carbon stocks were computed for 2166 sites from the French soil quality monitoring network (RMQS) (Fig. 1). The network is based on a 16 km × 16 km square grid. The sampling sites are located at the centre of each grid cell, except when settling a homogeneous 20 m × 20 m sampling area is not possible at this specific location (because of the soils being sealed or strongly disturbed by anthropogenic activities, for instance). In that case, another site is selected within 1 km from the centre of the cell depending on soil availability for sampling (Arrouays et al., 2002). Some of the 2166 sites of our dataset were actually replicates of the regular cells sites: some cells had two sites located in them, one close to the centre of the cell as described above, and another one located at another position within the cell.

At each site, 25 individual core samples were taken from the (0–30 cm) and the subsoil (30–50 cm) using a hand auger according

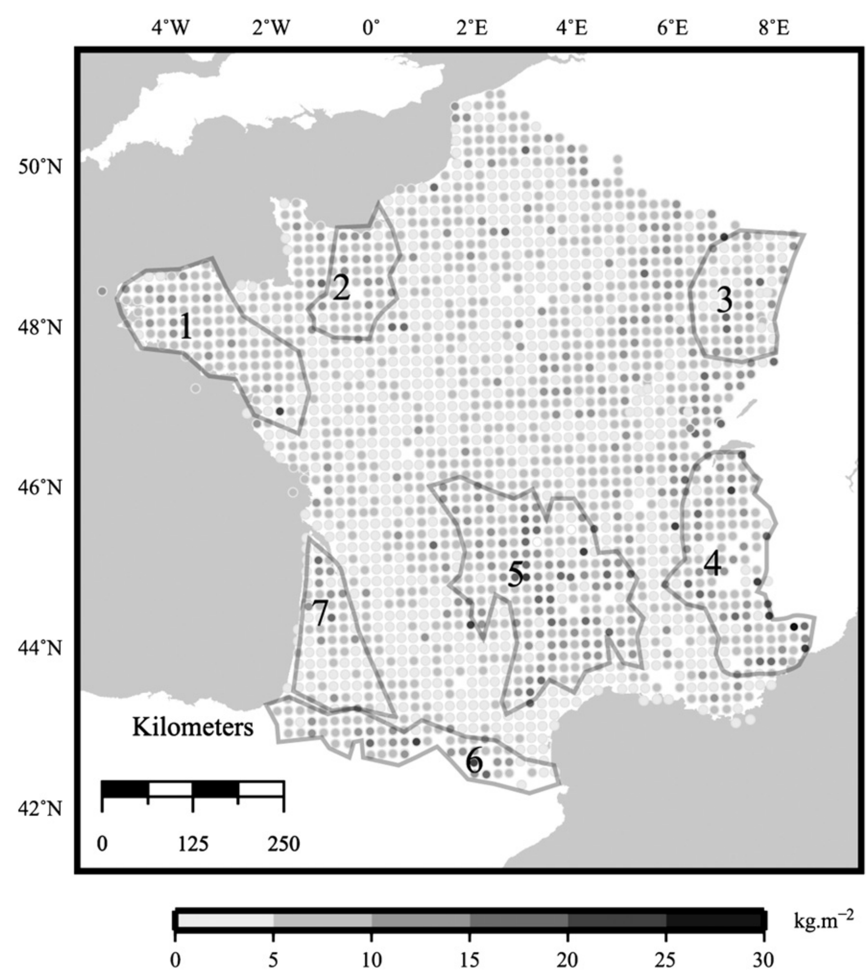

Fig. 1. SOC stocks (0–30 cm) values on the French monitoring network, which were used in the present study. Areas from 1 to 7 represent various different areas that are mentioned later in the text. 1: south–west Brittany. 2: part of Basse Normandie. 3: Alsace and part of Lorraine. 4: part of French Alps. 5: Massif Central. 6: French Pyrenean mountain range. 7: part of Aquitaine.