Contents lists available at ScienceDirect

Ecological Indicators

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

Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem

Ren-Min Yang[a,b], Gan-Lin Zhang[a,b,*], Feng Liu[a], Yuan-Yuan Lu[a,b], Fan Yang[a,b], Fei Yang[a,b], Min Yang[a,b], Yu-Guo Zhao[a], De-Cheng Li[a]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] University of the Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 10 April 2015
Received in revised form 17 August 2015
Accepted 21 August 2015
Available online 9 September 2015

Keywords:
Digital soil mapping
Soil organic carbon
Boosted regression tree
Random forest
Environmental variables
Tibetan Plateau

ABSTRACT

Soil organic carbon (SOC) plays an important role in soil fertility and carbon sequestration, and a better understanding of the spatial patterns of SOC is essential for soil resource management. In this study, we used boosted regression tree (BRT) and random forest (RF) models to map the distribution of topsoil organic carbon content at the northeastern edge of the Tibetan Plateau in China. A set of 105 soil samples and 12 environmental variables (including topography, climate and vegetation) were analyzed. The performance of the models was evaluated using a 10-fold cross-validation procedure. Maps of the mean values and standard deviations of SOC were generated to illustrate model variability and uncertainty. The results indicate that the BRT and RF models exhibited very similar performance and yielded similar predicted distributions of SOC. The two models explained approximately 70% of the total SOC variability. The BRT and RF models robustly predicted the SOC at low observed SOC values, whereas they underestimated high observed SOC values. This underestimation may have been caused by biased distributions of soil samples in the SOC space. Vegetation-related variables were assigned the highest importance in both models, followed by climate and topography. Both models produced spatial distribution maps of SOC that were closely related to vegetation cover. The SOC content predicted by the BRT model was clearly higher than that of the RF model in areas with greater vegetation cover because the contributions of vegetation-related variables in the two models (65% and 43%, respectively) differed significantly. The predicted SOC content increased from the northwestern to the southeastern part of the study area, average values produced by the BRT and RF models were 27.3 g kg^{-1} and 26.6 g kg^{-1}, respectively. We conclude that the BRT and RF methods should be calibrated and compared to obtain the best prediction of SOC spatial distribution in similar regions. In addition, vegetation variables, including those obtained from remote sensing imagery, should be taken as the main environmental indicators and explicitly included when generating SOC maps in Alpine environments.

1. Introduction

Soil is an important and the largest reservoir of organic carbon in terrestrial ecosystems (Batjes, 1996). Soil stores more carbon than the atmosphere and vegetation and thus plays an important role in the global carbon cycle (Bohn, 1982; Schlesinger, 1997; Grace, 2004). An important function of the soil organic carbon (SOC) pool is its role as a potential sink of greenhouse gases (Davidson and Janssens, 2006; Gal et al., 2007). Minor changes in the amount of SOC could greatly affect atmospheric CO_2 concentrations due to its sensitivity to climate changes and human activities (Bellamy et al., 2005). In addition, SOC is closely related to the soil quality, fertility, biological processes, structure and hydraulic properties of soil (Zhang et al., 2006). Therefore, a better understanding of SOC content and its controls is necessary for soil resource management and sustainable usage. Furthermore, accurate estimates of SOC are quite essential for analyzing regional carbon cycling and potential responses of soils to global environmental change.

It is difficult to sample and analyze a large number of points and then map the distribution of SOC across large areas, particularly in areas of rugged terrain such as Alpine environments. Digital soil mapping (DSM) is an efficient method for predicting soil properties and classes over large areas based on discrete samples (McBratney et al., 2003). Most DSM methods were

* Corresponding author at: State Key Laboratory of Soil and Sustainable Agriculture Institute of Soil Science, Chinese Academy of Sciences, 71 East Beijing Road, Nanjing 210008, China. Tel.: +86 25 8688 1279; fax: +86 25 8688 1000.
E-mail address: glzhang@issas.ac.cn (G.-L. Zhang).

http://dx.doi.org/10.1016/j.ecolind.2015.08.036
1470-160X/© 2015 Elsevier Ltd. All rights reserved.

developed based on a soil-landscape model that characterizes soils as a function of environmental variables including climate, biota, relief, parent material and time (Jenny, 1941). Numerous techniques have been used to SOC estimations, including multiple linear regression (Arrouays et al., 1995), linear mixed models (Zhao et al., 2014), artificial neural networks (Minasny et al., 2006), support vector machines (Were et al., 2015), kriging (Mishra et al., 2009), the boosted regression tree (BRT) model (Martin et al., 2011) and the random forest (RF) model (Grimm et al., 2008). Despite the various applications of statistical techniques in SOC estimations, comparisons of various methods are still rarely reported in the literature (i.e., Were et al., 2015).

Of these DSM techniques, tree models have been widely used to map the spatial distribution of SOC (Henderson et al., 2005; Grimm et al., 2008; Martin et al., 2011). In general, appropriate single trees are difficult to construct for several reasons, including incorrect parameter settings, simplicity rules and tree instability, and such issues have led to the development of bagging, boosting and random methods to improve predictive performance (Skurichina and Duin, 2002). In bagging, the models are fitted using random independent bootstrap replicates and are then combined by averaging the output for regression (Efron and Tibshirani, 1993). The BRT and RF models are two relatively new tree-based models that have been developed to optimize predictive performance by combining a large number of simple trees into a powerful model rather than using a single tree model based on traditional regression trees (Breiman, 2001; Skurichina and Duin, 2002; Friedman, 2001, 2002). In the BRT model, the fitted model is a simple linear combination of many trees that are fitted iteratively and boosted to reweight poorly modeled observations (Elith et al., 2008). The RF model is constructed in a random vector of the data feature space sampled independently (Breiman, 2001). Being data mining methods, the BRT and RF models have several common advantages, including a limited number of user-defined parameters and the ability to model non-linear relationships, manage qualitative and quantitative variables, remain robust despite missing data and outliers, reduce overfitting, and evaluate, summarize and interpret final models (Breiman, 2001; Friedman and Meulman, 2003).

Owing to these merits, BRT and RF models have been widely applied in various scientific fields, including ecological modeling (Peters et al., 2008; T. Froeschke and F. Froeschke, 2011), remote sensing (Lawrence et al., 2004, 2006; Pouteau et al., 2011), environmental science (Carslaw and Taylor, 2009), epidemiology (Friedman and Meulman, 2003) and digital soil carbon mapping (Grimm et al., 2008; Martin et al., 2011; Wiesmeier et al., 2011; Sreenivas et al., 2014; Wiesmeier et al., 2014). However, a comparison of the performance of BRT and RF models has not yet been attempted in recent SOC mapping studies.

Therefore, we evaluated the performance of and differences between the BRT and RF models in mapping the variability of organic carbon content in the topsoil (0–20 cm) at the northeast edge of the Tibetan Plateau in China. The specific objectives were to (1) develop BRT and RF models to predict the SOC content based on 105 soil samples and 12 environmental variables, (2) quantify the effects of various environmental variables on the SOC variation, and (3) map the spatial distribution of SOC by comparing the predictive qualities of the BRT and RF models.

2. Materials and methods

2.1. Study area

The study area measures approximately 30,000 km^2 and is located in northwestern China, specifically on the northeastern edge of the Tibetan Plateau (latitude 37.71°–40.03° North, longitude 96.78°–101.2° East) (Fig. 1). This region is dominated by the Qilian Mountains, which range in elevation from 1684 to 4600 m above sea level. The study area has a typical plateau continental climate, and the mean annual temperature (MAT) and mean annual precipitation (MAP) range from −12.3 to 6.6 °C and 72 to 480 mm, respectively. The vegetation consists of alpine grasslands. Grassland types, in order of descending elevation, vary as follows: cold desert alpine meadow, sub-alpine shrub grassland, mountain forest grassland, dry shrub grassland and desert grassland (Jin et al., 2009). The study area has a long history of land use as pasture. Due to Chinese policies regarding environmental protection, effects of human activities on land use in this area are very limited in recent years. The dominant soil types are Inceptisols and Mollisols, according to Soil Taxonomy (Soil Survey Staff, 2014).

2.2. Datasets

2.2.1. Soil samples

Soil surveys were conducted in 2012 and 2013 and produced 105 soil sampling profiles (Fig. 1). In the study area in the Tibetan Plateau, field sampling is difficult due to access constraints. To characterize high spatial variability of soil properties on such a large scale, a sampling strategy needs to be carefully considered. A purposive sampling strategy (Zhu et al., 2008) was used in our study. Briefly, sample sites were selected based on the variability of soil-forming factors that were expected to represent the heterogeneity of the soil in the study area, including elevation, climate, land use and parent material. This approach can result in a small number of typical soil samples. In addition, the accessibility of each sample site was evaluated based on traffic data to improve the sampling efficiency. Soil profiles were described down to a depth of 1.2 m or to bedrock. Samples of pedologic horizons were collected for physical and chemical analyses (Cooperative Research Group on Chinese Soil Taxonomy, 2001). Each sample contained approximately 1 kg of soil. In the laboratory, the samples were air dried and passed through a 2-mm sieve. Analyses for SOC content (g kg^{-1}) were performed using the classic Walkley-Black method (Zhang and Gong, 2012). The SOC content of the topsoil (i.e., depths of 0–20 cm) was then calculated using a depth-weighted average function for each profile and log-transformed to improve linear modeling by minimizing the rightward skew of the untransformed variable.

2.2.2. Environmental variables

Environmental variables were collected and transferred to a geographic information system (GIS) raster layer at a 90-m resolution using ArcGIS 10.0 (ESRI Inc., USA). To accurately map the SOC spatial distribution, environmental predictors with a high resolution are necessary. Although higher spatial resolution would provide more-detailed information (i.e., resolution as good as 10 m), predictors at a 90-m resolution used in this study are acceptable considering the widespread extent of the data and the low computational efficiency that would result from the use of a large data set. The relationships between SOC and the environmental variables are presented in Fig. 2. These variables are listed below.

1. A digital elevation model (DEM) and its derivatives. A 90 m resolution freely available DEM from the Shuttle Radar Topography Mission (SRTM, version 4.1) was used. Five derivatives were determined, including elevation, slope, aspect, catchment area (CA) and SAGA topographic wetness index (TWI). SAGA topographic wetness index is based on a modified catchment area, and it tends to predict a more realistic and higher potential soil wetness than conventional topographic wetness index for cells situated in valley floors with a small vertical distance to a channel (Boehner et al., 2002). Aspect was expressed in absolute values in the range of 0 to 180°, representing north and south,