Geoderma 286 (2017) 35–45

Contents lists available at ScienceDirect

Geoderma

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

Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region

Calogero Schillaci [a,b], Luigi Lombardo [c,b], Sergio Saia [d,*], Maria Fantappiè [e], Michael Märker [b,f], Marco Acutis [a]

[a] Department of Agricultural and Environmental Science, University of Milan, Italy
[b] Department of Geosciences, University of Tübingen, Germany
[c] Physical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Jeddah, Saudi Arabia
[d] Council for Agricultural Research and Economics (CREA), Cereal Research Centre (CREA-CER), Foggia, Italy
[e] Council for Agricultural Research and Economics (CREA), Centre for Agrobiology and Pedology (CREA-ABP), Florence, Italy
[f] Department of Earth and Environmental Sciences, University of Pavia, Italy

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form 10 October 2016
Accepted 17 October 2016
Available online xxxx

Keywords:
Soil organic carbon
Stochastic modelling
Terrain analysis
Remote sensing
Climate data

ABSTRACT

Efficient modelling methods to assess soil organic carbon (SOC) stocks have a pivotal importance as inputs for global carbon cycle studies and decision-making processes. However, laboratory analyses of SOC field samples are costly and time consuming. Global-scale estimates of SOC were recently made according to categorical variables, including land use and soil texture. Remote sensing (RS) data can contribute to the better modelling of the spatial distribution of SOC stock at a regional scale. In the present study, we used Stochastic Gradient Treeboost (SGT) to estimate the topsoil (0–30 cm) SOC stock of a Mediterranean semiarid area (Sicily, Italy, 25,286 km^2). In particular, our study examined agricultural lands, which represent approximately 64% of the entire region. An extensive soil dataset (2202 samples, 1 profile/7.31 km^2 on average) was acquired from the soil database of Sicily. The georeferenced field observations were intersected with remotely sensed environmental data and other spatial data, including climatic data from WORLDCLIM, land cover from CORINE, soil texture, topography and derived indices. Finally, the SGT was compared to published global estimates (GSOC) and data from the International Soil Reference and Information Centre (ISRIC) Soil Grids by comparing the pseudo-regressions of the SGT, GSOC and ISRIC with soil observations. The mean SOC stock across the entire region that was estimated by GSOC and ISRIC was 3.9% lower and 46.2% higher compared to the SGT. The SGT efficiently predicted SOC stocks that were <70 t ha^{-1} (corresponding to the 90th percentile of the observed values). On average, the coefficient of variation of the SGT model was 3.6% when computed on the whole dataset and remained lower than 23% when computed on a distribution basis. The SGT mean absolute error was 14.84 t ha^{-1}, 18.4% and 36.3% lower than GSOC and ISRIC, respectively. The mean annual rainfall, soil texture, land use, mean annual temperature and Landsat 7 ETM+ panchromatic Band 8 were the most important predictors of SOC stock. Finally, SOC stocks were estimated for each land cover class. SGT predicted SOC stock better than GSOC and ISRIC for most data. This resulted in a percentage of data in the prediction confidence interval $\pm$ 50% compared to the observed values of 71.4%, 65.8%, and 50.7% for SGT, GSOC, and SGT, respectively. This consisted of a higher R^2 and a slope (β) that was closer to 1 for the pseudo-regression constructed with SGT compared to GSOC and ISRIC. In conclusion, the results of the present study showed that the integration of RS with climatic and soil texture spatial data could strongly improve SOC prediction in a semi-arid Mediterranean region. In addition, the panchromatic band of Landsat 7 ETM+ was more predictive compared to the conventionally used NDVI. This information is crucial to guiding decision-making processes, especially at a regional scale and/or in semi-arid Mediterranean areas. The model performance of the SGT could be further improved by adopting predictors with greater spatial resolutions. The results of the present experiment yield valuable information, especially for assessing climate change or land use change scenarios for SOC stocks and their spatial distribution.

1. Introduction

Agricultural land plays a pivotal role in terms of carbon sequestration ability due to soil organic carbon (SOC) being in both topsoil and subsoil. SOC is recognized as the most important indicator of soil quality and determines plant productivity (Lal, 2004) through a wide range of mechanisms, including its activity as a main source of energy for microbial processes (Hudson, 1994); soil cation exchange capacity (Chan et al., 1992; Riffaldi et al., 1994); the effect on water holding capacity and infiltration rate through the soil profile (MacRae and Mehuys, 1985); and the reduction of the soil bulk density and cohesion (Soane,

* Corresponding author.
E-mail address: sergio.saia@crea.gov.it (S. Saia).

http://dx.doi.org/10.1016/j.geoderma.2016.10.019

1990). In addition, SOC is one of the most important CO_2 sequestration sources (Post et al., 1982). Rules and regulations were established by a number of countries to change the existing carbon balance, including the reduction of CO_2 emissions in the atmosphere from both soil and other sources. To achieve this, the CO_2 sequestration abilities of soils also need to be increased. The creation of a CO_2 accounting system, in which CO_2 levels are counted as C credits and debits, is presently a key agenda point in Europe and around the world. However, the creation of such accounting systems relies on correct knowledge of the C content of soils at the regional (Martin et al., 2011) or farm level (de Gruijter et al., 2016). In turn, such knowledge is fundamental to monitoring changes in the C stocks and a better understanding of the global C cycle (Martin et al., 2014). Different global or regional estimates were produced (Batjes, 2009, 1996; Hiederer and Köchy, 2012; Nachtergaele et al., 2008; Viscarra Rossel et al., 2016) based on various modelling and data-mining algorithms (see Minasny et al., 2013 and reference therein). Based on these efforts, various global estimates of SOC stock in the soil were provided, including the International Soil Reference and Information Centre (ISRIC) Soil Grids (Batjes, 2016, 2009; Hengl et al., 2014) and the Global Soil Organic Carbon Estimates (GSOC) (Hiederer and Köchy, 2012) by the Joint Research Centre (JRC) of the European Commission.

De Brogniez et al. (2015) created a topsoil organic carbon map using Generalized Additive Models (GAM) for the entire European Union (EU). In this study, topography and land use were recognized as key indicators to assess SOC stock and its distribution. In addition, Novara et al. (2013, 2014) also highlighted that some Mediterranean soils see increases in SOC levels when they are no longer cultivated. This suggests that soil cultivation can play a major role in affecting SOC under Mediterranean conditions. However, SOC dynamics also depend on other factors, including climate, soil type and texture, soil moisture, temperature regimes, lithology, morphology, land use history and management (Fantappiè et al., 2011b, 2010; Pisante et al., 2015). The knowledge of soil quality is a priority to support agricultural productivity and environmental quality. However, field sampling and laboratory analyses of SOC are costly and time consuming. Remote sensing (RS) data can contribute to modelling SOC information on a large scale (Gomez et al., 2008). In addition, RS predictors can also reduce uncertainties in SOC mapping through geographical soil unit classification (Köchy et al., 2015). However, in complex terrain, the high number of ecological determinants of the topsoil organic carbon can reduce the outcomes of the prediction (Yao et al., 2013), especially if samples from some areas are lacking. Nonetheless, the use of a high number of topographic and other ecological indices as predictors can increase the ability of the models to explain large parts of the amount of plant residues that were returned to the soil as well as SOC variation (Ferrara et al., 2009; Grimm et al., 2008).

Strategies for modelling SOC for large areas, e.g., at a regional scale, often rely on data mining approaches. However, these methodologies require a correction for spatial heterogeneity, outliers or correct sampling design to achieve a highly precise estimation of SOC stocks (Brus, 2015; Friedman et al., 2000; Schapire and Freund, 2012; Viscarra Rossel et al., 2016). The Stochastic Gradient Treeboost (SGT; Friedman, 2002), which is also referred to as Boosted Regression Trees (BRT; Elith et al., 2008), is an improvement of the Classification And Regression Trees (CART; Breiman et al., 1984). The SGT aims at identifying group membership to classes (the SOC stock value at a given cell or pixel) by sequentially partitioning the predictors' hyperspace into random trees (Lombardo et al., 2015). In particular, the SGT binary-splits the observations in homogeneous groups of the target variable as a function of combined explanatory variables and afterward combines several additive regression models in a forward stepwise procedure (Elith et al., 2008). The SGT was previously applied to an SOC model at a regional scale in organic soils (Bou Kheir et al., 2010) and in temperate environments (Martin et al., 2014). However, little information is available concerning SOC modelling in Mediterranean areas.

The Italian peninsula is characterized by a complex terrain with a high incidence of hilly and mountainous areas. As mentioned above, this can limit the accuracy of interpolations if the availability of data in the region is low. Conversely, the use of machine learning approaches allows for the recognition of causative relationships between SOC, topographic attributes and RS indices (e.g., McBratney et al., 2003). Lugato et al. (2014) provided an SOC stock estimation on a European scale using modelling techniques and validated them using the European Environment Information and Observation Network for Soil (EIONET-SOIL data by Panagos et al., 2013a, b). However, no specific regional modelling examples were provided for semi-arid Mediterranean regions.

The aim of the present work was to estimate the SOC stock through the SGT using a set of topographical and environmental covariates. Sicily was chosen for the model application since more than half of its surface is extensively cultivated and extensively sampled. In addition, across the island, there is a strong heterogeneity of agro-ecosystems in terms of soil type, texture, land use and microclimates. In the present study, we also compared the SGT results to those obtained from the GSOC estimate (Hiederer and Köchy, 2012; Panagos et al., 2012) and to those obtained from the International Soil Reference and Information Centre (ISRIC) Soil Grids (Batjes, 2016, 2009; Hengl et al., 2014). In particular, the SGT implemented in the present study was constructed using a 3-arcsec spatial resolution, whereas GSOC and ISRIC are freely available for scientific purposes as a spatial layer with a 30-arcsec spatial resolution and offer a benchmark regarding the overall SOC stock pattern in the agroecosystems.

2. Material and methods

2.1. Study area

Sicily is an Italian island in the middle of the Mediterranean Sea and has an area of 25,286 km^2 (36.64° to 38.30° N; 12.42° to 15.66° E), excluding its 37 ancillary islands. According to the CORINE land cover 2000 (CLC2000; Bossard et al., 2000), 64.1% of its territory is cropped. The remaining 35.9% encompasses non-agricultural ecosystems, including urban areas, Mediterranean maquis, dunes, coastal systems, forests, and industrial complexes. Sicily has several sub-climatic zones, all of which are included in the temperate Mediterranean belt, with mean annual temperatures usually higher than 15.8 °C, where summer is the driest period of the year. According to the climate classification of the Italian territory (Costantini et al., 2013), most of Sicily has a Mediterranean to subtropical climate that is partly semi-arid and is characterized by low rainfall, high air temperatures and high evapo-transpiration. Mediterranean subcontinental to continental climates, which are partly semi-arid to arid, typify the hinterland of the island. The mountain areas (Madonie, Sicani, Nebrodi and Peloritani ridges) are barely cultivated and are characterized by Mediterranean sub-oceanic to Mediterranean subtropical climates that are influenced by mountains. The continentality index, which is determined by the difference between the mean air temperature of summer and winter, is similar in all climatic regions. According to the World Reference Base for soils (IUSS Working Group WRB, 2014), the dominant soils in Sicily are Calcaric Regosols, Haplic Calcisols, Calcic Vertisols, Vitric or Silandic Andosols, Calcaric and/or Mollic Leptosols, Calcaric Phaeozems, and Fluvic Cambisols (Fantappiè et al., 2011a).

2.2. SOC stock analysis and database

The soil database of Sicily was the source of information for SOC (dag kg^{-1}, Fig. 1) and bulk density (BD; g cm^{-3}). It stores information for approximately 5658 georeferenced observations (soil profiles and minipits), of which 2891 are analysed for SOC following Walkley-Black (1934) and on fine-earth fraction (FEF); 1049 of them were also