Geoderma 335 (2019) 47–56

Contents lists available at ScienceDirect

Geoderma

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

Baseline map of soil organic matter in China and its associated uncertainty

Zongzheng Liang[a,b], Songchao Chen[c,d], Yuanyuan Yang[a], Ruiying Zhao[a], Zhou Shi[a,*], Raphael A. Viscarra Rossel[b]

[a] Institute of Agricultural Remote Sensing and Information Technology Application, Zhejiang University, 310058 Hangzhou, China
[b] CSIRO Land & Water, Bruce E. Butler Laboratory, PO Box 1700, Canberra, ACT 2601, Australia
[c] INRA Unité InfoSol, 45075 Orléans, France
[d] UMR SAS, INRA, Agrocampus Ouest, 35000 Rennes, France

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Soil organic matter
Spatial modeling
Cubist machine learning algorithm
Soil map
Uncertainty assessment

ABSTRACT

Accurate digital soil maps of soil organic matter (SOM) are needed to evaluate soil fertility, to estimate stocks, and for ecological and environment modeling. We used 5982 soil profiles collected during the second national soil survey of China, along with 19 environment predictors, to derive a spatial model of SOM concentration in the topsoil (0–20 cm layer). The environmental predictors relate to the soil forming factors, climate, vegetation, relief and parent material. We developed the model using the Cubist machine-learning algorithm combined with a non-parametric bootstrap to derive estimates of model uncertainty. We optimized the Cubist model using a 10-fold cross-validation and the best model used 17 rules. The correlation coefficient between the observed and predicted values was 0.65, and the root mean squared error was 0.28 g/kg. We then applied the model over China and mapped the SOM distribution at a resolution of 90 × 90 m. Our predictions show that there is more SOM in the eastern Tibetan Plateau, northern Heilongjiang province, northeast Mongolia, and a small area of Tianshan Mountain in Xinjiang. There is less SOM in the Loess Plateau and most of the desert areas in northwest China. The average topsoil SOM content is 24.82 g/kg. The study provides a map that can be used for decision-making and contribute towards a baseline assessment for inventory and monitoring. The map could also aid the design of future soil surveys and help with the development of a SOM monitoring network in China.

1. Introduction

Soil organic matter (SOM) is an important component of soil that helps to determine crop yield and carbon sequestration (Manlay et al., 2007). It is a key property that affects soil quality and the assessment of soil resources. The amount of carbon stored in soil is three times that in the atmosphere (Post and Kwon, 2000), and thus, small losses of soil carbon to the atmosphere can have a significant impact on the overall emissions of greenhouse gases and the greenhouse effect (Raich and Potter, 1995).

Soil in China, like elsewhere, is subject to complex soil forming environments, with persistent soil erosion and degradation, and long-term intensive farming. As a consequence, the spatial distribution of soil properties is very heterogeneous and existing soil property maps have considerable uncertainty. There is a growing demand for fine-resolution soil property maps for applications in environmental modeling and monitoring. Traditional polygon-based soil maps are less useful for these purposes because they do not adequately characterize the spatial variation of continuous soil properties. For instance, there is a need for precise spatially explicit estimates of SOM at the national scale for providing baselines for monitoring and to inform national greenhouse gas inventories (Viscarra Rossel et al., 2014).

Dokuchaev firstly developed a scientific classification of soils, methods for soil mapping and established the foundation for the study of both soil genesis and soil geography (Buol et al., 2011). Later Jenny proposed the well-known State Factor Equation of soil, where soil is described as a function of CLimate, Organisms, Relif, Parent material and Time, referred to as *CLORPT* (Jenny, 1941). McBratney et al. (2003) and Scull et al. (2003) reviewed methods for soil mapping which they defined as a spatial soil information system using field and laboratory observational methods, coupled with spatial and non-spatial soil inference systems. Lagacherie and McBratney (2007) formalized the "*SCORPAN*" framework of McBratney et al. (2003) and in a collection of manuscripts described "digital soil mapping". Since then, much of the work on soil mapping with linear regression, geostatistical methods, and data mining methods have fallen under the "digital soil mapping" umbrella (Adhikari et al., 2014; Chen et al., 2018; Grunwald, 2009; McBratney et al., 2003; Sun et al., 2012; Viscarra Rossel and

* Corresponding author at: Environmental & Resource Sciences College, Zhejiang University, Hangzhou 310058, China.
E-mail address: shizhou@zju.edu.cn (Z. Shi).

https://doi.org/10.1016/j.geoderma.2018.08.011
Received 18 October 2017; Received in revised form 4 August 2018; Accepted 7 August 2018
0016-7061/

Chen, 2011; Zhou et al., 2016).

In the last decade, much attention has been focused on soil carbon storage at the national scale. In China, most researchers have used classification statistics and interpolation methods to obtain average soil organic carbon content, soil depth, and other information (Pan, 1999; Wang et al., 2000; Wu et al., 2003; Xie, 2004; Yu et al., 2005). However, due to the inaccuracies of the input point data, the studies have produced diverse results. Other studies have used statistical models to map the spatial distribution of soil organic carbon in China, such as multiple regression combined with high accuracy surface modeling (HASM) and neural networks (Li et al., 2010; Q.Q. Li et al., 2013), and land surface modeling (Shangguan et al., 2013). But the spatial resolution of these studies is relatively coarse ($> 1 \times 1$ km). Large area, country, continental and global scale mapping at a fine resolution is now a major research emphasis that will allow for a better understanding of the soil resource and our environment (Arrouays et al., 2014).

Our aim here was to use a machine-learning model with data from the second national soil survey of China and covariates that represent the environmental factors, to map the spatial distribution of topsoil (0–20 cm) organic matter in China and its uncertainty at 90×90 m spatial resolution.

2. Materials and methods

2.1. Collection and processing of soil profile data

The study used a dataset of 5982 soil profiles derived from the second national soil survey of China (SNSSC), which was undertaken in the 1980s and is mainly recorded in the Soil Series of China (National Soil Survey Office, 1993, 1994a, 1994b, 1995a, 1995b, 1996) and the Soil Series of Provinces (National Soil Survey Office, 1998). The carbon determination was carried out by rational wet combustion (Pan et al., 2004). The soil data covers most geographical regions of China and is the most detailed soil survey available at the national scale.

The soil profiles were sampled and analyzed by genetic horizons, and thus the depth intervals for each soil profile are inconsistent. As variation in SOM down a profile is usually continuous, we used equal-area splines (Ponce-Hernandez et al., 1986; Bishop et al., 1999; Malone et al., 2009) to harmonize the SOM content of the topsoil, which we defined as the 0–20 cm depth layer. To fit the splines to the SOM values in the profiles we tested different tuning parameter values, λ: 10, 1, 0.1, 0.01, 0.001, 0.0001, and 0.00001. We found that $\lambda = 0.01$ produced the best fits with the smallest root mean square error (RMSE). The splines were fitted to a maximum depth of 1 m, and we aggregated the spline predictions of SOM over the 0–20 cm to represent topsoil. Fig. 1 displays SOM depth function curves for three random soil profiles under different land uses.

The soil profiles from the SNSSC lack precise geographical registration and have no data on latitude and longitude. However, they do have detailed sampling location information that can be accurate to the villages, fields. To verify the spatial location accuracy of the digitized soil profiles, we compared elevation, mean annual precipitation, and mean annual temperature (below) recorded in each soil profile with data extracted from a high-resolution digital elevation model (DEM) and digital climate map using linear regression analysis. Coefficients of determination (R^2) for both elevation and temperature were larger than 0.90, and it was 0.80 for precipitation (Fig. 2). These results confirmed that the spatial accuracy of the digitized soil profiles was adequate for our study. The spatial distribution of the profiles is shown in Fig. 3.

Statistical analysis of the dataset showed that the distribution of SOM was skewed, with a mean of 24.82 g/kg, maximum of 560.1 g/kg for a peat soil type, found in Ganzi in Sichuan Province, minimum of 0.6 g/kg for a clay soil type found in Gaolan in Gansu Province. SOM of topsoil showed strong spatial variation, with a coefficient of variation (CV) of 140%, which is mainly attributed to diverse soil types, land uses, ecosystems, etc., at the national scale. To ensure the data is normally distributed, the SOM data was log-transformed prior to modeling with logs to the base 10.

2.2. Environmental covariates

Following soil formation theory, a number of environmental covariates were chosen for our modeling. They include covariates that represent terrain, climate, biota, geology, and human activities (Table 1). Terrain information was derived from the 90-m shuttle radar topographic mission (STRM) DEM. All terrain attributes, including elevation, slope, aspect, curvature, slope length (LS), slope steepness, mass balance index (MBI), terrain ruggedness index (TRI), topographic wetness index (TWI), and multiresolution index of valley bottom flatness (MrVBF) were derived from the DEM with the System for Automated Geoscientific Analyses (SAGA) geographic information system (GIS) (http://www.saga-gis.org).

Data on daytime land surface temperature (LST_D), nighttime land surface temperature (LST_N), normalized difference vegetation index (NDVI), evapotranspiration (ET), and net primary productivity (NPP) were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Justice et al., 1998). The resolution of different data is shown in Table 1.

Monthly precipitation products were obtained from the Tropical Rainfall Measuring Mission (TRMM), which measures tropical and subtropical precipitation (http://trmm.gsfc.nasa.gov/data_dir/data.html).These data were at a coarse resolution (0.25° resolution) (Huffman et al., 2007) and so we downscaled it using a geographically weighted regression (Ma et al., 2017) to derive 90-m resolution mean annual precipitation that we could use here.

Daily air temperature data for the period 1951–2014 from 754 base meteorological observation stations distributed throughout mainland China were used to calculate annual mean temperature (http://cdc.nmic.cn/home.do). Mean annual temperature maps were produced at 90-m resolution using a regression-kriging approach with elevation, latitude, and longitude as the auxiliary variables.

Annual mean solar radiation data (1950–1980) were derived from the National Earth System Science Data Sharing Infrastructure (http://www.geodata.cn). Land use and land cover data were obtained from the Resource and Environment Data Center of the Chinese Academy of Sciences (http://www.resdc.cn/).

These environmental covariates (LST_D, LST_N, ET, NDVI, NPP and Solar Radiation) with resolution coarser than 90 m were resampled to 90-m using bilinear method in ArcGIS 10.0.

2.3. Digital soil mapping

2.3.1. Modeling

The sites recorded covariates (precipitation, temperature, and elevation) as well as other environmental covariates at the sampling locations are easy to obtain by overlaying the sampling locations with the covariates maps and can be used as predictors from which to predict soil properties such as SOM. We took advantage of this by setting up a model in the form of a decision tree at the sites for which we had data and then using the model to predict SOM elsewhere. Decision trees have become one of the most commonly used data mining algorithms and are ideally suited for dealing with complex nonlinear relationships and missing values (Quinlan, 1992).

We used the algorithm that is implemented in the 'cubist' library (Kuhn et al., 2014) in the R software (R Core Team, 2013). Cubist uses conditional functions to build rules that partition the data into regions that are similarly defined by the characteristics of the predictor variables. If the condition is true, then an ordinary least-squares linear model predicts the response. If the condition is false, then the rule defines the next node in the tree. The sequence if, then, else is repeated. The result is that the regression equations, although general in form, are