Geoderma 266 (2016) 98–110

Contents lists available at ScienceDirect

Geoderma

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

Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran

R. Taghizadeh-Mehrjardi [a,*], K. Nabiollahi [b], R. Kerry [c]

[a] Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran
[b] Faculty of Agriculture, University of Kurdistan, Sanandaj, Iran
[c] Department of Geography, Brigham Young University, Provo, UT, USA

ARTICLE INFO

Article history:
Received 14 November 2014
Received in revised form 17 November 2015
Accepted 7 December 2015
Available online xxxx

Keywords:
Artificial neural network
Support vector regression
k-nearest neighbor
Random forest
Regression tree model
Genetic programming

ABSTRACT

This study aimed to map SOC lateral, and vertical variations down to 1 m depth in a semi-arid region in Kurdistan Province, Iran. Six data mining techniques namely; artificial neural networks, support vector regression, *k*-nearest neighbor, random forests, regression tree models, and genetic programming were combined with equal-area smoothing splines to develop, evaluate and compare their effectiveness in achieving this aim. Using the conditioned Latin hypercube sampling method, 188 soil profiles in the study area were sampled and soil organic carbon content (SOC) measured. Eighteen ancillary data variables derived from a digital elevation model and Landsat 8 images were used to represent predictive soil forming factors in this study area. Findings showed that normalized difference vegetation index and wetness index were the most useful ancillary data for SOC mapping in the upper (0–15 cm) and bottom (60–100 cm) of soil profiles, respectively. According to *5-fold* cross-validation, artificial neural networks (ANN) showed the highest performance for prediction of SOC in the four standard depths compared to all other data mining techniques. ANNs resulted in the lowest root mean square error and highest Lin's concordance coefficient which ranged from 0.07 to 0.20 log (kg/m^3) and 0.68 to 0.41, respectively, with the first value in each range being for the top of the profile and second for the bottom. Furthermore, ANNs increased performance of spatial prediction compared to the other data mining algorithms by up to 36, 23, 21 and 13% for each soil depth, respectively, starting from the top of the profile. Overall, results showed that prediction of subsurface SOC variation needs improvement and the challenge remains to find appropriate covariates that can explain it.

1. Introduction

Baneh area located in Kurdistan province, Iran has suffered from deforestation in recent decades due to population growth. Forest areas were cleared to create land for cultivation to feed the growing population, which caused land degradation. SOC maps are useful for several reasons, namely; increasing crop production, land degradation management and designing an effective C sequestration program for the area. However, there are no high-resolution maps which describe SOC in the topsoil and subsoil for Iran. Conventional soil mapping techniques have been criticized in the scientific literature for being subjective and qualitative in character, where soil maps are developed based on a mental model developed by the soil surveyors (Taghizadeh-Mehrjardi et al., 2015). Such qualitative maps, while helpful, can lead to ill-informed management decisions. To overcome these problems, the application of digital soil mapping (DSM) techniques could be an efficient alternative approach. In DSM, soil properties are mapped digitally based on their relationship with cheaper-to-measure ancillary data (McBratney et al., 2003). Previous studies indicated that digital elevation models (DEM) and remotely sensed data are the most common ancillary data for SOC prediction (Malone et al., 2009; Mulder et al., 2011; Minasny et al., 2013; Dai et al., 2014; Were et al., 2015).

Numerous prediction methods have been developed and introduced to correlate ancillary variables and soil organic carbon (SOC) through the DSM framework proposed by McBratney et al. (2003). Minasny et al. (2013) give a comprehensive review of SOC modeling. Most commonly, multiple and linear regression have been used for relating SOC to ancillary variables (Hengl et al., 2015). The later technique is simple in application and easy in interpretation. Fewer studies used generalized linear models (Karunaratne et al., 2014), regression tree models (Martin et al., 2011), random forest (Were et al., 2015; Hengl et al., 2015), artificial neural networks (Malone et al., 2009; Dai et al., 2014), support vector regression (Were et al., 2015), *k*-nearest neighbor (Mansuy et al., 2014) or genetic programming to construct the relationships between SOC content and other ancillary variables. However, such modeling techniques have the potential for detecting non-linear relationships and might therefore prove more powerful for digital SOC mapping. Unfortunately, a major drawback of these machine learning approaches is that they only show SOC spatial variability mapped at

* Corresponding author.
E-mail address: rtaghizadeh@ardakan.ac.ir (R. Taghizadeh-Mehrjardi).

http://dx.doi.org/10.1016/j.geoderma.2015.12.003

specified depths or a combination of depth intervals while SOC generally varies continuously within a typical soil profile. Soil carbon has been observed to decline rapidly with depth (Minasny et al., 2013). Therefore, this variation can be modeled using continuous soil depth functions (Malone et al., 2009) to create a 3D map, describing vertical and lateral variation of SOC. Many attempts have been made to derive some functions of soil variation with depth (Mishra et al., 2009; Kempen et al., 2011). However, Bishop et al. (1999) suggested that equal-area quadratic splines are more flexible and practicable depth functions compared to other methods.

With regard to the potential of soil depth functions (Malone et al., 2009) and the capabilities of digital soil mapping (Mulder et al., 2011), the only way to predict lateral and vertical variation of soil properties seems to be a combination of both methods. So this paper aims to predict spatial SOC variation using different digital soil mapping techniques (i.e. artificial neural network, support vector regression, *k*-nearest neighbor, random forest, regression tree model, and genetic programming) together with a depth function (the equal-area smoothing spline) in a semi-arid area of Iran.

2. Material and methods

2.1. Study area

The study area is located in Kurdistan Province, about 12 km northwest of Baneh, Iran (Fig. 1). It lies between the latitudes of 36.01 and 36.08 ° North and the longitudes of 45.66 and 45.83° East and covers 3000 ha. The climate is semiarid with distinct differences between dry (July–September) and wet (Oct–May) seasons. Average annual rainfall and temperature are 700 mm and 13.8 °C, respectively. Soil moisture and temperature regimes are Xeric and Mesic, respectively. The geomorphologic units consist of piedmont, plateau and hills with steep slopes. Elevation varies from 1400 to 1600 m.s.l. Soil textures are mainly in the clay to sandy loam classes. Main land use types consist of forest, cropland (croplands occupy approximately 70% of the total area), and rangeland. Forests are more degraded in low areas than in high areas due to ease of accessibility. The main agricultural crops are wheat and barley. Agricultural practices on the steeper slopes cause soil erosion as no conservation practice is applied by the farmers except that they leave the crop residues after harvest.

The flowchart, shown in Fig. 2, illustrates the procedures employed in this study.

2.2. Ancillary data

2.2.1. Acquiring ancillary data

The digital soil mapping approach follows the *scorpan* spatial prediction function (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n) + e \tag{1}$$

where S, the soil organic carbon to predict, is a function of soil (s), climate (c), organisms (o), relief (r), parent materials (p), age (a), and spatial position (n); and where e is the error. Because of uniformity of climate and parent materials in our study area, SOC is mainly influenced by variations in vegetation and relief. These factors could be characterized by Landsat spectral data and a digital elevation model, respectively.

A Landsat 8 ETM$^+$ image acquired on March 28, 2013 was analyzed using ERDAS Imagine image processing software (Leica Geosystems Geospatial Imaging, 2008). These data were projected in the same geographic space using the Universal Transverse Mercator system, clipped to the extent of the study area, and converted into ERDAS Imagine file format (.img). Then, some band ratios found to be the best representation of the vegetation variable (Mulder et al., 2011) were calculated (Normalized difference vegetation index; ratio vegetation index; soil adjusted vegetation index). Other band ratios were also computed to represent parent material and soil factors at the study area (Clay Index, gypsum index and salinity ratio). Table 1 shows which Landsat bands contribute to each index or band ratio. In addition to band ratios a principal component analysis (PCA) of individual bands was calculated to summarize the variation in spectral reflectance in the dataset as a whole. Summary statistics were computed prior to PCA analysis to check for normality of the data and the use of the covariance and correlation matrix was compared. The computed eigenvalues were used to determine the relationship of the original bands to one another in multivariate space and identify those bands that were not important and could be ignored in future analysis. The Eigenvectors gave a multivariate summary of each observation location in terms of principal components and showed how they relate to one another in terms of groupings in multivariate space and were used to derive new brightness values. PCA is widely used to discriminate SOC in semi-arid regions (Taghizadeh-Mehrjardi et al., 2014). For this paper, the first three principal components were included as environmental covariates as suggested by Mulder et al., 2011 and because they accounted for the majority of variation (98%) in the dataset.

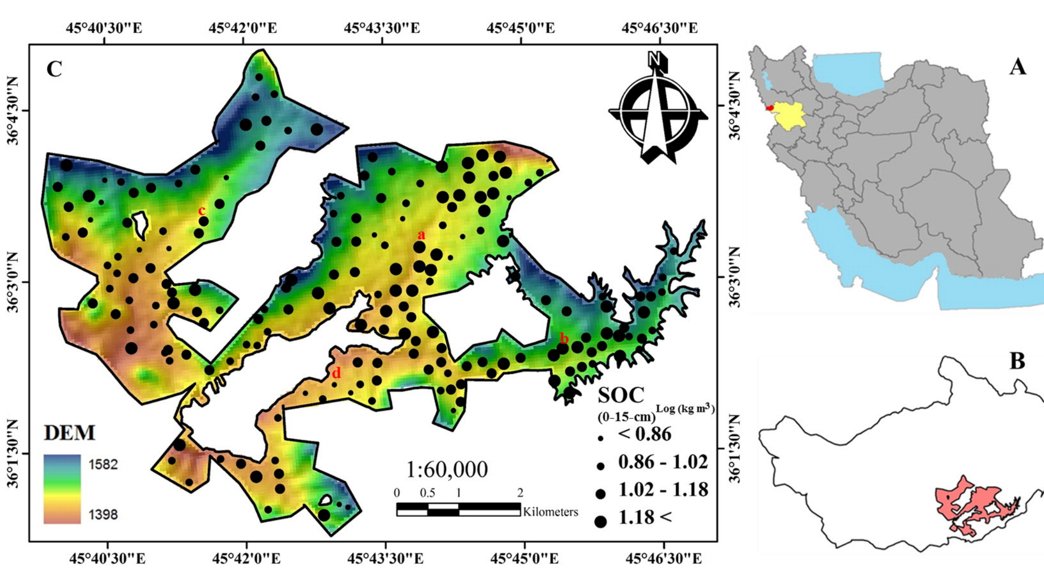

Fig. 1. Location of Kurdistan province (A), the Baneh region, and study area (B) in western Iran and the spatial distribution of soil samples (C) (a, b, c and d are the four validation data points selected randomly). SOC: soil organic carbon content log (kg/m^3) at 0–15-cm depth; DEM: digital elevation model (m above sea level).