Geomorphology 285 (2017) 186–204

Contents lists available at ScienceDirect

Geomorphology

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

Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran

Mojtaba Zeraatpisheh [a,c], Shamsollah Ayoubi [a,*], Azam Jafari [b], Peter Finke [c]

[a] Department of Soil Science, College of Agriculture, Isfahan University of Technology, 84156-83111 Isfahan, Iran
[b] Department of Soil Science, College of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran
[c] Department of Soil Management, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 28 August 2016
Received in revised form 9 February 2017
Accepted 10 February 2017
Available online 24 February 2017

Keywords:
Digital soil mapping
Conventional soil mapping
Map purity
Modelling efficiency
Multinomial logistic regression
Random forest

ABSTRACT

The efficiency of different digital and conventional soil mapping approaches to produce categorical maps of soil types is determined by cost, sample size, accuracy and the selected taxonomic level. The efficiency of digital and conventional soil mapping approaches was examined in the semi-arid region of Borujen, central Iran. This research aimed to (i) compare two digital soil mapping approaches including Multinomial logistic regression and random forest, with the conventional soil mapping approach at four soil taxonomic levels (order, suborder, great group and subgroup levels), (ii) validate the predicted soil maps by the same validation data set to determine the best method for producing the soil maps, and (iii) select the best soil taxonomic level by different approaches at three sample sizes (100, 80, and 60 point observations), in two scenarios with and without a geomorphology map as a spatial covariate. In most predicted maps, using both digital soil mapping approaches, the best results were obtained using the combination of terrain attributes and the geomorphology map, although differences between the scenarios with and without the geomorphology map were not significant. Employing the geomorphology map increased map purity and the Kappa index, and led to a decrease in the 'noisiness' of soil maps. Multinomial logistic regression had better performance at higher taxonomic levels (order and suborder levels); however, random forest showed better performance at lower taxonomic levels (great group and subgroup levels). Multinomial logistic regression was less sensitive than random forest to a decrease in the number of training observations. The conventional soil mapping method produced a map with larger minimum polygon size because of traditional cartographic criteria used to make the geological map 1:100,000 (on which the conventional soil mapping map was largely based). Likewise, conventional soil mapping map had also a larger average polygon size that resulted in a lower level of detail. Multinomial logistic regression at the order level (map purity of 0.80), random forest at the suborder (map purity of 0.72) and great group level (map purity of 0.60), and conventional soil mapping at the subgroup level (map purity of 0.48) produced the most accurate maps in the study area. The multinomial logistic regression method was identified as the most effective approach based on a combined index of map purity, map information content, and map production cost. The combined index also showed that smaller sample size led to a preference for the order level, while a larger sample size led to a preference for the great group level.

1. Introduction

Soil information of good quality and high spatial resolution is essential for adequate support of land use management practices, precision agriculture, and ecosystem research. In spite of more than 50 year of soil survey history in the world, in Iran there are just few maps at scales appropriate for land use planning and agricultural practices. As an example, the conventional soil map of Iran (1:1,000,000) recently was prepared by the Soil and Water Research Institute of Iran (Mohammad, 2000; Banaei et al., 2005) based on landform delineations of the main physiographic regions, is not sufficiently informative (Hengl et al., 2007). Detailed maps supporting many applications, however, exist in some countries with soil maps at spatial resolutions of 100 m (The Netherlands; De Vries et al., 2003; Kempen et al., 2015), 10 m (one-third of Germany; Lösel, 2003), 100–400 m (Germany; McBratney et al., 2003) and 200–500 m (France; King et al., 1999). Therefore, it is necessary for Iranian soil scientists and decision makers to produce soil maps at finer scales that provide more detailed information.

Conventional methods of soil mapping are currently considered to be ineffective to produce detailed soil maps at a reasonable cost and time (Kempen et al., 2012). Digital soil mapping (DSM) is a powerful technique which is increasingly applied by soil scientists and

* Corresponding author.
E-mail address: ayoubi@cc.iut.ac.ir (S. Ayoubi).

http://dx.doi.org/10.1016/j.geomorph.2017.02.015
0169-555X/

environmentalists to map soil types and/or properties using ancillary data (McBratney et al., 2003; Lagacherie and McBratney, 2007). These ancillary data, termed environmental covariates, can be obtained from digital elevation models (DEM), satellite imagery (remote sensing data), maps of geology and geomorphology, and legacy soil maps (categorical maps) (Krasilnikov et al., 2011).

The basis of DSM is the application of pedometric techniques that predict the spatial distribution of soil types and soil properties (Wulf et al., 2015). Here, we focus on making maps of soil types because these have been mapped at incomplete coverage until now and the desire exists to finalize soil mapping in the most economically feasible way. Recently, several novel models have been developed to produce soil type maps from profile observations by utilizing auxiliary data (Nelson and Odeh, 2009; Heung et al., 2014; Brungard et al., 2015). Many such methods have been investigated for digital soil mapping of soil types, including Random forest (RF) (Pahlavan Rad et al., 2014; Brungard et al., 2015), Multinomial logistic regression (MLR) (Abdel-Kader, 2011; Jafari et al., 2012; Kempen et al., 2012; Brungard et al., 2015), Artificial Neural Networks (Jafari et al., 2013; Brungard et al., 2015), Support Vector Machine (Kovačević et al., 2010), Nero-Fuzzy approach (Viloria et al., 2016), and Genetic Algorithms (Nelson and Odeh, 2009).

DSM models are divided into simple, intermediate, and complex models (Brungard et al., 2015) based on their interpretability and the number of parameters required. In the present study, two DSM models including RF (a complex model), MLR (a simple model), and the conventional soil mapping method were compared for predicting soil types. RF and MLR compared favourably to other methods in earlier studies in Iran (Jafari et al., 2013; Pahlavan Rad et al., 2014).

RF can be regarded as an ensemble of classification and regression trees (CART) which are aggregated to provide the final prediction (Breiman et al., 1984; Breiman, 2001; Cutler et al., 2007). RF has several advantages over other statistical modelling approaches (Breiman, 2001; Liaw and Wiener, 2002). Its input and output variables can be both continuous and categorical (Grimm et al., 2008). Moreover, RF has the advantage of incorporating 'randomness' into its predictions through reiterative bootstrap sampling and randomized variable selection when generating each decision tree (Heung et al., 2014). The RF algorithm is considered a powerful modelling technique for predicting soil types because (i) it is quite robust to noise in predictors, (ii) it shows no over-fitting, (iii) it produces predictions with low bias and low variance, and (iv) since it is also fairly fast, it does not require the pre-selection of variables (Díaz-Uriarte and De Andres, 2006; Prasad et al., 2006; Wiesmeier et al., 2011). RF also identifies the most important covariates (Hua et al., 2005; Archer and Kimes, 2008).

Abdel-Kader (2011) reported that the MLR model is the most frequently used statistical model for spatial prediction of soil types and spatial modelling in land use and ecology studies (Rhemtulla et al., 2007; May et al., 2008; Suring et al., 2008). However, in recent years only some studies have used MLR for digital soil mapping (Abdel-Kader, 2011; Jafari et al., 2012; Kempen et al., 2012).

DSM and CSM approaches are similar in that they both make use of relationships between soil properties and more readily observable land surface properties (shape, position, and reflectance). Conventional soil maps are limited by the scale of the base map, their inability to represent continuous soil classes and spatial variation (Roecker et al., 2010). Production of maps using CSM techniques is also labour-intensive and expensive. DSM-based maps suffer less from these limitations, thus DSM is generally assumed to be more efficient than CSM (Kempen et al., 2012).

Although several papers have been published on the benefits of DSM compared to CSM in recent years, few examples exist that compare DSM techniques with CSM approaches for predicting soil types in the same area, especially in arid and semi-arid regions. The objective of this study, therefore, was to compare two different DSM techniques (MLR and RF) with a conventional soil survey for producing soil maps at different taxonomic levels in a semiarid region of Iran. This comparison assessed not just map accuracy, but also information content and production cost, with the purpose of selecting the most efficient method as a function of the taxonomic level of the maps. Because results may depend on sampling density, we evaluated the effect of three different sample sizes on our conclusions.

2. Materials and methods

2.1. Description of the study area

The study area is located between 51° 19′ 9″ to 51° 20′ 45″ E longitude and 31° 41′ 00″ to 32° 00′ 00″ N latitude, and area covering approximately 86,000 ha in the Borujen region, Chaharmahal-Va-Bakhtiari Province, Central Iran (Fig. 1). The mean annual precipitation is 255 mm, mean annual temperature is 10.7 °C, and mean elevation of the selected area is 2277 m a.s.l. The main land uses in this area include irrigated wheat cropping, dryland farming, and pasture. According to the US Soil Taxonomy (Soil Survey Staff, 2014), the study area has a Xeric soil moisture regime and a Mesic soil temperature regime. Major landscape units in the study area consist of mountains, hills, piedmonts, and lowlands.

2.2. Soil sampling scheme and profile description

The soil sampling scheme was carried out by applying the conditioned latin hypercube sampling (Minasny and McBratney, 2006) algorithm using Matlab software (MathWorks, 2009) with all covariates mentioned in Section 2.3 (Table 1). Location coordinates of 100 soil profiles were acquired by latin hypercube sampling and 25 legacy profiles were added to our dataset. Fig. 2 shows the distribution of the soil profiles described in the study area. All locations were excavated to a depth of 100–150 cm, described, sampled, analysed, and classified up to the subgroup level of the US Soil Taxonomy (Soil Survey Staff, 2014).

2.3. Environmental covariates

Environmental covariates were represented by categorical maps of geomorphology and geology (scale of 1:100,000), by quantitative maps representing topographic attributes, and by remote sensing data. Topography and parent material are the main soil forming factors in arid and semi-arid regions (Florinsky et al., 2002; Tajik et al., 2012; Mehnatkesh et al., 2013). Therefore, to obtain the topographic attributes, we downloaded a DEM with the cell size of 30 × 30 m derived from the Aster GDEM database (Ministry of Economy, Trade and Industry of Japan and the National Aeronautics and Space Administration, 2009). The terrain attributes obtained from the DEM included elevation, the topographic wetness index, the SAGA (System for Automated Geoscientific Analysis) wetness index, a multi-resolution of ridge top flatness index, a multi-resolution valley bottom flatness index (Gallant and Dowling, 2003), curvature, profile curvature, plan curvature, aspect, and slope (Table 1).

Remote sensing auxiliary variables included the normalized difference vegetation index (NDVI; Boettinger et al., 2008), the ratio vegetation index (Pearson and Miller, 1972), the perpendicular vegetation index (Richardson and Wiegand, 1977), the clay index (Boettinger et al., 2008), and the soil adjusted vegetation index (SAVI; Huete, 1988). These indices were derived from the Landsat Enhanced Thematic Mapper acquired in 2008 (U.S. Geology Survey, 2004). All extracted environmental covariates were used in the latin hypercube sampling scheme and soil type prediction (Table 1). The SAGA GIS was used to derive environmental covariates (Olaya, 2004).