Geoderma 232–234 (2014) 148–163

Contents lists available at ScienceDirect

Geoderma

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

Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran

Azam Jafari [a], Hossein Khademi [b,*], Peter A. Finke [c], Johan Van de Wauw [c], Shamsollah Ayoubi [b]

[a] Department of Soil Science, College of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran
[b] Department of Soil Science, College of Agriculture, Isfahan University of Technology, 84156-83111 4 Isfahan, Iran
[c] Department of Geology and Soil Science, Ghent University, B9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 10 January 2012
Received in revised form 11 March 2014
Accepted 28 April 2014
Available online 28 May 2014

Keywords:
Boosted regression tree
Soil great groups
Soil diagnostic horizons
Limited dataset

ABSTRACT

We evaluated the suitability and performance of boosted regression trees (BRT) as a potential technique for soil mapping using a limited point dataset in an arid region of Iran. The model was applied using two approaches: logistic-BRT as an indirect approach and multiclass-BRT as a direct approach to produce soil class maps. In indirect prediction, the occurrence of relevant diagnostic horizons was first mapped and various maps were then combined for a pixel-wise classification by combining the presence or absence of diagnostic horizons. To allow combination of the indicator maps for classification into great groups, a decision tree was defined for linking the occurrence of diagnostic horizons to a soil great group. In direct prediction, the dependent variable was the great group itself and; therefore, the probability distribution of the soil great groups was directly predicted. Auxiliary data used in this study to represent predictive soil forming factors were terrain attributes and Landsat data as quantitative variables and a geomorphology map as categorical variable. To assess the added value of the geomorphology map, the most laborious auxiliary variable to obtain, the BRT-performance of 2 data situations was compared: (i) using only the DEM and remote sensing covariates and (ii) additionally using the geomorphology map. Results showed that the geomorphology map contributed importantly to the prediction accuracy. When it was removed from the predictors, the prediction accuracy strongly decreased. The maximum reduction in predictive performance of both indirect and direct models occurred in the absence of geomorphology map, particularly for soil classes whose presence has direct relationship with geomorphic surfaces. Generally, poor predictions seemed to be mainly due to low sample size, high variability of ancillary predictors and the inability of the geomorphology map to differentiate the strata in detail. In most predictions, the purity was better for the direct model. The indirect method predicted a high probability of salic horizon in playa landform, gypsic horizon in gypsiferous hills and calcic horizon in alluvial fans. The indirect method could provide an insight into the causes of errors in prediction at the level of diagnostic horizons, which could help in selecting better covariates.

1. Introduction

Numerical information of soils based on new processing tools and different digital data is continuously increasing. In the context of a growing demand of high-resolution spatial soil information for environmental protection and management, fast and accurate prediction methods are needed. Recent publications indicate that digital soil mapping has been tested in a wide range of soils and the mapping scales throughout the world (Dobos et al., 2001; Grunwald, 2006; Hengl et al., 2007; McBratney et al., 2003). In digital soil mapping, soil observations are related to readily available ancillary spatial data. The relationship is quantified by different prediction methods using geographic information science, statistics and pedological approaches. Therefore, digital soil mapping has been facilitated by the advances in computing and information processing that occurred over the last 30 years. Recent soil landscape predictive algorithms such as neural networks, fuzzy logic or tree model develop mainly from machine learning fields (Fayyad et al., 1996; Grinand et al., 2008).

The classification and regression tree (CART) algorithm was applied for predictive soil mapping using data and maps from a reference area by Lagacherie (1992). Recent statistical advances were implemented on decision tree models, namely stochastic gradient boosting (Freidman, 1999). Boosted Regression Trees (BRT) is one of the several new techniques which aim to improve the performance of a single model by fitting many models and combining them for prediction. Boosting, or more

* Corresponding author. Fax: +98 311 3913471.
E-mail address: hkhademi@cc.iut.ac.ir (H. Khademi).

http://dx.doi.org/10.1016/j.geoderma.2014.04.029
0016-7061/

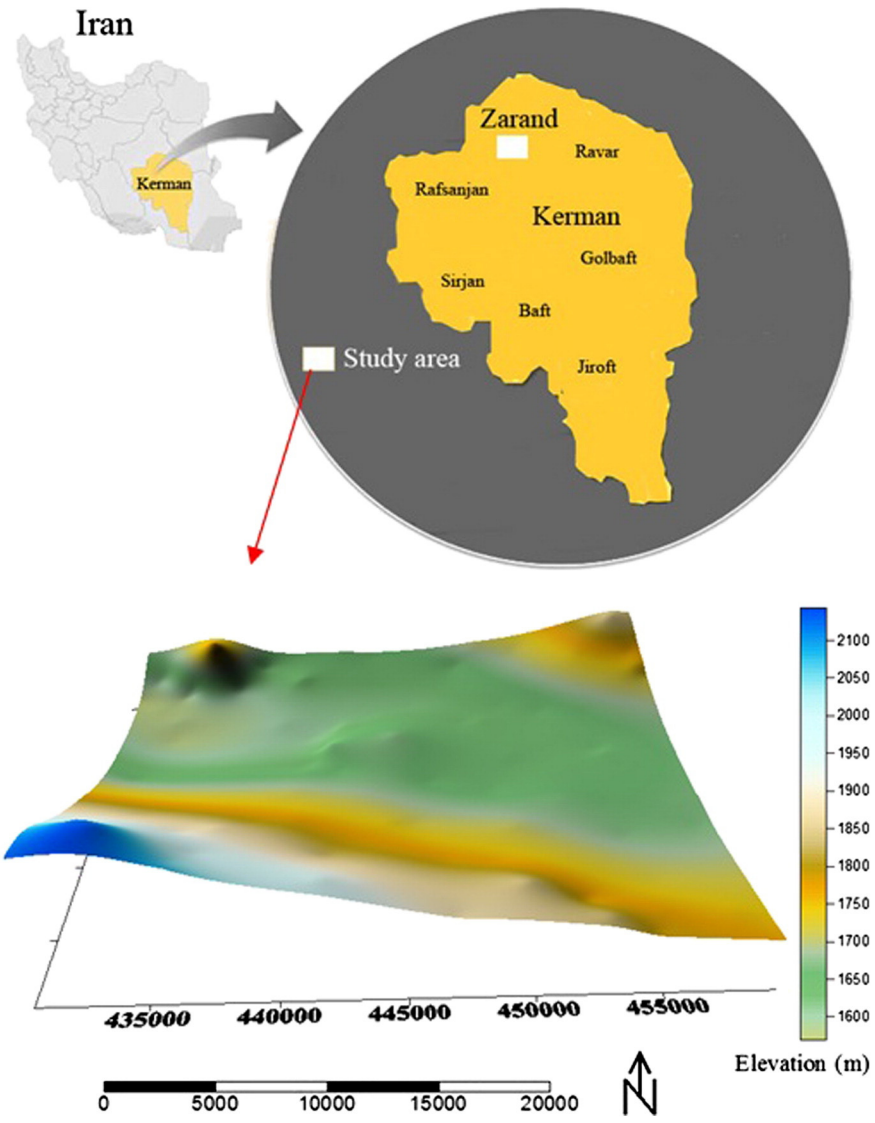

Fig. 1. The study area located near the city of Kerman, Iran.

precisely, stochastic gradient boosting, increases the predictive performance by reducing the over-learning, or overfitting, that commonly occurs with simple regression trees. Fitted BRT functions may be linear, curvilinear or non-linear, where the choice of error distribution includes normal, binomial and Poisson (De'ath, 2007; Elith et al., 2008). However, unlike the GLM (generalized linear model), in fitting a BRT model, there is no concern regarding outliers, the number or order of predictors, missing predictor values and variable selection. Given these advantages of the BRT method, there has been recent interest in tree-based models for soil mapping applications (Brown et al., 2006; Grinand et al., 2008). Recent studies have recognized advantages of using boosted trees as compared with simple trees which include the improvement of accuracy (Lawrence et al., 2004; Moran and Bui, 2002), little tuning needed and high robustness (Friedman and Meulman, 2003). Because it is more flexible, a boosted model tends to fit more realistic than a linear model and; therefore, inferences made based on the model may have more credibility.

Bauer and Kohavi (1999) made an extensive comparison of boosting to several other competitors on 14 datasets and found boosting as the best algorithm. Friedman et al. (2000) compared several boosting variants to the CART method and found that all the boosting variants outperform the CART algorithm on eight datasets. This technique showed significant improvements in the classification accuracy compared to unboosted classification and regression trees. Grinand et al. (2008) evaluated the ability of boosted tree model to provide accurate soil landscape prediction at an unsampled area. They found that the predictive capacity of models was quite low when extrapolated to an independent validation area.

In Iran, most soil survey studies have been carried out based on traditional methods and some areas have not yet been mapped at any scale. Recently, industry, agriculture, and mining sectors have increasingly focused on the application of geographic information systems and, as a result, digital soil data are now being collected more systematically. Foreseen intensive applications in agriculture and hydrological management demand high quality soil maps. Although predictive soil mapping studies are still in an introductory stage in Iran (Jafari et al., 2012), they provide a light start point, particularly in arid regions where traditional soil survey is difficult to perform.

To carry out digital soil mapping, Iranian soil scientists are faced with areas without any data and soil map. On the other hand, based on the principal of digital soil mapping techniques, there should be soil observations (soil profile data). Under such circumstances and also, due to the high cost of field work, digital soil mapping should be performed