Geoderma 232–234 (2014) 97–106

Contents lists available at ScienceDirect

Geoderma

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

Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran

Mohammad Reza Pahlavan Rad [a], Norair Toomanian [b,*], Farhad Khormali [a], Colby W. Brungard [c], Chooghi Bayram Komaki [d], Patrick Bogaert [e]

[a] Department of Soil Science, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran
[b] Department of Soil Science, Agricultural and Natural Resources Research Center of Isfahan, Isfahan, Iran
[c] Department of Plants, Soils and Climate, Utah State University, Logan, UT, USA
[d] Department of Arid Zone Management, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran
[e] Earth and Life Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

A R T I C L E I N F O

Article history:
Received 23 January 2014
Received in revised form 25 April 2014
Accepted 28 April 2014
Available online 23 May 2014

Keywords:
Digital soil mapping
Random forests
Soil series
Conditioned Latin hypercube sampling
Golestan

A B S T R A C T

Many Iranian soil surveys need to be updated. Conventional soil survey methods are expensive and time-consuming. Digital soil mapping (DSM) can be used for updating soil surveys. Many sampling and modeling techniques exist for DSM. In this paper we investigate the use of conditioned Latin hypercube sampling and random forest modeling for mapping Soil Taxonomy great group, subgroup and series levels for ~85,000 ha in Golestan Province, Iran. Overall error was 48.5, 51.5 and 56.6% for great group, subgroup and series levels, respectively. Estimated individual soil type error was between 8 and 100%. Soil types with larger sample sizes were predicted over a greater area at each taxonomic level. The soil adjusted vegetation index, the conventional soil series map and geomorphology were the most important covariates for each taxonomic level. Taxonomic classes with important covariates had low OOB error. The updated soil series map was 13.4% more accurate than the existing conventional soil series map.

1. Introduction

Many Iranian soil maps need to be updated, but conventional methods of soil mapping are expensive and time consuming. Digital soil mapping can reduce the costs associated with updating soil maps by identifying relationships between expensive soil observations and inexpensive auxiliary information (Kempen et al., 2012b; McBratney et al., 2003). Such auxiliary information, termed environmental covariates, can be obtained from digital elevation models, satellite imagery, geology maps and legacy soil maps.

To constrain costs an efficient method of identifying sampling locations is necessary. Conditioned Latin hypercube sampling (Minasny and McBratney, 2006) is an efficient sampling method because it captures the variability of multiple input environmental covariates (Minasny and McBratney, 2007) and can do so with a limited number of samples (Brungard and Boettinger, 2010).

Many techniques have been used to link soil observations and environmental covariates for updating soil maps including, decision trees (Adhikari et al., 2014), multinomial logistic regression (Kempen et al., 2009; Marchetti et al., 2011), possibilistic decision trees (Subburayalu et al., 2014) and the soil land inference model (SoLIM) (Yang et al., 2011). Subburayalu and Slater (2013) compared a classification tree and random forests for updating soil series maps using disaggregated soil map units and found that random forests produced higher accuracy. Häring et al. (2012) accurately disaggregated complex soil units with random forests in Germany.

Random forest (RF) can be viewed as an ensemble of classification and regression trees which are aggregated to provide a final prediction (Breiman et al., 1984). Each classification tree is constructed using bootstrap sampling and each node is split based on a random subset of the environmental covariates. Good reviews of RF can be found in Gislason et al. (2006) and Peters et al. (2007). Random forest is an attractive modeling technique for updating soil maps because it has high prediction performance, is robust to noise, has low bias and variance and is not sensitive to overfitting (Diaz-Uriarte and de Andres, 2006). Random forest also identifies the most important covariates (Archer and Kimes, 2008).

Although RF has been used for predicting soil classes in unmapped areas (Barthold et al., 2013; Stum et al., 2010) few studies have

* Corresponding author at: Department of Soil Science, Agricultural and Natural Resources Research Center of Isfahan, Iran. Tel.: +98 311 6255002; fax: +98 311 7757022.
E-mail address: notoomanian@yahoo.com (N. Toomanian).

http://dx.doi.org/10.1016/j.geoderma.2014.04.036
0016-7061/

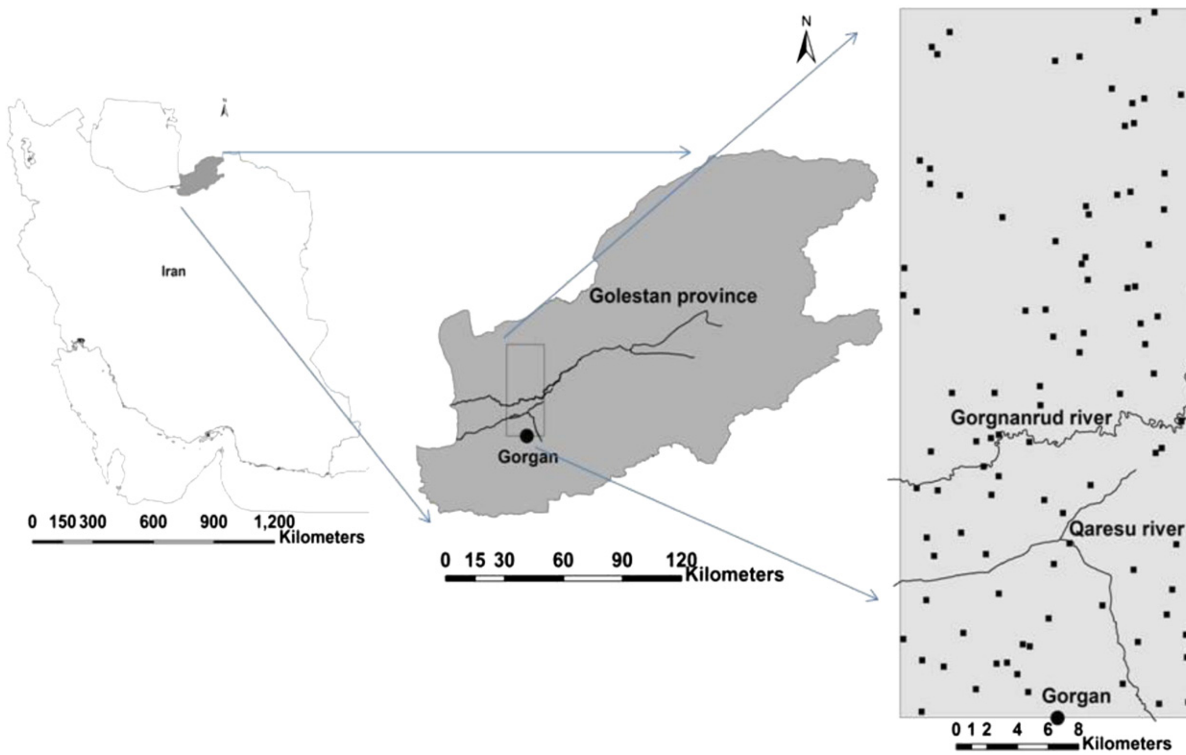

Fig. 1. The location of the study area and visited soil sampling sites.

investigated the use of random forests for updating soil maps, and none in combination with point observations obtained using cLHS. The objective of this study was to investigate the use of cLHS and RF for mapping soil types a three levels of taxonomic classification.

2. Materials and methods

2.1. Study area

The study area is located in Golestan province, a major agricultural region in northern Iran (Fig. 1). It covers approximately 85,000 ha and extends 45 km northward from Gorgan City. The elevation ranges from 158 m above m.s.l. to about 18 m below m.s.l. Annual precipitation ranges from about 600 mm in the south to under 200 mm in the north. Mean annual temperature is about 18 °C. The Gorganrud River divides the study area into north and south parts. Landcover varies from farmland in the south to rangelands in the north. Farmlands occupy approximately 85% of the total area, while the rest of the study area consists of rangelands. Parent materials are mainly loess and reworked loess. Geomorphologic units consist of piedmont, alluvial plains, river alluvial plains, depressions and loess hills (Table 1). Dominant soil orders in this area were expected to be Mollisols, Alfisols, Inceptisols, Aridisols and Entisols. Mollisols and Alfisols were expected to occur in the southern piedmont where the precipitation is high. Aridisols were expected to occur in central and northern areas that have low precipitation and near-surface ground water. Entisols were expected to exist around the Gorganrud River. Inceptisols were expected in both piedmont and loess landscapes.

2.2. Environmental covariates

A digital elevation model (DEM) with a spatial resolution of 27 × 27 m was downloaded from the Aster GDEM database (http://gdem.ersdac.jspacesystems.or.jp), reprojected to Universal Transverse

Table 1
Hierarchical description of geomorphic units in study area.

Landscape	Landform	Lithology	Geomorphic surface	Code
Piedmont	Alluvial fan	Alluvium of loess	Moderately flat, cultivated	Pi111
	Piedmont plain	Alluvium of loess	Flat, cultivated	Pi211
Alluvial plain	Alluvial flat	River alluvium	Flat, none saline, cultivated	Ap111
			Flat, low saline, cultivated	Ap112
			Flat, moderately saline, cultivated	Ap113
Loess	Undulating loess	Alluvial fine sediments	Low slope uncultivated	Lo111
			Low slope, cultivated	Lo112
			Low slope, rangeland	Lo113
	Depression	Alluvial fine sediments	Flat, cultivated	Lo211
			Flat, salty, cultivated	Lo212
			Flat, very salty, uncultivated	Lo213
Mud volcano	Mud volcano sediment	Silty, clayey	Undulating	Mu111
Lake	Lake sediments	Silty, clayey		La111