Geoderma 279 (2016) 141–148

Contents lists available at ScienceDirect

Geoderma

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

Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran

Mohammad Reza Pahlavan-Rad [a], Farhad Khormali [a,*], Norair Toomanian [b], Colby W. Brungard [c], Farshad Kiani [a], 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, Iran
[c] Department of Plants, Soils, and Climate, Utah State University, Logan, UT, United States
[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

ARTICLE INFO

Article history:
Received 1 January 2016
Received in revised form 6 May 2016
Accepted 22 May 2016
Available online 13 June 2016

Keywords:
Digital soil mapping
Legacy soil survey
Random forests
Multinomial logistic regression
Iran

ABSTRACT

Digital soil mapping (DSM) can be used for updating soil surveys. Legacy soil survey maps are often used as a covariate for updating soil surveys because such soil survey maps are logically assumed to contain significant information about the spatial distribution of soil classes. In the present study the usefulness of including conventional soil survey maps as a DSM covariate was investigated. Random forest and multinomial logistic regression models were built using two different covariate sets: covariate set 1 included the legacy soil survey, covariate set 2 excluded the soil survey. Soil Great Groups, Subgroups, and Series taxonomic classes were modeled using both models and covariate sets for an area of ~85,000 ha in Golestan Province, northern Iran. Overall model accuracy, the Kappa statistic, and individual covariate importances were used to assess the influence of including the legacy soil survey.

Including the conventional soil map as covariate generally increased model accuracy, but the improvement in model accuracy was surprisingly small at all taxonomic levels. This may be due to soil change or the mapping scale of the legacy soil survey. Random forests was found to be more accurate than multinomial logistic regression at all taxonomic levels. Multinomial logistic regression models at the soil Series level were less accurate than the legacy soil survey.

1. Introduction

Soil survey maps are needed to guide efficient agricultural and land management practices in Iran, but about 75% of the country lacks soil survey information. Given historical soil survey background, many years would be required before Iran is completely surveyed. Additionally, many existing legacy soil surveys require update, because soil change due to shifting land management practices, erosion, salinization and the change of groundwater levels should be considered over time. Because traditional soil survey methods are likely infeasible given current logistical and financial constraints, alternative solutions are required. Digital soil mapping (DSM) (Kempen et al., 2012; McBratney et al., 2003) is a solution. Digital soil mapping is the application of numerical models to link soil observations with quantitative proxies of the factors driving soil formation. The resulting outputs are predictive maps of soil distribution and associated uncertainty.

The selection of appropriate numerical models to link soil observations with quantitative proxies is required for accurate digital soil mapping. This is an active research topic and many techniques have been investigated (Adhikari et al., 2014; Häring et al., 2012; Kempen et al., 2009; Stum et al., 2010). In Iran, Jafari et al. (2012) used multinomial logistic regression to predict soil taxonomic Great Groups in southeast Iran. Pahlavan-Rad et al. (2014) applied random forests to model soil series in an area of loess in northern Iran. Taghizadeh-Mehrjardi et al. (2015) compared several models for predicting soil family classes in an area of northwest Iran including: multinomial logistic regression, artificial neural networks, support vector machines, K-nearest neighbors, random forests, and decision trees. These studies successfully used different models suggesting that the choice of numerical model is dataset specific (Grunwald, 2010). However; in a semi-arid region of the western USA, Brungard et al. (2015) found that complex models were more accurate than simple models.

In addition to an appropriate numerical model, quantitative proxies of soil forming factors, termed environmental covariates, are required for accurate digital soil mapping. In previously surveyed areas, existing legacy soil survey maps can be used as an environmental covariate

* Corresponding author at: Department of Soil Sciences, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran.
E-mail address: khormali@yahoo.com (F. Khormali).

http://dx.doi.org/10.1016/j.geoderma.2016.05.014
0016-7061/© 2016 Elsevier B.V. All rights reserved.

(Grunwald, 2009). There are two general approaches to including the existing legacy soil maps in DSM. The first approach samples directly from legacy soil maps to obtain soil class observations (Collard et al., 2014; Nauman and Thompson, 2014; Odgers et al., 2014). The second approach obtains field soil samples and then uses the legacy soil survey as a covariate (Kempen et al., 2009; Pahlavan-Rad et al., 2014), or derives categorical covariates from the original soil survey map (Kempen et al., 2015). In both approaches, the legacy soil survey is used because the soil maps are logically assumed to contain significant information regarding the spatial distribution of the soil classes.

To test this assumption we collected field soil samples and then predicted three soil taxonomic levels (Great Group, Subgroup, and Series) using a simple (multinomial logistic regression) and a complex (random forests) model. Each taxonomic level was predicted using two covariate sets: covariate set 1 included the legacy soil map, covariate set 2 excluded the legacy soil map. Our hypothesis was that including the legacy soil map as a covariate would lead to more accurate digital soil maps than excluding the legacy soil map.

2. Materials and methods

2.1. Study area

The study area was located in Golestan province in northern Iran, extending 45 km northward from Gorgan City and covers approximately 85,000 ha, (Fig. 1). The elevation ranges from 158 m above m.s.l. to about 18 m below m.s.l. Annual precipitation ranges from approximately 600 mm in the south to under 200 mm in the north. Mean annual temperature is about 17 °C. The Gorganrud River divides the study area into northern and southern parts. Landcover varies from farmland in the south to saline rangelands in the north. Farmlands occupy approximately 85% of the total area, while the rest of the study area is rangeland. Most of the farmlands are flat, and the main parent materials are mainly loess and reworked loess (Pahlavan-Rad et al., 2014).

2.2. Environmental covariates

2.2.1. Legacy soil maps

Two conventional soil surveys cover the area. A1:50,000 soil series map (Banaei, 1972) exists for the area south of the Gorganrud River, while a 1:100,000 soil series map covers the area north of the Gorganrud River (Farmanara, 1975). There were twelve soil series mapped in the southern part of the study area and four soil series in the northern part. Each conventional soil survey was digitized, merged into a single soil map, and rasterized to a spatial resolution of 30 m (Fig. 3).

2.2.2. Additional covariates

Terrain Analysis System 2.05 software (Lindsey, 2005) was used to derive aspect, maximum downward slope, down slope flowpath length, mean upslope slope, profile curvature, plan curvature, surface curvature, sediment transport capacity index, slope, and the topographical wetness index (Wilson and Gallant, 2000) from a 30 m^2 digital elevation model. The soil adjusted vegetation index (SAVI) (Huete, 1988) was derived from a March 2011 Landsat 5 TM image. Six main land use types were identified using supervised classification of the Landsat 5 TM image (Leica Geosystems Geosp). Land use types were: rangeland, farmland, built up and barren land, water body, and wetland (Fig. 2). Visual interpretation of aerial photography was used to delineate 13 geomorphic surfaces (Table 1) (Toomanian et al., 2006). Further details regarding covariate development can be found in Pahlavan-Rad et al. (2014).

2.3. Soil sampling

Conditioned Latin hypercube sampling (Minasny and McBratney, 2006) was used to identify 105 soil sampling locations using all covariates mentioned in Section 2.2, except for land use. Because of logistical constraints only 99 of these locations were actually sampled in the field (Fig. 3).

At each sampling location a soil profile was excavated to a depth of 100–150 cm. Each soil profile was analyzed, classified to the family level of Soil Taxonomy (Soil Survey Staff, 2010) and assigned to an existing soil series. Due to the relatively few observations of some series, thirteen series were combined with similar, but more common series, to reduce the total number of soil series to fifteen. Reducing the number of series was done to address problems which can affect modeling accuracy (Subburayalu et al., 2014; Kempen et al., 2009). Final soil taxonomic classes and the number of observations per class are shown in Table 2.

2.4. Experimental design

2.4.1. Modeling

Two covariate sets were created from the environmental covariates listed in Section 2.2. Covariate set one included the digitized legacy conventional soil survey (CSS +). Covariate set two excluded the digitized legacy conventional soil survey, but retained all other covariates

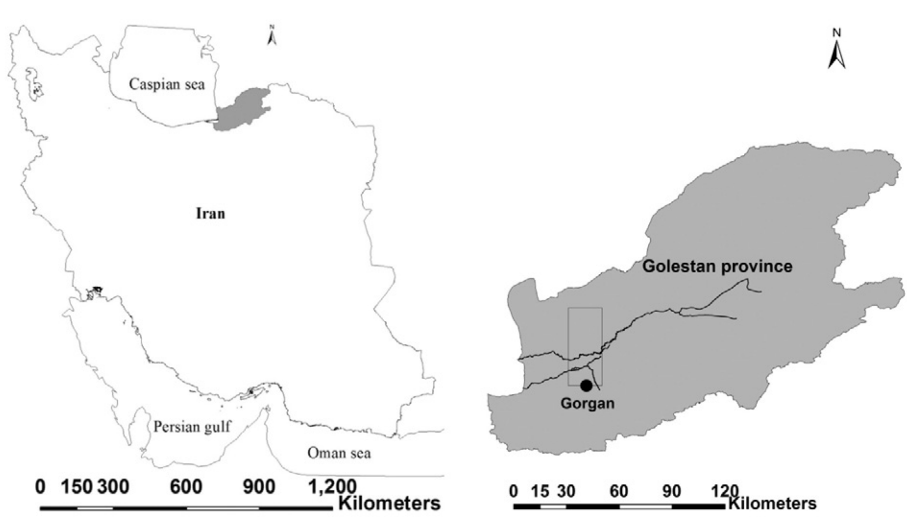

Fig. 1. The location of the study area.