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 d, 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 xxxx

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

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.

© 2016 Elsevier B.V. All rights reserved.

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.
(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 m² 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.
The influence of including the conventional soil survey as a DSM covariate was tested by modeling three taxonomic levels (Great Group, Subgroup, and Series) with both covariate sets and two model types. Tested models were multinomial logistic regression (a simple model) and random forests (a complex model). This resulted in four models for each taxonomic level. At each taxonomic level, model accuracy metrics were used to quantify the influence of including the conventional soil survey map as a covariate. A reduction in model accuracy metrics when using the covariate set that excluded the conventional soil survey (CSS−) indicated that the conventional soil survey map was an important covariate.

Model accuracy was assessed using leave-group-out cross validation as no independent validation dataset existed for this area. Leave-group-out cross validation splits the soil observations into a training and a test set. Models were constructed using the training set and validated with the test set. We used 70% of the observations for the training set and 30% of the observations for the test set (Brungard et al., 2015). This process was repeated 100 times for each model.
Model accuracy was assessed with the overall accuracy (Brus et al., 2011) and the Kappa statistic (Congalton and Green, 1998). The overall accuracy is a measure of classification accuracy and is defined as the proportion of the predicted area which equals the true soil class. Higher overall accuracy indicates a more accurate soil map (Brus et al., 2011). The Kappa statistic accounts for chance agreement when dealing with imbalanced classes (Marchetti et al., 2011).

All analysis was performed using the caret package in R 3.2.0 (R Development Core Team, 2015) and RStudio 0.98.1103 (RStudio, 2015).

### 2.4.2. Covariate importance

The influence of the conventional soil survey as a covariate was also quantified using covariate importance metrics from the individual models which retained the conventional soil survey (CSS+). Covariate importance metrics provide insight into the relative importance of each covariate.

Random forests identifies important covariates by generating multiple classification trees using bootstrap sampling, randomly scrambling the covariates in each bootstrap sample, and reclassifying the bootstrap sample. The misclassification error between the bootstrap sample using the scrambled covariate is then compared to the misclassification error of the original covariate (Peters et al., 2007). Covariates that result in a larger misclassification error when scrambled are more important. Covariate importance from multinomial logistic regression were derived by summing the absolute coefficient values over all modeled factor classes (e.g., Aquisalids vs. Calcixerolls) for each taxonomic level (Kuhn et al., 2015; Kuhn, 2015, Gevrey et al., 2003). All Model specific covariate importance values were converted to relative importance by scaling model specific importance values from 0 to 100%. Covariates with higher relative importance values are more important.

### 3. Results and discussion

#### 3.1. Model accuracy

At the Great Group level, the accuracy and kappa values were highest when the conventional soil map was excluded (CSS−) (Figs. 4 & 5). At the Subgroup and Series levels, the most accurate models were obtained when including the conventional soil survey (CSS+) (Figs. 4 & 5). However; the improvement in accuracy when including the conventional soil survey map at these two taxonomic levels was slight. At the Subgroup level, including the conventional soil map (CSS+) only increased the overall accuracy by 1.2% and kappa by 1.4%, while at the Series level the improvement in accuracy and kappa were 1.9% and 2.4%, respectively.

We suggest two possible reasons for explaining the increase in model accuracy when excluding the conventional soil map at the Great Group level, and the surprisingly small improvement in model accuracy at the Subgroup and Series levels when including the legacy soil survey map. First, the legacy conventional soil maps were produced approximately 40 years ago and since this time the soils have changed due to decreasing ground water levels and shifting land management activities (Pahlavan-Rad et al., 2014). This change has likely lead to a decreased accuracy in the legacy soil map, which is currently only 30% accurate (Pahlavan-Rad et al., 2014) and it is likely that this relatively low accuracy legacy soil map no longer contains highly relevant information about the spatial distribution of soil classes. Secondly, it is impossible to discount that the less detailed mapping (1:100,000 scale) in the northern part of the study area was simply too coarse to capture important differences in soil types. More than 50% of the soil observations used to build the models were located in areas covered by this 1:100,000 scale map. It is possible that if all of the observations has been in the area covered by the more detailed soil map (1:50,000) then the inclusion of the legacy soil map as a covariate may have resulted in more drastic differences in model accuracy.

These results are interesting when compared with Kempen et al. (2015) who found the legacy soil map an important covariate source for updating a 1:50,000 legacy map of peat soils in areas suspected of soil change resulting from shifting groundwater and land use. However; Kempen et al. (2015) did not directly model soil taxonomic classes, but modeled key diagnostic soil properties, which were then used to build a soil class map. Additionally, they did not directly use the soil map as a covariate. Instead they derived classes of soil characteristics from the legacy soil map and used these as covariates in the predictive model. It may be that soil characteristics derived from legacy soil maps are a better choice for digital soil mapping than the map classes themselves.

#### 3.2. Covariate importance results

The conventional soil survey was the second most important covariate across all taxonomic levels for the random forests models (Fig. 6) and ranged between the 5th and 7th most important covariate for multinomial logistic regression models (Fig. 6). Although ranked relatively highly in both models (particularly random forests), large differences between the first most important covariate (usually SAVI) and the conventional soil survey existed. This large difference (e.g., compare importance values in Fig. 6C.) and the relatively modest changes in model accuracy when the conventional soil survey was excluded (Figs. 4 & 5), suggests that either the most important variable is dominating each model or that other covariates were able to capture the information provided by the soil survey.

The most important covariate for all models except the multinomial logistic regression model at the Great Group level, was the soil adjusted vegetation index (SAVI). The importance of SAVI may be related to the influence of water on both vegetation and the taxonomic classification...
of the soils in this area. It is expected that areas with surface water, or near-surface groundwater will have higher vegetation cover and thus higher SAVI values. Because approximately half of the taxonomic groups at each taxonomic level indicate the presence of water (aqui- or fluv-prefixes) it is possible that SAVI captures important patterns in soil taxonomic classes.

3.3. Comparison between model types

Although not directly related to our hypotheses, two patterns in model accuracy metrics between the most accurate models at the different taxonomic levels were interesting. Firstly, random forests consistently had much higher overall accuracy and kappa values than did
multinomial logistic regression, regardless of the taxonomic level (Figs. 4 & 5). Secondly, overall accuracy and kappa decreased as taxonomic level increased from Great Group to Series for multinomial logistic regression, while overall accuracy and kappa remained roughly constant for random forests. These patterns suggest that random forests are more appropriate for capturing complex (and likely non-linear relationships) for updating soil maps in this landscape than is multinomial logistic regression. This is further supported as the accuracy of multiple logistic regression model at the Series level was actually less than the legacy conventional soil survey (25.4% v. 30.0%). We caution the use of multinomial logistic regression for soil class mapping, although we acknowledge that Kempen et al. (2015) found multinomial logistic regression useful. Additionally, we note that at all taxonomic levels differences in model accuracy metrics were larger between model types than between covariate sets.

3.4. Spatial prediction

Spatial predictions of taxonomic classes using the model with the highest accuracy and Kappa are presented in Fig. 7. The majority of the area was predicted to be Haploxerepts and Aquisalids Great Groups (Fig. 7A). Typic Calcixerepts, Gypsic Aquisalids and Aquic Calcixerepts were dominant at the subgroup level (Fig. 7B). At the series level, the majority of the area was predicted to be the Aqtapeh and Aghghala series (Fig. 7C). Prediction area appears related to sampling frequency (Table 2) (Pahlavan-Rad et al., 2014; Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015).

4. Conclusions

Our hypothesis was generally correct; the inclusion of legacy soil maps as a digital soil mapping covariate did increase model accuracy, but we were surprised at the relatively minor differences in model accuracy magnitude between models which included and excluded the

<table>
<thead>
<tr>
<th>Order</th>
<th>n Suborder</th>
<th>n Great Group</th>
<th>n Subgroup</th>
<th>n Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aridisols</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>Aquisalids 22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typic Aquisalids 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aquic Calcixerepts 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gypsic Calcixerepts 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typic Calcixerepts 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fluventic Haploxerepts 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gypsic Haploxerepts 22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typic Haploxerepts 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Halpoxerepts 35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aquic Halloys 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typic Xeroxluvents 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aquic Calcixerepts 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typic Calcixerepts 3</td>
</tr>
</tbody>
</table>

Table 2
Soil taxonomic classes and number of observations per class.

Fig. 4. Overall accuracy at different taxonomic levels when including the conventional soil series map (CSS+) and when excluding the conventional soil series map (CSS−). RF is random forests. MLR is multiple logistic regression.
legacy soil map. These relatively small differences may be attributed to soil change between the time of the legacy soil survey and the current soil sampling, or the use of a coarse-scale (1:100,000) legacy soil map. These results suggest that legacy soil survey information be included as a digital soil mapping covariate in other geographical locations, but that caution be taken when legacy soil survey maps (particularly coarse-scale soil surveys) are used for soil survey update and/or disaggregation activities in areas where land management activities may have resulted in soil change.

Digital soil mapping models were more accurate than the conventional soil survey in this study. Thus we conclude that digital soil survey in this study. Thus we conclude that digital soil mapping is an attractive option for soil survey update in Iran and that digital soil mapping can likely be applied to the remaining unmapped areas in Iran.
References


Marchetti, A., Piccini, C., Santucci, S., Chiuciarelli, I., Francaviglia, R., 2011. Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data. Catena 85, 267–274.


