Geoderma 290 (2017) 51–68

Contents lists available at ScienceDirect

Geoderma

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

Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes

Brandon Heung [a,*], Matúš Hodúl [b], Margaret G. Schmidt [a]

[a] Soil Science Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[b] Department of Geography, University of Ottawa, 75 Laurier Ave E, Ottawa, ON, K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form 20 November 2016
Accepted 1 December 2016
Available online 29 December 2016

Keywords:
Digital soil mapping
Machine-learning
Soil classification
Data-mining
Model comparison
Ensemble-learning

ABSTRACT

Machine-learners used for digital soil mapping are generally trained using either data derived from field-observed soil pits or from soil survey polygons - although no direct comparison of the accuracy resulting from the two methods has yet to be undertaken. This study examined such a comparison over the Okanagan Valley and Kamloops region of British Columbia where good quality soil pit and soil survey data were available. A standard set of environmental variables including vegetative, climatic, and topographic indices were used to predict soil Great Groups in accordance with the Canadian System of Soil Classification. The pit-derived training dataset was developed using $n = 478$ points from the British Columbia Soil Information System while the polygon-derived training dataset was developed through random sampling of single-component soil survey map units based on an area-weighted approach. In both cases, the training points were intersected with a suite of 18 environmental covariates, reduced from 27 covariates using principal component analysis, and submitted to a machine-learner for predictions at a 100 m spatial resolution. Four single-model learners (CART, *k*-nearest neighbor, multinomial logistic regression, and logistic model tree) and five ensemble-model learners (CART with bagging, *k*-nearest neighbor with bagging, multinomial logistic regression with bagging, logistic model trees with bagging, and Random Forest) were compared. Surfaces of prediction uncertainty were produced using ignorance uncertainty and results were validated using a 5-fold cross-validation procedure. Predictions made using polygon-derived training data were consistently higher in accuracy across all models where the Random Forest model was the most effective learner with $C = 61\%$ accuracy when using pit-derived training data and $C = 68\%$ accuracy when using polygon-derived training data. Comparing single-model and ensemble-learner models, the bagging algorithm resulted in a 2–11% increase in accuracy when using pit-derived training data. Ensemble-models allowed for the visualization of prediction uncertainty. This study provides further insight into the use of legacy soil data and the development of training data for digital soil mapping.

1. Introduction

The soil-environmental variables identified in Jenny (1941) codified the concept of soil-environmental relationships, where easily measurable environmental properties could be used to predict soil properties. In digital soil mapping (DSM), the environmental-correlation concept (McKenzie and Ryan, 1999), later formalized within the *scorpan* model (McBratney et al., 2003), takes spatial soil data and co-locates it to readily available environmental data such as digital elevation models (DEM) and remotely sensed data in order to form the training dataset for a type of supervised learning. The relationships between soil and environmental conditions are correlated through the fitting of a model using machine-learning and/or geostatistical techniques where the soil-environmental relationships are then used to predict the soil properties for areas that have not been sampled. Furthered with increasing computational power, advancing remote sensing and GIS technologies, and the availability of accurate soil-environmental data, the application of the environmental-correlation concept has been applied for the mapping of soil classes and attributes over progressively larger spatial extents and data sizes (Chaney et al., 2016; Hengl et al., 2014, 2015; McBratney et al., 2003; Mulder et al., 2011, 2016).

Within the DSM literature, training data for mapping categorical soil properties has typically come from one of two sources: soil pit data or soil polygon data that has been digitized from conventional soil survey maps (Brungard et al., 2015; Heung et al., 2016). When using soil pit data for mapping soil taxonomic units, geolocated soil profile information is either recovered from a legacy soil database (Bui et al., 2006; Hengl et al., 2014) or based on field data that were collected for specific studies (Brungard et al., 2015; Rad et al., 2014). The use of pit data is particularly useful for situations when there is limited soil survey data

* Corresponding author.
E-mail addresses: brandon_heung@sfu.ca (B. Heung), mhodul@sfu.ca (M. Hodúl), margaret_schmidt@sfu.ca (M.G. Schmidt).

http://dx.doi.org/10.1016/j.geoderma.2016.12.001
0016-7061/

available, when there is an existing database of field observations, or when the spatial resolution of existing soil surveys is too coarse.

When using polygon data for model training purposes, the generic procedure typically involves the generation of training points within polygons where values from environmental covariates are extracted. This method has been used to map properties such as surficial geology and soil parent material (i.e. Bui and Moran, 2001) but has most commonly been used to map soil taxonomic units (i.e. Bui and Moran, 2001, 2003; Collard et al., 2014; Grinand et al., 2008; Odgers et al., 2014; Subburayalu and Slater, 2013; Subburayalu et al., 2014). The methods for generating training points have varied amongst studies – some of which included an area-weighted approach where the number of randomly generated sample points for each class were proportional to the class' areal extent (i.e. Moran and Bui, 2002); a by-polygon approach where a set number of training points were randomly generated within each polygon (i.e. Odgers et al., 2014); equal-class sampling where the number of randomly generated training points for each class were equal (i.e. Moran and Bui, 2002); and a sampling approach that integrated expert knowledge in the selection of points (i.e. Bulmer et al., 2016). Studies that have compared some of these methods have typically identified an area-weighted approach to produce more accurate predictions, relative to other methods, as the spatial extent and variability of the largest classes were better represented within the training data (Moran and Bui, 2002; Heung et al., 2014, 2016).

The main advantages of the polygon method include the ability for users to select an arbitrarily large sample size, which is beneficial for capturing more of the landscape's variability and the multivariate feature space of a categorical variable (Moran and Bui, 2002; Heung et al., 2014, 2016); furthermore, this approach has also been shown to be effective for the refinement and improvement of existing maps through the disaggregation of complex map units (Collard et al., 2014; Häring et al., 2012; Holmes et al., 2015; Odgers et al., 2014; Subburayalu et al., 2014). A concern with this approach has typically been related to the issue of map scale and the variability and purity within individual map units at given scales (Lin et al., 2005). For instance, in Heung et al. (2016), it was visually observed that as the map scale decreased from one region of the study area to another, there was a noticeable decrease in the diversity of soils that were predicted. Furthermore, soils developed from local-scale colluvial and fluvial processes were poorly predicted. The relationship between soil survey scale and the accuracy of predictions has also been observed in studies such as Bui and Moran (2003); in addition, that study also identified that the accuracy of predictions varied greatly even when soil surveys that were mapped at similar scales were used as training data due to differing survey methods and the time given to complete the survey.

Although these two approaches have commonly been used in the DSM literature, studies such as Brungard et al. (2015), Heung et al. (2016), and Lacoste et al. (2011) have identified a potential research gap where these approaches have yet to be directly compared using the same suite of machine-learners and environmental covariates over a study area. As such, the primary objective of this study was to address the comparison between pit-derived and polygon-derived data as training data for predicting soil classes at the Great Group level of the taxonomic hierarchy, based on the Canadian System of Soil Classification (Soil Classification Working Group, 1998), for the Okanagan-Kamloops region of British Columbia. Here, the pit-derived training data were obtained from legacy soil pit data taken from the British Columbia Soil Information System (BCSIS) (Sondheim and Suttie, 1983) and the polygon-derived training data were derived from legacy soil survey polygons using the framework provided in Heung et al. (2014). An identical suite of 27 environmental covariates and nine machine-learning algorithms for classification were tested on each of the two training datasets and as such, differences in the results could be constrained to the differences in training data. Secondary objectives included the comparison of nine machine-learning algorithms: four single-model learners and five ensemble-model learners. Validation of the predictions was performed using soil pit data and a cross-validation procedure.

2. Methodology

2.1. Study area

The study area was chosen due to the availability of both soil pit data and soil survey polygon data of high reliability and spatial quality, as well as a relatively high sampling density in the case of pit data. The study area represents a 47,350 km^2 portion of south-central British Columbia (Figure 1) and is located at approximately 49.0°N to 51.1°N latitude, 117.5°W to 120.8°W longitude, with an elevation range of 280–2720 m.

There is a great diversity of ecosystem types within the study area with the Interior Douglas Fir (IDF) and Ponderosa Pine (PP) biogeoclimatic zones making up much of the valleys and the Bunchgrass (BG) zone at the lowest elevations. The IDF is the largest zone in the valleys, with a mean annual temperature of 1.6–9.5 °C, and 300–750 mm of precipitation, 15–40% of which falls as snow (Hope et al., 1991b). The zone is largely covered by mature stands of Douglas Fir (*Pseudotsuga menziesii*) although grasslands occur in some places. Soils here are primarily Luvisols and Brunisols, with Chernozems occurring in the grasslands. Due to the basic volcanic parent material and low leaching rates in the arid environment, the soils are considered to have a high nutrient status (Hope et al., 1991b). The PP zone occurs below the IDF zone, and is the driest and warmest forested zone in British Columbia, with a mean annual temperature of 4.8–10 °C and 280–500 mm of precipitation. Soils here are much the same as in the IDF zone, consisting mostly of Chernozems and Brunisols. The lowest elevations, along valley bottoms of major rivers in the region, are occupied by the BG zone and is characterized by its warm, dry climate with sparse shrubs and grass cover, and Chernozemic soils (Hope et al., 1991a).

Higher elevations are characterized by the forested Montane Spruce (MS) and Interior Cedar Hemlock (ICH) zones with Engelmann Spruce Subalpine Fir (ESSF) and Interior Mountain Heather Alpine (IMA) zones located at the highest points (Ketcheson et al., 1991). The ICH zone has a mean annual temperature of 2.0–8.7 °C, and 500–1200 mm of precipitation of which 25–50% of it falls as snow. Humo-Ferric Podzols dominate at drier areas while Ferro-Humic Podzols and Gleysols occur in wetter areas. The MS zone occurs at slightly higher elevations, leading to lower mean annual temperatures of 0.5–4.7 °C and 380–900 mm of precipitation. The soils of the MS zone are mostly of the Brunisolic and Luvisolic orders formed from clayey volcanic parent material; however, Humo-Ferric Podzols can be found in areas that are moist with coarse parent materials (Hope et al., 1991c). The ESSF and IMA zones occur only at the highest elevations in the northeast portion of the study area, and represent only a small proportion of its total area.

2.2. Environmental covariates

27 environmental variables were derived from remote sensing, climate and digital elevation model (DEM) data (Table 1). In order to decrease multi-collinearity between the variables and computational demand, principal component analysis was performed on the topographic and vegetation data. The analysis resulted in a total of 18 covariates, which were then scaled in order to convert the covariate values into distributions with similar ranges – a procedure that is recommended for machine-learners (such as *k*-nearest neighbors) where the decision boundaries for classes are defined based on the distance in feature space between observed and unobserved points.

2.2.1. Topographic indices

Topographic indices were derived from a 100 m spatial resolution DEM of the study area, obtained from HectaresBC.org – a provincial repository of freely available environmental data. Consecutive smoothing was applied to the DEM in order to minimize the effects of spatially non-correlated noise on the calculation of topographic indices, in the form of three consecutive mean filters of 3×3, 3×3, and 5×5 pixels (Heung et