Machine learning for predicting soil classes in three semi-arid landscapes

Colby W. Brungard[a] Department of Plants, Soils and Climate, 4820 Old Main Hill, Utah State University Logan, UT, 84322, USA. Email: envsoilco@gmail.com. Corresponding author. Ph: +14357973404
Janis L. Boettinger[a] Department of Plants, Soils and Climate, 4820 Old Main Hill, Utah State University Logan, UT, 84322, USA. Email: janis.boettinger@usu.edu.
Michael C. Duniway[b] U.S. Geological Survey, Southwest Biological Science Center, 2290 SW Resource Blvd, Moab, UT 84532, USA. Email: mduniway@usgs.gov.
Skye A. Wills[c] National Soil Survey Center, Natural Resources Conservation Service – United States Department of Agriculture. 100 Centennial Mall North, Lincoln, NE 68508, USA. Email: skye.wills@lin.usda.gov.
Thomas C. Edwards, Jr.[d] U.S. Geological Survey, Utah Cooperative Fish and Wildlife Research Unit and Department of Wildland Resources, Utah State University, Logan, UT 84322, USA. Email: t.edwards@usu.edu.

Abstract

Mapping the spatial distribution of soil taxonomic classes is important for informing soil use and management decisions. Digital soil mapping (DSM) can quantitatively predict the spatial distribution of soil taxonomic classes. Key components of DSM are the method and the set of environmental covariates used to predict soil classes. Machine learning is a general term for a broad set of statistical modeling techniques. Many different machine learning models have been applied in the literature and there are different approaches for selecting covariates for DSM. However, there is little guidance as to which, if any, machine learning model and covariate set might be optimal for predicting soil classes across different landscapes.

Our objective was to compare multiple machine learning models and covariate sets for predicting soil taxonomic classes at three geographically distinct areas in the semi-arid western United States of America (southern New Mexico, southwestern Utah, and northeastern Wyoming). All three areas were the focus of digital soil mapping studies. Sampling sites at each study area were selected using conditioned Latin hypercube sampling (cLHS). We compared models that had been used in other DSM studies, including clustering algorithms, discriminant analysis, multinomial logistic regression, neural networks, tree based methods, and support vector machine classifiers. Tested machine learning models were divided into three groups based on model complexity: simple, moderate, and complex. We also compared environmental covariates derived from digital elevation models and Landsat imagery that were divided into three different sets: 1) covariates selected *a priori* by soil scientists familiar with each area and used as input into cLHS, 2) the covariates in set 1 plus 113 additional covariates, and 3) covariates selected using recursive feature elimination.

Overall, complex models were consistently more accurate than simple or moderately complex models. Random forests (RF) using covariates selected via recursive feature elimination was consistently the most accurate, or was among the most accurate, classifiers between study areas and between covariate sets within each study area. We recommend that for soil taxonomic class prediction, complex models and covariates selected by recursive feature elimination be used.

Overall classification accuracy in each study area was largely dependent upon the number of soil taxonomic classes and the frequency distribution of pedon observations between taxonomic classes. Individual subgroup class accuracy was generally dependent upon the number of soil pedon observations in each taxonomic class. The number of soil classes is related to the inherent variability of a given area. The imbalance of soil pedon observations between classes is likely related to cLHS. Imbalanced frequency distributions of soil pedon observations between classes must be addressed to improve model accuracy. Solutions include increasing the number of soil pedon observations in classes with few observations or decreasing the number of classes. Spatial predictions using the most accurate models generally agree with expected soil-landscape relationships. Spatial prediction uncertainty was lowest in areas of relatively low relief for each study area.

Keywords:

Digital soil mapping; Machine Learning; Recursive feature elimination; Random forests; Brier score

1. Introduction

Maps that predict the spatial distribution of soil taxonomic classes are of interest in many countries because they inform soil use and management decisions. Digital soil mapping (DSM) may have advantages over conventional soil mapping approaches as it may better capture observed spatial variability and reduce the need to aggregate soil types based on a set mapping scale (Zhu et al., 2001). A key component of any DSM activity is the method used to define the relationship between soil observations and environmental covariates. Many such methods have been investigated including expert systems (Smith et al., 2012, Van Zijl et al., 2012, Zhu et al., 2001), unsupervised classification (Boruvka et al., 2008; Triantifilis et al., 2012), and machine learning (Behrens and Scholten, 2006; Bui and Moran, 2003; Kim et al., 2012; Stum et al., 2010).

Machine learning is a general term for a broad set of models used to discover patterns in data and to make predictions (Witten et al., 2011). Although machine learning is most often applied to large databases, it is an attractive tool for learning about and making spatial predictions of soil classes because knowledge about relationships between soil classes and environmental covariates is often poorly understood (Grunwald, 2006). Machine learning techniques have been used to model soil depth classes (Boer et al., 1996), biological soil crust classes (Brungard and Boettinger, 2012), soil drainage classes (Campling et al., 2002; Liu et al., 2008) and the presence of diagnostic soil horizons (Jafari et al., 2012).

Several broad types of machine learning models have been applied for digital soil mapping of soil types, such as logistic regression (Hengl et al., 2007; Jafari et al., 2012; Kempen et al., 2012; Marchetti et al., 2011), classification trees (Bui and Moran, 2003; Kim et al., 2012; Scull et al., 2005), random forests (Barthold et al., 2013; Pahlavan Rad et al., 2014; Poggio et al., 2013; Stum et al., 2010), neural networks (Behrens et al., 2005; Jafari et al., 2013; Moonjun et al., 2010), and support vector machines (Kovačević et al., 2010). Although machine learning models have been tested in different landscapes around the world, it is rare for multiple models to be tested on the same landscape.

Two general approaches have been applied to predicting soil taxonomic classes using machine learning. The first approach attempts to find and extract soil class-landscape relationships from existing digitized soil polygon maps when the exact locations (GPS coordinates) of soil pedon observations are unknown (Behrens et al., 2005; Grinand et al., 2008; Subburayalu and Slater, 2013). The second approach attempts to construct soil class-landscape relationships from soil pedon observations made by field sampling at known locations (Barthold et al., 2013; Hengl et al., 2007; Jafari et al., 2012; Kempen et al., 2012; Kim et al., 2012; Stum et al., 2010). The choice of approach largely depends on the availability of soil pedon observations with known locations.

There have been few studies that compare DSM methods for categorical data such as soil types or classes, especially when soil-landscape relationships were developed from soil pedon observations. Of the studies that used soil pedon observations to construct soil class-landscape relationships (e.g., Barthold et al., 2013; Jafari et al., 2012; Kempen et al., 2012) few compared more than two machine learning models, and none compared multiple machine learning models at more than one study area. To address this knowledge gap, we compared multiple machine learning models for predicting soil classes in multiple study areas using soil pedon observations. Specifically, we compared eleven machine learning models for predicting subgroup classes in Soil Taxonomy (Soil Survey Staff, 1999) using soil pedon observations at three geographically distinct areas in the western United States of America (southern New Mexico, southwestern Utah, and northeastern Wyoming; Fig. 1). Each study area was the focus of a digital soil mapping study and represented a broad range of semi-arid landscapes with different soil-landscape relationships.

Model performance depends on the covariates used to represent soil-landscape relationships and covariate selection is an important aspect of digital soil mapping (Vasques et al., 2012; Xiong et al., 2014). Therefore, we also compared the influence of three groups of environmental covariates on machine learning model performance in each of the three study areas: 1) covariates selected *a priori* by soil scientists familiar with each area (expert knowledge; Zhu et al., 2001), 2) the covariates in set 1 plus 113 additional covariates derived from digital elevation models and Landsat imagery at several resolutions that represented a large suite of potentially useful covariates, and 3) a subset of covariates identified using recursive feature elimination (Guyon et al., 2002) from covariate sets 1 and 2.

Identifying which of the many available machine learning models and which of the many available covariates are appropriate for predicting soil classes from soil pedon observations in a given landscape would be useful where efficiencies are necessary for operational DSM. In this paper, we demonstrate that complex models using covariates selected by recursive feature elimination resulted in the most accurate predictions.

2. Methods

2.1. Study Areas

2.1.1 New Mexico (NM)

The New Mexico (NM) study area is located on Otero Mesa in the northern reaches of the Chihuahuan Desert, approximately 130 km northeast of El Paso, TX, USA. Centered at 105.6° W longitude, 32.5° N latitude (Fig. 1), the area is approximately 190 km^2. The underlying geology is primarily limestone and sandstone (Green and Jones, 1997). Soil parent material is primarily calcareous alluvium but also includes eolian sands and residuum. Vegetation is a mix of shrublands (primarily creosote bush [*Larrea tridentata*] and tar bush [*Florencia cernua*]) and grasslands (primarily black grama [*Boutaluoa eriopoda*] and tobosa [*Pleuraphis mutica Buckley*]). Elevation ranges from 1430 to 1915 m. The soil moisture regime is aridic bordering on ustic. Mean annual precipitation is 354 mm, the majority of the precipitation arrives between June and December, and mean annual temperature is approximately 15 °C (PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu/, accessed 4 March 2014).

2.1.2 Utah (UT)

The Utah (UT) study area is located in the eastern Great Basin physiographic province, approximately 14 km southwest of Milford, UT, USA. Centered at 113° W longitude and 38° N latitude, the area is approximately 300 km^2 and consists of mountainous terrain and associated alluvial fans formed from a complex mix of limestone, dolomite, quartzite, basalt, quartz monzonite, quartz latite, shale, sandstone, andesite, rhyolite, granite, and ash flows (Best et al., 1989). Elevation ranges from 1540 to 2100 m. Vegetation consists of shrubs (primarily Wyoming big sagebrush [*Artemisia tridentata*] and black sagebrush [*Artemisia nova*]) and bunch grasses (Indian ricegrass [*Achnatherum hymenoides*]) at lower elevations, while trees (primarily Utah Juniper [*Juniperus osteosperma*] and Singleleaf Pinyon [*Pinus monophylla*]) dominate higher elevations. The soil moisture regime is aridic bordering on xeric in lower elevations and xeric in higher elevations. Mean annual temperature and precipitation for the nearest weather station (Milford, UT) is 11°C and 200 mm, respectively, the majority of the precipitation arrives in April and October (Western Regional Climate Center, 2013).

2.1.3 Wyoming (WY)

The Wyoming (WY) study area is located in the Powder River Basin of Wyoming, USA, part of the Northern Rolling High Plains (United States Department of Agriculture, 2006), approximately 43 km southwest from Gillette, WY. Centered at approximately 106° W longitude and 44° N latitude, the area is approximately 296 km^2. Geology in the area consists of variegated mudstone, sandstone, conglomerate, limestone, shale and coal (Cole and Boettinger, 2006; Green and Drouillard, 1994) . Topography is a mix of bedrock-controlled, low rolling hills and badlands (locally known as the “Powder River Breaks”) a system of steep, bedrock-controlled hills and gullies (gullies commonly > 6 m deep) with extremely high rates of erosion and low vegetation cover (Cole, 2004). Vegetation is characterized by a mixture of mid-stature cool season grasslands (bluebunch wheatgrass [*Pseudoroegneria spicata*] and needle-and-thread [*Hesperostipa comata*]) and sagebrush shrublands (Wyoming big sagebrush [*Artemisia tridentata*]) (United States Department of Agriculture, 2006). Elevation ranges from 1220 and 1600 m. The soil moisture regime is aridic bordering on ustic. Mean annual temperature and precipitation is 8°C and 310 mm, respectively, with the majority of the precipitation falling between April and October (Western Regional Climate Center, 2013).

2.2 Sampling

Sampling locations for each study area were selected using conditioned Latin hypercube sampling (cLHS) (Minasny and McBratney, 2006). Covariates used for input into cLHS were chosen by soil scientists familiar with each study area and assumed to best represent soil-landscape relationships and anticipated soil forming processes in each area (covariate set 1). The soil scientists who selected cLHS input covariates for the NM study area had worked inside the study area and in similar landscapes for approximately ten years. The soil scientist who selected cLHS input covariates for the UT study area had visited the area, performed three months of field sampling in an adjacent area, and conducted a literature review to identify important covariates in similar landscapes. The soil scientists who selected cLHS input covariates for the WY area were Natural Resource Conservation Service (NRCS) soil scientists who were conducting traditional soil surveys in similar landscapes around the study area.

In each area, soils were manually excavated to a depth of at least 100 cm, or root limiting layer if shallower, and were sampled and described according to Schoeneberger et al. (2003). Soil Taxonomy (Soil Survey Staff, 1999) defines the following hierarchical levels of classification: order, suborder, great group, subgroup, family, and series. We chose to model at the subgroup class as this level of classification existed for the soils described in each study area. Rock outcrop and Badland were also included at the subgroup level. For each area, subgroup classes with only 1 observation were grouped with the most similar subgroup class.

2.2.1 New Mexico cLHS

Covariates used for cLHS were derived from an October 2006 Landsat 5 TM image and a 5-m Lidar digital elevation model (DEM). Imagery covariates from Landsat were band 5 (short wave infrared) plus band 2 (green), band 5 minus band 2, and a normalized band 5/2 ratio ([Band 5-Band 2]/[Band 5+Band 2]). Terrain attributes were aspect in degrees, elevation, slope, and a multipath wetness index (Shi, 2013) calculated at four slope resolutions (5, 10, 25, 35 m) from the DEM. A categorical terrain classification was also used. Imagery covariates were chosen for use in cLHS because they had been shown to correlate with soil surface properties. Slope and the multipath wetness index, were chosen to represent potential soil moisture distribution. Aspect and elevation were chosen to represent microclimate and potential soil moisture (higher elevation, north-facing areas often have more potential soil moisture than lower elevation, south-facing areas. The terrain classification consisted of seven classes related to elevation and slope.

Initially 200 potential sampling sites were identified, but because of logistical constraints it was impossible to visit all 200 sites. To select a smaller set of representative sampling locations cLHS was used to produce a hierarchical nested set (each smaller sample size was a subset of the previous larger sample, Webster et al., 2006) of 175, 150, 125 and 100 potential sampling sites from the original 200 sites. All sites in the 100 subset were visited, plus an additional three sites. In total 103 soil sampling locations were observed (Fig. 2). Each soil observation was classified to family level in Soil Taxonomy. Ten subgroup classes were extracted from family names (Table 1).

2.2.2 Utah cLHS

Covariates used for cLHS were derived from an atmospherically corrected (Chavez, 1996) July 31st 2000 Landsat 7 ETM+ image and a 10-m hydrologically correct DEM. A soil adjusted vegetation index (SAVI) was derived from the imagery using an L value of 0.5 (Heute, 1988). Terrain attributes were slope, inverse wetness index (Tarboton, 2013) and transformed aspect (a measure of northness vs. southness). Land cover and geologic type were also used. Land cover type was obtained from the Southwest Regional Gap Analysis Program (Lowry et al., 2007). Geology was obtained from a United States Geological Survey 1:50,000 geology map (Best et al., 1989). Land cover and SAVI were chosen because it was anticipated that vegetation type and density was correlated with soil properties such as soil depth. Geologic type was chosen because the highly complex geology in this area was anticipated to exert a strong control on potential pedogenesis. Terrain covariates were chosen to represent microclimate, because microclimate heavily influences soil moisture, which in turn influences pedogenesis.

Three hundred locations were visited. Soil pedons were excavated, described, and classified to family level. Subgroup classes were extracted from family names. Three soil observations were excluded from modeling as they were located in highly disturbed areas. This resulted in 297 soil observations in 15 subgroup classes (Fig. 3, Table 1).

2.2.3 Wyoming cLHS

Covariates used for cLHS were derived from a Landsat 5 TM image and a 2-m Lidar DEM. Imagery covariates were Normalized Difference Vegetation Index (NDVI) and band ratios 5/2 and 5/7. Terrain derivatives were topographic wetness index, topographic position index, stream power index (Wilson and Gallant, 2000) and distance to the nearest road. All covariates for cLHS, except distance to the nearest road, were selected using the Optimum Index Factor (OIF). OIF identifies the combination of input covariates that maximize variability, with the lowest correlation among covariates (Kienast-Brown and Boettinger, 2010). Distance to the nearest road was included for a vegetation sampling project not directly related to soil mapping.

Similar to the NM study area, cLHS was used to select hierarchical nested sets of 150, 100, and 50 potential sampling sites from 200 original sampling sites. Fifty-seven soil pedon observations were made: the set of 50 nested cLHS samples plus an additional seven pedon observations (Fig. 4). Each soil pedon was excavated, described, and assigned to a soil series. Subgroup classes were extracted from each series using official soil series descriptions (https://soilseries.sc.egov.usda.gov/osdname.asp). This resulted in 5 subgroup classes (Table 1).

2.3 Additional covariates

Additional terrain covariates were created from a 5-m Lidar derived DEM for the NM study area, a 5-m auto-correlated DEM (Utah Automated Geographic Reference Center, 2013) for the UT study area and resampling the 2-m WY Lidar DEM to 5-m. Terrain covariates were created in R (R Core Team, 2012) with the RSAGA package (Brenning, 2008). For each area the following terrain covariates were created: slope, total curvature, plan and profile curvature, SAGA wetness index, catchment area, catchment slope, modified catchment area, convergence index, morphometric protection index (Yokoyama et al., 2002), multi-resolution index of valley bottom flatness and multi-resolution index of ridge top flatness (Gallant and Dowling, 2003), topographic position index, and terrain ruggedness index. Definitions of individual terrain covariates can be found in Wilson and Gallant (2000) and Hengl and Reuter (2008).

Estimated potential direct, diffuse, total, and the duration of incoming solar radiation of the approximate growing season in each area were also calculated. All potential incoming solar radiation was calculated for clear sky and standard atmosphere conditions, and represent potential solar radiation in the absence of clouds or significant amounts of atmospheric aerosols. All terrain and potential solar radiation covariates were calculated at 5, 10, 30, 50, and 100 m cell sizes. Digital elevation models with 10, 30, 50, and 100 m cell sizes were created from the 5-m DEMs by averaging over blocks of cells at these resolutions. The morphometric protection index calculated at 100-m cell size was not used because at this resolution there was no variance in the covariate. This resulted in 89 terrain covariates for each area.

For each area, we selected Landsat 5 TM imagery from 2 different dates. Each image pair consisted of an image acquired during a season of peak vegetation growth and a season of dormant vegetation. Each image was atmospherically corrected using the "Cost without Tau" method (Chavez, 1996) in the R Landsat package (Goslee, 2011). From each image the following covariates were created: normalized band ratios 5/2, 5/7, 3/1, and 1/7; NDVI; six bands of the tasseled cap transformation (Crist and Kauth, 1986); and greenness above bare soil (GRABS) index (Jensen, 2005). This resulted in 24 imagery covariates for each area. Total additional terrain and imagery covariates for each area were 113 (covariate set 2).

These covariates represent a wide range of topographic and spectral derivatives commonly used for DSM in the western USA (Boettinger, 2010), but these additional covariates are not exhaustive of the potentially available covariates. For example, in other DSM studies, Heung et al. (2014) included distance to the nearest stream/river and relative hydrological slope position. Behrens et al. (2010) used elevation differences from the center pixel of a DEM as predictor covariates. Xiong et al. (2012) used covariates such as LANDFIRE (Landscape fire and resource management tools project) vegetation maps and geospatial land cover maps as vegetation related covariates. Poggio et al., (2013) used multi-temporal MODIS (Moderate Resolution Imaging Spectroradiometer) vegetation and drought indices. Taylor et al. (2013) used potential evapotranspiration from ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) imagery. Although a wide range of potential covariates exist, we chose to incorporate the specific terrain and imagery covariates in covariate set 1+2, because they were easily calculated with the available software with which we were familiar, and because we anticipated these covariates to adequately characterize soil distribution in these areas. While relatively coarse-resolution (3rd order soil survey; Soil Survey Division Staff, 1993) soil maps were available for the NM and WY study areas (the UT area was previously unmapped), we did not include existing soil maps in covariate set 1+2 for these areas in an effort to keep all covariate sets as consistent as possible. Geological maps were not included in covariate set 1+2, because only a single geological unit was mapped in the NM and WY areas.

2.4 Covariate selection

Recursive feature elimination (Guyon et al., 2002; Kuhn and Johnson, 2013) was used to identify an optimal subset of covariates from the set of all available covariates (covariate set 1+2) for each area (Fig. 5). Recursive feature elimination identifies optimal subsets (lowest misclassification error) of predictor covariates by constructing a classification model with all predictor covariates, ranking each predictor covariate, eliminating the covariate(s) with the lowest importance, and repeating this procedure until a predefined threshold is reached or only one predictor covariate remains. Xiong et al. (2012) used recursive feature elimination to identify important predictors for digital soil mapping of soil carbon in Florida.

For each study area, random forests (Liaw and Wiener, 2002; parameters mtry = default and ntree = 1000) was used to calculate covariate importance, as random forests is not highly sensitive to non-informative predictors (Kuhn and Johnson, 2013). Random forests identifies important covariates by generating multiple classification trees (a forest) using bootstrap sampling, randomly scrambling the covariates in each bootstrap sample and reclassifying the bootstrap sample. The misclassification error of the bootstrap sample (termed the “out-of-bag” error) using the scrambled covariate is compared to the misclassification error using the original covariate and the percent difference is used as a measure of covariate importance (Peters et al., 2007). Important covariates will have a large increase in “out-of-bag” error. For each area, the optimal subset of covariates was identified as the subset of covariates with the minimum OOB error (Fig. 5).

For the UT study area, although a set of twelve covariates returned the absolute lowest misclassification error (OOB error = 0.512), we selected a set of six covariates (OOB error = 0.520) as optimal for a more parsimonious model. Selected covariates ranked by importance (covariate set 3) are listed in Table 3.

2.4 Modeling

All modeling was performed using the caret package (Kuhn et al., 2013) in R (R Core Team, 2012). We tested eleven classification models for each area (Table 2). Each model was chosen based on a review of machine learning methods used in other published DSM literature. Selected machine learning models represented several broad classes of machine learning techniques and included multinomial logistic regression, tree based classifiers, neural networks, support vector machines, and clustering methods. An accessible explanation of all tested models can be found in Kuhn and Johnson (2013) and James et al. (2014). Tested models were divided into three groups based on model complexity: simple, moderate, and complex (Table 2). Models were assigned to one of the three groups based on the interpretability and number of parameters of each model. Complex models (e.g., support vector machines and neural networks) were difficult to interpret (i.e., black-box models) and had many parameters. Simple models were interpretable and had few parameters, while medium complexity models were between simple and complex models.

The goal of machine learning is to find a useful approximation of the function that underlies the predictive relationship between input covariates and desired outcomes (Hastie et al., 2001). In this study input covariates were derived from DEM’s and Landsat imagery and the desired outcomes were subgroup classes. Each type of model (e.g., support vector machines, neural networks) has specific and different required parameters (referred to as tuning parameters) that control how the relationship between input covariates and outcomes is defined. These parameters must be optimized to generate the best “fit” possible between covariates and outcomes.

For each model leave-group-out cross-validation was used to select optimal tuning parameters (Kuhn, 2014). Leave-group-out cross validation involved randomly splitting the pedon observations into training and test sets, using the training set for model construction and the test set for model validation, then repeating this process. We used a 70%/30% training/testing split (70% of the pedon observations were used for model training and 30% for model testing) repeated 100 times for each area. Although splitting observations into separate training and test sets (no cross validation) is a standard approach taken in other DSM studies (e.g., Henderson et al., 2005; Tesfa et al., 2009; Pahlavan Rad et al., 2014) we observed that use of a single training/test set resulted in accuracy metrics (e.g., Kappa and the Brier Score; Section 2.5) with high variance. Ninety-five-percent confidence intervals were used to assess the variability in accuracy metrics over the repeated test sets.

For each required model parameter (the number of required model parameters ranged between 0 and 2) ten potential candidate values were defined. This resulted in an *n* x 10 matrix of potential model tuning parameters, where *n* = the number of required parameters. Models were tuned using each set of parameters, and the average Kappa (Section 2.5) was calculated over the 100 repeated training/test splits. Optimal parameters were chosen using the one-standard-error rule (James et al., 2014), where the simplest (smallest) tuning values within one standard error of the tuning parameters which produced the maximum kappa, were selected as optimal (Kuhn, 2008). For those models that did not require tuning parameters (bagged classification tree, linear discriminant analysis) no optimization was possible.

Each model was applied to three sets of covariates for each area: the soil scientist selected covariates used as input into cLHS (covariate set 1), the covariates in set 1 plus the 113 additional terrain and imagery covariates that we created (covariate sets 1 + 2), and those covariates that were selected by recursive feature elimination from all available covariates (covariate set 3). Because some models required covariates to have similar ranges (e.g., K-nearest neighbors), all environmental covariates were centered and scaled to have mean = 0 and standard deviation = 1 before use. Multinomial logistic regression and linear discriminant analysis could not be fit using covariate set 1+2.

When using covariate sets 1+2, any cLHS covariate that was duplicated by the additional terrain and imagery covariates (e.g., slope) was removed. Additionally, geologic type and distance to roads were removed from covariate sets 1 and 2 for the UT and WY study areas, respectively; because the geology covariate did not cover the entire study area, and distance to roads was included for another purpose not thought to be related to soil taxonomic classes (impact of disturbance on vegetation) in the initial cLHS.

2.5 Model accuracy comparison

Kappa analysis and Brier scores (Brier, 1950) were used to compare model accuracy. The kappa statistic (κ) (Congalton, 1991) is a measure of classification accuracy accounting for chance agreement (Congalton and Green, 1998). Accounting for change agreement is an important consideration when dealing with highly imbalanced classes as high classification accuracy could result from classifying all observations as the largest class (Congalton and Green, 1998). The κ of a random classifier would be 0 whereas a κ of 1 would indicate perfect classification (Congalton, 1991). Kappa values greater than 0.80 represent strong agreement, values between 0.4 and 0.8 represent moderate agreement, and values below 0.4 represent poor agreement (Congalton and Green, 1998).

Brier scores account for the difference between the true class and probability (or probability-like) estimates of the true class (Johansson et al., 2010) as:

$$BS = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}\left(F_{ij} - E_{ij}\right)^2$$

where r = number of taxonomic classes, n = number of observations in the test set, F_{ij} is the probability estimate that observation n_i belongs to class r_j, and E_{ij} is an indicator covariate such that E_{ij} = 1 if n_i was the subgroup class and 0 otherwise. Brier scores range between zero and two, with lower scores indicating better model performance. A Brier score of 1.25 indicates that each taxonomic class was predicted with the same probability. Brier scores for both linear and radial support vector machines were not calculated, because support vector machines require more than three observations per class to calculate probability estimates and several subgroup classes in each area had three or less observations (Table 1).

Models with the highest κ and lowest Brier scores were determined to be the most accurate model for each site. T-tests were performed to determine if differences in Kappa and Brier scores between models were statistically significant at the 0.05 level. The percent correctly classified (PCC) and producer's accuracy of individual subgroup classes from the most accurate model, averaged over all cross-validation repetitions, were also calculated.

In addition to Kappa, and Brier scores, spatial predictions from each model identified as potentially optimal were visually inspected for pedologically meaningful patterns. The uncertainty associated with each cell of the spatial predictions was assessed using the confusion index (Burrough et al., 1997; Odgers et al., 2011):

$$CI = \left[1 - \left(\mu_{max} - \mu_{(max-1)}\right)\right]$$

Where μ_{max} is the probability value of the class with the maximum probability at each cell and μ_{max-1} is the second largest probability value at the same cell. The confusion index ranges between 0 and 1; high CI values indicate greater uncertainty in subgroup class predictions.

3. Results

Models built using covariate set 3 had the highest κ for all three study areas (Figs. 6, 8, & 10). The model with the highest average κ for the NM study area (κ = 0.32 ± 0.09) was support vector machines using a radial basis function (SVMR; Fig. 6); however, t-tests indicated no significant difference in κ existed between SVMR and random forests (RF). Random forests (RF) had the highest average κ for both the UT (κ = 0.19 ± 0.06; Fig. 8) and the WY study areas (κ = 0.53 ± 0.14; Fig. 10). Kappa values for the WY study area represent moderate agreement between observed and predicted subgroup classes, while κ for the NM and UT study areas represent low agreement between observed and predicted subgroup classes. The models with the highest κ also had the highest percent correctly classified (PCC) for each area; PCC was 47 ± 0.07%, 43 ± 0.04%, and 72 ± 0.08%, for the NM, UT, and WY study areas, respectively.

Models constructed using covariate set 3 had the lowest Brier scores for the NM and WY study areas (Figs. 7 & 11). The lowest Brier score for the UT area was obtained using covariate set 1+2 (Fig. 9), but t-tests indicated no significant difference existed between models with the lowest Brier scores from covariate set 1+2, and covariate sets 1 (t = 5.34, df = 198, p-value = 2.537e-07) and 3 (t = -2.52, df = 198, p-value = 0.0126) for this area. Random forests (RF) was the model with the lowest average Brier score for the NM (BS = 0.70 ± 0.05; Fig. 7), UT (0.70 ± 0.01; Fig. 9) and WY study areas (0.46 ± 0.08; Fig. 11).

Average individual subgroup class producer's accuracy ranged between 0 and 86 percent (Table 1). The number of optimal covariates as determined by recursive feature elimination for each study area ranged between six and ten and included terrain derivatives at multiple cell sizes as well as several Landsat derivatives (Table 3). Spatial predictions using the model identified as the most accurate for each area generally met expected soil-landform patterns (Figs. 12A, 13A, & 14A). Confusion index values ranged between 0.46 and 0.99 for the NM study area (Fig. 12B), 0.53 and 0.99 for the UT study area (Fig. 13B), and 0.04 and 0.98 for the WY study area (Fig. 14B).

4. Discussion

4.1 Model performance

Random forests (RF) models using covariates selected by recursive feature elimination (covariate set 3) were consistently the most accurate, or was among the most accurate, classifiers (had the highest κ and lowest Brier score), between study areas and between covariate sets within each study area (Figs. 5-10). Although, single-hidden-layer neural networks (NNET), multilayer-perceptron neural networks (MLP), and nearest shrunken centroids (NSC) had slightly lower average Brier scores than random forests (RF) for the UT study area (Fig. 9) the differences were minimal. The consistency of random forests (RF) and covariate set 3 for producing the most accurate subgroup classifications across all study areas is likely because random forests was used in the recursive feature elimination procedure, which optimized covariates for subgroup class prediction (Section 2.4).

In addition to random forests (RF), radial-basis support vector machines (SVMR) and single-hidden-layer neural networks (NNET) had competitive accuracy metrics for subgroup class prediction in the NM (Fig. 6) and UT (Fig. 9) study areas, respectively. If multiple models are applied for a digital soil mapping project and accuracy metrics are approximately equivalent between models, then model averaging (Malone et al., 2014) may be appropriate.

Across all study areas, complex models (Table 2) were better classifiers than simple models. As recursive feature elimination (RFE) does not require a specific model (although random forests is convenient for RFE) and as complex models produced more accurate predictions than did simpler models, this suggests that the most accurate soil taxonomic class predictions will be produced using a combination of RFE and complex models. Covariate reduction methods similar to RFE, also resulted in the most accurate soil carbon models in Florida, USA (Xiong et al., 2014).

As the model with the highest classification accuracy for each study area is of most interest for predicting soil subgroup classes we restrict further discussion to random forests models using covariate set 3 when discussing differences in classification accuracy between study areas. Differences in classification accuracy between study areas can be partially attributed to the number of soil subgroup classes and the frequency distribution (the balance of observations between subgroup classes) of soil pedon observations. The UT study area was the least accurately modeled, had the most soil subgroup classes (n = 15), and the most skewed frequency distribution of soil pedon observations between subgroup classes. Two subgroup classes for the UT study area contained approximately 70% of the total soil pedon observations (Table 1). In contrast, the WY study area, the most accurately classified study area, had the fewest soil subgroup classes (n = 5) and a somewhat more balanced soil pedon observation distribution frequency. The classification accuracy, number of soil subgroup classes (n = 10) and soil pedon observation distribution frequency for the NM study area was between those of the UT and WY study areas. This suggests that overall classification accuracy will be highest when there are many soil observations, few soil classes, and the frequency distribution of soil observations between classes is approximately equal.

The frequency distribution of soil pedon observations heavily influenced individual subgroup class accuracies (Table 1). In general, classes with lower sampling frequencies were modeled less accurately. This finding is consistent with data presented by others (Barthold et al., 2013; Hengl et al., 2007; Kim et al., 2012; Marchetti et al., 2011; Stum et al., 2010; Taghizadeh-Mehrjard et al., 2012) and is likely because there are simply not enough observations to separate such classes in feature space.

The number of soil subgroup classes per study area appears related to the inherent variability of the given landscape. Areas with high geological and topographical complexity likely experience complex relationships between soil forming factors that lead to increased diversity in soil types. For example, the geologically and topographically complex UT study area had more subgroup classes than did the less complex NM or WY sites (Table 1).

The frequency distribution of soil pedon observations between subgroup types in a study area is likely a result of the sampling strategy used to select sites. Conditioned Latin hypercube sampling is a sampling method designed to identify sampling sites which represent the multivariate distribution of input environmental covariates and assumes that the input environmental covariates are related to the covariate of interest (Minasny and McBratney, 2006). Environmental covariates used as input to cLHS for each study area were selected to represent broad soil-landscape relationships. Our results suggest that in complex landscapes where likely many different soil types exist, such input environmental covariates result in adequate sampling of the most frequent soil types, but not of rare soil types (e.g., the UT study area).

As accurate modeling of soil classes depends on the number of classes and the frequency distribution of soil pedon observations between classes (many classes with few observations = poor model performance) such imbalance must be addressed for accurate modeling. There are two options to address such challenges: 1) increase observation number in classes with few observations and 2) decrease the number of classes.

Increasing the number of observations in classes with few observations may be difficult given financial and logistical constraints, and because it is likely difficult to identify *a priori* which classes will need to be more intensively sampled. However, this might be addressed using a combination of cLHS and targeted sampling or case-based reasoning (Shi et al., 2009), where the soil surveyor could manually identify likely locations of rare soil types. This may be especially useful in arid and semi-arid regions where small, localized areas often contain significant diversity when compared to the majority landscape.

The second option is to decrease the number of taxonomic classes. This could be accomplished by: 1) combining similar classes and 2) modeling separate sub-areas. Combining similar subgroup classes could be accomplished by using higher taxonomic levels such as great group or suborder. Modeling higher taxonomic levels would likely increase model accuracy (Jafari et al., 2013), but a trade-off between taxonomic level and soil information usefulness exists. Many decisions about soil use and management are based on soil differences not captured by higher taxonomic levels (i.e., order, suborder, and great group), so combining subgroup classes into higher taxonomic levels may miss important differences in soil function and likely not provide useful information for soil management decisions.

Ideally, digital soil mapping would be able to accurately model all levels of Soil Taxonomy including soil series. Soil series are the finest level of Soil Taxonomy (Soil Survey Staff, 1999) and two levels finer that what was predicted in this study. However, accurate predictions of soil series may not possible, because soil series are often defined by soil morphological diagnostic criteria that may not be well represented by environmental covariates. For example, the difference between Xeric Haplocalcid and Durinodic Xeric Haplocalcid subgroup classes (UT study area, Table 1) is based on the occurrence of cemented silica masses (durinodes). Such differences may not be identifiable with the terrain and spectral covariates commonly used for digital soil mapping.

Similar classes could also be combined based on a particular soil property (e.g., bedrock contact). This would result in a focus on the specific property while excluding other potentially important soil properties. Likely any such decision to group classes by soil property types would best be made by the user of the soil information. Additional options may be to combine classes with few observations into a single class denoted as “other soil classes”, or to add rare soil class observations to larger taxonomic classes. This approach has been taken by others (Pahlavan Rad et al., 2014), but we decided against doing so, because we suspected that classes with few observations might be topographically and spectrally distinct and thus be accurately predicted. Although, several subgroup classes with relatively few observations were predicted with moderate accuracy (e.g., Xeric Torriorthents in the UT study area (*n* = 6, average producer’s accuracy = 40%) and Lithic Ustic Haplocambids in the NM study area (*n* = 3, average producer’s accuracy = 50%); Table 1) individual class accuracies (Table 1) generally do not indicate this to be the case, and so in retrospect such a pragmatic approach is probably wise.

Modeling separate sub-areas might also decrease the number of taxonomic classes by reducing the area and thus the number of soil types considered in a model. For example, it is likely that the number of subgroup classes in one model would have been fewer had the UT study area been segregated into uplands and alluvial fan sub-areas. Although such an approach would increase the number of required models in proportion to the number of chosen sub-areas, this is theoretically appealing as different pedo-geomorphic sub-areas are likely to have different relationships between subgroup classes and environmental covariates (McBratney et al., 2003).

Another option to increase model accuracy could be to apply a weighting scheme to soil classes with few observations during model construction. This might improve classification accuracy, but for highly imbalanced datasets weighting can severely decrease the accuracy of the majority classes and result in apparent over prediction of the small classes (Stum et al., 2010). Additionally, using taxonomic distance (Minasny and McBratney, 2007) instead of misclassification error as the loss function to minimize during model training may result in increased model accuracy. We did not incorporate taxonomic distance in this study as it does not currently exist for Soil Taxonomy subgroup classes. Overall, increasing model accuracy is likely to require several of these options (increasing observation numbers, reducing class numbers, the use of a weighting scheme, and incorporation of taxonomic distance), and that applicable options will best be identified on a project-by-project basis.

4.2 Covariate set comparison

Surprisingly, models using all available covariates (covariate set 1+2) were as accurate, or slightly more accurate (higher κ, lower Brier scores), than models using the covariates selected by soil scientists (covariate set 1) for each area (Figs. 6-10). As covariate set 1 was selected by soil scientists anticipating how soil-landscape relationships would be best represented for modeling, the fact that this covariate set did not result in the most accurate models suggests that soil scientists may be unable to *a priori* identify optimal covariates for predicting taxonomic classes. In hindsight, this is not entirely surprising given the complexity of soil taxonomic classes and the disparate kinds of knowledge needed to predict these relationships *a priori*. Soil taxonomic classes represent multiple soil forming factors operating over long periods of time (likely decades to millennia) at several scales. Thus choosing optimal predictive covariates for modeling requires knowing both 1) how, and the scale at which, multiple soil forming factors vary across the landscape to produce soil taxonomic classes and 2) how those factors are best distinguished using spectral and topographic covariates. Being able to do both requires extensive familiarity with the local landscape and an understanding of terrain modeling and remote sensing. This suggests a pressing need for further investigation into relationships between specific environmental covariates and soil forming processes.

In addition to producing models with the highest accuracy, covariate set 3 may also provide information about the processes controlling soil type distribution across each study area landscape. The NM area mostly consists of broad, gently sloping, southward facing alluvial surfaces. More than half of the optimal covariates for this study area were related to catchment-scale patterns of potential soil moisture (multipath wetness index, catchment area and catchment slope; Table 3). We attribute this to the correlation of run-on/run-off relationships, landscape stability, and soil formation observed in this region (Gile et al., 1981). Vegetation related covariates (tasseled cap greenness transformation and the GRABS index) selected in covariate set 3, were likely related to the strong control of soils in determining vegetation cover and composition in the study area (Bestelmeyer et al., 2006, Duniway et al., 2010). Thus covariates related to catchment scale patterns of potential soil moisture and vegetation indices may be the best predictors in similar landscape settings. Similar settings include the large alluvial fans and bajadas (coalesced alluvial fans) that extend from mountain fronts into the valleys of many semi-arid and arid landscapes. Interestingly, topographic shading is an important covariate for both the UT and WY areas, but not the NM area. This is likely because landforms in the NM area are mostly southward facing with little vertical relief.

The optimal covariates for the UT study area were related to topographic shading (diffuse insolation), slope, slope position, and terrain ruggedness (Table 3). The UT area was highly variable in topographic relief. This local topography strongly influences soil erosion and deposition as well as the amount of incoming solar radiation, which in turn influences soil distribution (Beaudette and O'Geen, 2009). As the UT area had the greatest geologic complexity between the three study areas, it is surprising that covariates related to geology (Landsat band ratios 5/2, 5/7) were not among those identified as optimal. This may be because the influence of local topography exerted a stronger effect on soil development than did the relatively larger scale influence of geology. In semi-arid steeply sloping uplands and mountainous erosional landscapes, covariates related to soil erosion/deposition processes and solar radiation may be the most useful for predicting soil distribution.

The WY area is generally composed of rounded hills which have been dissected by numerous small drainages and lacks the topographic relief of the UT area or the broad alluvial slopes of the NM area. The optimal covariates for this area were plan and total curvature, topographic shading (diffuse insolation), catchment slope and Landsat band ratio 5/2 (Table 3). As three of the seven optimal covariates were related to slope curvature which approximates flow convergence/divergence (Wilson and Gallant, 2000) and as topographic shading was also an important covariate, it is likely that differences in soil moisture control soil development in this area. Landsat band ratio 5/7 is useful for distinguishing differences in geologic parent material (Inzana et al., 2003) and likely helps distinguish differences in inter-bedded geologic types. For much of the northern rolling high plains and possibly for other areas with rolling hills, curvature, potential solar radiation and geological type are likely useful for modeling soil distribution.

4.3 Spatial predictions

Spatial predictions of subgroup classes using the most accurate model visually correspond to expected soil-landscape relationships for each study area (Figs. 12A, 13A, & 14A). For the NM and WY study areas spatial predictions generally agree with published soil surveys (data not shown) although our predictions show much finer spatial detail. For the NM study area (Fig. 12A), soils with a bedrock contact (Lithic Ustic Haplocalcids) were predicted on steeply sloping uplands. Calcic Petrocalcids (subsurface cemented $CaCO_3$) were predicted on older, stable alluvial surfaces. Ustic Haplocambids (little soil development) were predicted on what are likely more active and recent geomorphic surfaces. Petronodic Ustic Haplocalcids (subsurface $CaCO_3$ concretions, possibly approaching cementation) were predicted on landforms intermediate between where Calcic Petrocalcids and Ustic Haplocambids were predicted. Ustic Haplocalcids (subsurface $CaCO_3$accumulation) were predicted to occur in an intermingled pattern with Calcic Petrocalcids and Ustic Haplocambids, but may be over-predicted on steeply sloping uplands. For the WY study area (Fig. 14A), both Ustic Torriorthents (minimal development) and Badlands (steep hills and gullies) were predicted on steeply sloping, dissected landforms near stream channels where active erosion may be occurring. Ustic Haplargids (subsurface clay accumulation) were predicted on flatter, more stable upland surfaces that likely had enough time for clay to form and/or translocate in the subsoil.

Although spatial predictions for the UT study area (Fig. 13A) must be treated with caution given the low accuracy metrics, the spatial patterns of predicted subgroup classes for the UT study area corresponded with our understanding of soil-landscape relationships. Lithic Xeric Haplocalcids (soils with a bedrock contact and subsurface accumulation of $CaCO_3$) were predicted on steeply sloping uplands. Lithic Calciargids (bedrock contact and subsurface accumulation of $CaCO_3$ and clay) were predicted on concave areas of these steeply sloping uplands where potential soil moisture accumulation is higher, resulting in greater development of subsurface clay. Rock Outcrops were predicted on the steepest mountain faces where many cliffs and talus fields were observed. Xeric Haplocalcids (subsurface $CaCO_3$) were predicted to occur on alluvial surfaces. Xeric Calciargids (subsurface $CaCO_3$ and clay) were predicted on older more stable alluvial surfaces and in some upland areas.

Spatial prediction uncertainty was generally lowest (lowest confusion index values) in relatively low relief alluvial channels and run-in areas in the NM (Fig. 12B) and UT study areas (Fig. 13B), and in lower relief portions of the WY area (Fig. 14B). This is likely because low relief areas had comparatively low covariate complexity and suggests that soil taxonomic class prediction will be least uncertain in relatively low relief areas.

5. Conclusions

This study provides insight into the use of machine learning for mapping the spatial distribution of soil taxonomic classes. We applied eleven machine learning models to three separate semi-arid study areas using three different sets of environmental covariates. Random Forests models using covariates identified by recursive feature selection were consistently the most accurate models between study areas and between covariate sets within each area. Complex models were consistently more accurate than simple or moderately complex models. We recommend that for predicting soil taxonomic classes, complex models and covariates selected by recursive feature elimination be used.

Machine learning models are most accurate when there are few soil classes and when the frequency distribution of soil pedon observations are approximately equal between classes. The number of soil subgroup classes depends on the inherent variability of each landscape. The frequency distribution of soil pedon observations depends on the sampling method. The use of cLHS results in many soil pedon observations in common soil classes and few observations in “rare” soil classes. Solutions to this problem could include increasing the number of samples in rare classes by targeted sampling or case-based reasoning. Spatial prediction uncertainty is likely to be lowest in relatively low relief areas.

Acknowledgements: This research was supported in large part by the USDI Bureau of Land Management (BLM) via Cooperative Ecosystem Study Unit (CESU) Agreement number L09AC15757 and the Utah Agricultural Experiment Station (UAES), Utah State University. Approved as UAES journal paper number 8655. We would like to thank Jeremiah Armentrout, Brook Fonnesbeck, and the Fort Bliss military base for providing the NM dataset. Brook Fonnesbeck, Mike Leno, Zamir Libohova, Shawn Nield and Amanda

Preddice collected the WY data. We would like to thank Rob Gentilion for assisting with covariate development. We would also like to thank two anonymous reviewers for their insightful comments, which substantially improved this article. Mention of any product by name does not constitute endorsement by the U.S. Geological Survey.

6. References

Barthold, F.K., Wiesmeier, M., Breuer, L., Frede, H.-G., Wu, J., Blank, F.B., 2013. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. J. Arid Environ. 88, 194–205. doi:10.1016/j.jaridenv.2012.08.004

Beaudette, D.E., O'Geen, A.T., 2009. Quantifying the aspect effect: an application of solar radiation modeling for soil survey. Soil Sci. Soc. Am. J. 73, 1345–1352. doi:10.2136/sssaj2008.0229

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168, 21–33. doi:10.1002/jpln.200421414

Behrens, T., Scholten, T., 2006. A Comparison of Data-Mining Techniques in Predictive Soil Mapping, in: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, pp. 353 – 617.

Behrens, T., Schmidt, K., Zhu, A.X., Scholten, T., 2010. The ConMap approach for terrain-based digital soil mapping. Eur. J. Soil Sci. 61, 133–143. doi:10.1111/j.1365-2389.2009.01205.x

Best, M.G., Lemmon, D.M., Morris, H.T., 1989. Geologic map of the Milford quadrangle and east half of the Frisco quadrangle, Beaver county, Utah. Miscellaneous Investigations Series Map I-1904. U.S. Geological Survey. Reston.

Bestelmeyer, B.T., Ward, J.P., Havstad, K.M., 2006. Soil-geomorphic heterogeneity governs patchy vegetation dynamics at an arid ecotone. Ecology 87, 963–973. doi:10.1890/0012-9658(2006)87[963:SHGPVD]2.0.CO;2

Boer, M., DelBarrio, G., Puigdefabregas, J., 1996. Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model. Geoderma 72, 99–118. doi:10.1016/0016-7061(96)00024-9

Boettinger, J.L., 2010. Environmental covariates for digital soil mapping in the western USA, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, Dordrecht, pp. 17–27.

Boruvka, L., Pavlu, L., Vasat, R., Penizek, V., Drabek, O., 2008. Delineating acidified soils in the jizera mountains region using fuzzy classification, in: Hartemink, A.E., McBratney, A., Mendonça-Santos, M. de L. (Eds.), Digital Soil Mapping with Limited Data. Springer, Netherlands, pp. 303–309.

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models, in: SAGA – Seconds Out (= Hamburger Beitraege zur Physischen Geographie und Landschaftsoekologie, Vol. 19). J. Boehner, T. Blaschke, L. Montanarella, pp. 23–32.

Brier, G.W., 1950. Verification of forecast expressed in terms of probability. Mon. Weather Rev. 78, 1–3. doi:http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brungard, C.B., Boettinger, J.L., 2012. Spatial prediction of biological soil crust classes; value added DSM from soil survey, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 57–60.

Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. Geoderma 111, 21–44. doi:10.1016/S0016-7061(02)00238-0

Burrough, P.A., Gaans, P.F.M. van, Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115 – 135. doi:http://dx.doi.org/10.1016/S0016-7061(97)00018-9

Campling, P., Gobin, A., Feyen, J., 2002. Logistic Modeling to spatially predict the probability of soil drainage classes. Soil Sci. Soc. Am. J. 66, 1390–1401.

Chavez Jr, P.S., 1996. Image-Based Atmospheric Corrections - Revisited and Improved. Photogramm. Eng. Remote Sens. 62, 1025–1036.

Cole, N.J., 2004. A pedogenic understanding raster classification model for mapping soils, Powder River Basin, Wyoming (MS Thesis). Utah State University, Logan, USA.

Cole, N.J., Boettinger, J.L., 2006. Pedogenic understanding raster classification methodology for mapping soils, Powder River Basin, Wyoming, USA, in: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, pp. 377–619.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46. doi:10.1016/0034-4257(91)90048-B

Congalton, R.G., Green, K., 1998. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. CRC/Taylor & Francis, Boca Raton.

Crist, E., Kauth, R., 1986. The tasseled cap de-mystified. Photogramm. Eng. Remote Sens. 52, 81–86.

Duniway, M.C., Herrick, J.E., Monger, H.C., 2010. Spatial and temporal variability of plant-available water in calcium carbonate-cemented soils and consequences for arid ecosystem resilience. Oecologia 163, 215–226. doi:10.1007/s00442-009-1530-7

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39. doi:10.1029/2002WR001426

Gile, L.H., Hawley, J.W., Grossman, R.B., 1981. Memoir 39—Soils and geomorphology in the Basin and Range area of southern New Mexico: Guidebook to the Desert Project. New Mexico Bureau of Mines and Mineral Resources, Soccoro.

Goslee, S.C., 2011. Analyzing remote sensing data in R: the landsat package. J. Stat. Softw. 43, 1–25.

Green, G.N., Drouillard, P.H., 1994. The Digital Geologic Map of Wyoming in ARC/INFO Format. U.S. Geological Survey Open-File Report 94-0425. http://geo-nsdi.er.usgs.gov/metadata/open-file/94-425/ (last accessed: 7/5/2014).

Green, G.N., Jones, G.E., 1997. The Digital Geologic Map of New Mexico in ARC/INFO Format. U.S. Geological Survey Open File Report 97-0052. http://pubs.usgs.gov/of/1997/ofr-97-0052/new_mex.htm (last accessed: 7/5/2014).

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143, 180–190. doi:10.1016/j.geoderma.2007.11.004

Grunwald, S., 2006. Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics. CRC/Taylor & Francis, Boca Raton.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422. doi:10.1023/A:1012487302797

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398. doi:10.1016/j.geoderma.2004.06.007

Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma 140, 417–427. doi:10.1016/j.geoderma.2007.04.022

Hengl, T., Reuter, H.I., 2008. Geomorphometry. Concepts, Software, Applications. Developments in Soil Science. Elsevier, Amsterdam.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214–215, 141–154. doi:10.1016/j.geoderma.2013.09.016

Heute, A.R., 1988. A Soil-Adjusted Vegetation Index (SAVI). Remote Sens. Environ. 25, 295–309.

Inzana, J., Kusky, T., Higgs, G., Tucker, R., 2003. Supervised classifications of Landsat TM band ratio images and Landsat. J. Afr. Earth Sci. 37, 59–72. doi:10.1016/S0899-5362(03)00071-X

Jafari, A., Finke, P.A., Van de Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Eur. J. Soil Sci. 63, 284–298. doi:10.1111/j.1365-2389.2012.01425.x

Jafari, A., Ayoubi, S., Khademi, H., Finke, P.A., Toomanian, N., 2013. Selection of a taxonomic level for soil mapping using diversity and map purity indices: a case study from an Iranian arid region. Geomorphology 201, 86–97. doi:10.1016/j.geomorph.2013.06.010

James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. An Introduction to Statistical Learning: with Applications in R. Springer, New York.

Jensen, J.R., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective. Pearson Prentice Hall, Upper Saddle River.

Johansson, U., König, R., Niklasson, L., 2010. Genetic rule extraction optimizing brier score, in: Pelikan, M., Branke, J. (Eds.), GECCO. ACM, pp. 1007–1014. http://bada.hb.se/bitstream/2320/6795/1/Gecco2010_GREX_Optimizing_Brier_Score.pdf

Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012. Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. Geoderma 189, 540–553. doi:10.1016/j.geoderma.2012.05.028

Kienast-Brown, S., Boettinger, J.L., 2010. Applying the optimum index factor to multiple data types in soil survey, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, Dordrecht, pp. 385–398.

Kim, J., Grunwald, S., Rivero, R.G., Robbins, R., 2012. Multi-scale modeling of soil series using remote sensing in a wetland ecosystem. Soil Sci. Soc. Am. J. 76, 2327–2341. doi:10.2136/sssaj2012.0043

Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154, 340–347. doi:10.1016/j.geoderma.2009.11.005

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Kuhn, M. Wing, J., Weston, S., Williams, A., Keefer, C., A. Engelhardt. 2012. caret: Classification and Regression Training. R package version 5.17-7. http://CRAN.R-project.org/package=caret

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York.

Kuhn, M., 2014. A short introduction to the caret package. http://cran.r-project.org/web/packages/caret/vignettes/caret.pdf (last accessed: 7/5/2014).

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Liu, J., Pattey, E., Nolin, M.C., Miller, J.R., Ka, O., 2008. Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. Geoderma 143, 261 – 272. doi:http://dx.doi.org/10.1016/j.geoderma.2007.11.011

Lowry, J., Ramsey, R.D., Thomas, K., Schrupp, D., Sajwaj, T., Kirby, J., Waller, E., Schrader, S., Falzarano, S., Langs, L., Manis, G., Wallace, C., Schulz, K., Comer, P., Pohs, K., Rieth, W., Velasquez, C., Wolk, B., Kepner, W., Boykin, K., O'Brien, L., Bradford, D., Thompson, B., Prior-Magee, J., 2007. Mapping moderate-scale land-cover over very large geographic areas within a collaborative framework: a case study of the Southwest Regional Gap Analysis Project (SWReGAP). Remote Sens. Environ. 108, 59 – 73. doi:http://dx.doi.org/10.1016/j.rse.2006.11.008

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232–234, 34–44. doi:10.1016/j.geoderma.2014.04.033

Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., Francaviglia, R., 2011. Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data. CATENA 85, 267–273. doi:10.1016/j.catena.2011.01.012

McBratney, A.B., Mendonça Santos, M.D.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. doi:10.1016/S0016-7061(03)00223-4

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388. doi:10.1016/j.cageo.2005.12.009

Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285–293. doi:10.1016/j.geoderma.2007.08.022

Moonjun, R., Farshad, A., Shrestha, D.P., Vaiphasa, C., 2010. Artificial Neural Network and Decision Tree in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Springer, Dordrecht, pp. 151–164.

Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma 163, 38–44. doi:10.1016/j.geoderma.2011.03.014

Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 232–234, 97–106. doi:10.1016/j.geoderma.2014.04.036

Peters, J., Baets, B.D., Verhoest, N.E.C., Samson, R., Degroeve, S., Becker, P.D., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207, 304–318. doi:10.1016/j.ecolmodel.2007.05.011

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209–210, 1–14. doi:10.1016/j.geoderma.2013.05.029

R Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D. (Eds.), 2003. Field Book for Describing and Sampling Soils, Version 2.0. Natural Resources Conservation Service. National Soil Survey Center, Lincoln.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model. 181, 1–15. doi:10.1016/j.ecolmodel.2004.06.036

Shi, X., Long, R., Dekett, R., Philippe, J., 2009. Integrating different types of knowledge for digital soil mapping. Soil Sci. Soc. Am. J. 73, 1682. doi:10.2136/sssaj2007.0158

Shi, X., 2013. ArcSIE user's guide. http://www.arcsie.com/Download/ArcSIE_UsersGuide_130319.pdf (last accessed: 7/5/2014).

Smith, C.A.S., Daneshfar, B., Frank, G., 2012. Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 215–220.

Soil Survey Staff, 1999. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys, 2nd ed. U.S. Department of Agriculture Handbook 436. Natural Resources Conservation Service, Lincoln.

Stum, A.K., Boettinger, J.L., White, M.A., Ramsey, R.D., 2010. Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, Springer, Dordrecht, pp. 179–190.

Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio county soil map. Soil Sci. Soc. Am. J. 77, 1254–1268. doi:10.2136/sssaj2012.0321

Taghizadeh-Mehrjard, R., Minasny, B., McBratney, A.B., Triantafilis, J., Sarmadian, F., Toomanian, N., 2012. Digital soil mapping of soil classes using decision trees in central Iran, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 197–202.

Tarboton, D., 2013. Terrain Analysis Using Digital Elevation Models (TauDEM). http://hydrology.usu.edu/taudem/taudem5/index.html (last accessed: 7/5/2014).

Taylor, J.A., Jacob, F., Galleguillos, M., Prévot, L., Guix, N., Lagacherie, P., 2013. The utility of remotely-sensed vegetative and terrain covariates at different spatial resolutions in modelling soil and watertable depth (for digital soil mapping). Geoderma 193–194, 83–93. doi:10.1016/j.geoderma.2012.09.009

Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil depth from topographic and land cover attributes. Water Resour. Res. 45, W10438. doi:10.1029/2008WR007474

Triantifilis, J., Earl, N.Y., Gibbs, I.D., 2012. Digital soil-class mapping across the Edgeroi district usings numerical clustering and gamma-ray spectrometry data, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 187–191.

United States Department of Agriculture, 2006. Land Resource Regions and Major Land Resource Areas of the United States, the Caribbean, and the Pacific Basin. ftp://ftp-fc.sc.egov.usda.gov/NSSC/Ag_Handbook_296/Handbook_296_low.pdf (last accessed: 7/5/2014).

Utah Automated Geographic Reference Center, 2013. 5 Meter Auto-Correlated Elevation Models. http://gis.utah.gov/data/elevation-terrain-data/5-meter-auto-correlated-elevation-models/ (last accessed: 7/5/2014).

Van Zijl, G.M., le Roux, P.A.L., Smith, A.B., 2012. Rapid soil mapping under restrictive conditions in Tete, Mozambique, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 335–339.

Vasques, G.M., Grunwald, S., Myers, D.B., 2012. Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. Landsc. Ecol. 27, 355–367. doi:10.1007/s10980-011-9702-3

Webster, R., Welham, S.J., Potts, J.M., Oliver, M.A., 2006. Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. Comput. Geosci. 32, 1320–1333. doi:10.1016/j.cageo.2005.12.002

Western Regional Climate Center, 2013. Cooperative Climatological Data Summaries. http://www.wrcc.dri.edu/climatedata/climsum/ (last accessed: 7/5/2014).

Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, New York.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington.

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2012. Which soil, environmental and anthropogenic covariates for soil carbon models in Florida are needed? In: Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 335–339.

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environ. Model. Softw. 57, 202–215. doi:10.1016/j.envsoft.2014.03.004

Yokoyama, R., Shirasawa, M., Pike, R.J., 2002. Visualizing topography by openness: a new application of image processing to digital elevation models. Photogramm. Eng. Remote Sens. 68, 257–266.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65, 1463–1472.

Figure Captions:

Fig. 1. Study area locations in western USA.

Fig. 2. Spatial distribution of pedon observation locations in the NM study area overlain on google physical map. Total number of pedon observations was 103.

Fig. 3. Spatial distribution of pedon observation locations in the UT study area overlain on google physical map. Total number of pedon observations was 297.

Fig. 4. Spatial distribution of pedon observation locations in the WY study area overlain on google physical map. Total number of pedon observations was 57.

Fig. 5. Optimal covariate subset selection using recursive feature elimination. Out-of-bag (OOB) error is random forests misclassification error. Random forests models were begun with the total available covariates and the least important covariate was iteratively removed. Optimal covariate subsets were selected as those covariates that returned the lowest OOB error and which had the fewest covariates. Arrows indicate optimal covariate subset.

Fig. 6. Average κ for the NM study area. Model with highest κ is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer-Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 7. Average Brier scores for the NM study area. Model with lowest Brier score is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 8. Average κ for the UT study area. Model with highest κ is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 9. Average Brier scores for the UT study area. Model with lowest Brier score is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 10. Average κ for the WY study area. Model with highest κ is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 11. Average Brier scores for the WY study area. Model with lowest Brier score is the most accurate classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

Fig. 12. Spatial predictions of subgroup classes (A), and the confusion index (B) for the NM study area using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown (5 of 10 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions; values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.

Fig. 13. Spatial predictions of subgroup classes (A), and the confusion index (B) for the UT study area using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown (5 of 15 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions; values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.

Fig. 14. Spatial predictions of subgroup classes (A), and the confusion index (B) for the WY study area using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown (3 of 5 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions; values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.