Published April 20, 2017

Neighborhood Size of Training Data Influences Soil Map Disaggregation

Matthew R. Levi*
USDA-ARS
Jornada Experimental Range
MSC 3JER Box 30003
New Mexico State Univ.
Las Cruces, NM 88003

Soil class mapping relies on the ability of sample locations to represent portions of the landscape with similar soil types; however, most digital soil mapping (DSM) approaches intersect sample locations with one raster pixel per covariate layer regardless of pixel size. This approach does not take the variability of covariate information adjacent to the training data into account. The objective here was to disaggregate a soil map in a semiarid Arizona rangeland (78,569 ha) by exploring different neighborhood sizes for extracting covariate data to points. Eight machine learning algorithms were compared to assess the influence of summarizing covariate data in 0-, 15-, 30-, 60-, 90-, 120-, 150-, and 180-m circular neighborhoods and a multiscale model. K values of all models ranged between 0.24 and 0.44 and increased with neighborhood size up to 150 m. Support vector machine and random forest algorithms performed best across all scales. The radial support vector machine model using a 150-m neighborhood had the highest K and produced a more generalized map compared with the best multiscale model (random forest), which resulted in a mix of general and detailed soil features. Evaluating a range of neighborhood sizes for aggregating covariate data provides a method of accounting for multiscale processes that are important for predicting soil patterns without modifying the pixel size of the final maps. Incorporating concepts from traditional soil surveys with DSM approaches can strengthen ties between them and optimize the extraction of landscape information for predicting soil properties.

Abbreviations: cLHS, conditioned Latin hypercube sample; DEM, digital elevation model; DSM, digital soil mapping; iPCA, iterative principal component analysis; NDVI, normalized difference vegetation index; RF.multiscale, random forest multiscale model; RFE, recursive feature elimination; SVMR.150, the support vector machine model with a radial kernel and a 150-m neighborhood.

Core Ideas

- **Focal summaries of covariate data around sampling points affect model performance.**
- **Support vector machine and random forest approaches produced the best results.**
- **A 150-m neighborhood emerged as the best model, albeit with a general soil map.**
- **Multiscale covariate data reflect realistic patterns of soil–landscape features.**

Most DSM approaches that use point data to develop relationships with environmental covariate data intersect sample locations with one raster pixel regardless of pixel size. An alternative approach is to extract covariate data for discernable landscape units surrounding sample locations with similar soils. Traditional soil surveys are conducted with the awareness that a given soil sample location often represents a larger portion of the landscape with similar soil (i.e., the soil–landscape paradigm; Hudson, 1992). One reason for this phenomenon is that soil forming environments can be explained by Tobler's first law of geography that "near things are more related than distant things" (Tobler, 1970). Therefore, it makes sense that incorporating information surrounding the sample points (i.e., neighborhood information) is appropriate for modeling soil classes because spatially adjacent classes are likely to be similar. One method of representing soil bodies is to use a circular neighborhood around sampling points to aggregate covariate data (Behrens et al., 2010a; Grinand et al., 2008; Silva et al., 2016), with the assumption that the size of the neighborhood will reflect a similar soil-forming environment. This study explored the effect of summarizing covariate data in different neighbor-

This article has supplemental material.
Soil Sci. Soc. Am. J. 81:354–368
doi:10.2136/sssaj2016.08.0258
Received 16 Sept. 2016.
Accepted 21 Dec. 2016.
*Corresponding author (matthew.levi@ars.usda.gov)

hood sizes on model performance for disaggregating a soil map with multi-component map units in a semiarid rangeland environment.

Significant effort has been dedicated to discerning unique soil types within multi-component soil map units in many soil-forming environments (Häring et al., 2012; Nauman et al., 2014; Odgers et al., 2014; Rad et al., 2014; Scull et al., 2005). Knowing the geographic locations of soil components within a soil map unit is useful for many management questions and is correlated to numerous interpretations (Soil Survey Staff, 1993). Recent studies have focused on various methods of disaggregating soil map units, including the use of newly collected point data (Brungard et al., 2015; Rad et al., 2014) and legacy data from previous soil survey efforts or other projects (Bui et al., 2006; Kempen et al., 2009; Heung et al., 2016). Other methods have focused on the embedded information from soil surveys to develop probability-based samples of soil types within map units to discern relationships between covariate data and soil types (Nauman and Thompson, 2014; Odgers et al., 2014). Although the majority of DSM efforts predict soil properties analyzed in a laboratory setting, recent efforts have shown that field descriptions provide valuable information that can be used to develop soil prediction maps (Balkovic et al., 2013; Hengl et al., 2007). Although positional error is an important consideration when using field soil profile descriptions, it is often much smaller than the error associated with covariates and model performance (Nelson et al., 2011). Many soil survey areas have archived soil information from previous survey efforts available in the local and regional offices of federal agencies that are rarely used after a survey is completed. In some cases, transect data are archived in national databases (e.g., the USDA-NRCS National Soils Information System). These data offer a wealth of information that can be incorporated with contemporary sampling schemes to facilitate improved soil map products.

Advances in data availability and computation power allow for the integration of local, regional, and supraregional landscape characteristics that are important for predicting soil properties (Behrens et al., 2014). Fine-scale environmental covariates are not always the best predictors for soil properties (Samuel-Rosa et al., 2015); hence, a range of pixel sizes and analysis neighborhoods should be explored to optimize models for a given region (Cavazzi et al., 2013; Nauman et al., 2014; Roecker and Thompson, 2010). In a given area, different soil classes may require different scales of the same attributes to predict their spatial distribution accurately (Behrens et al., 2010b). For example, a range of analysis scales for topographic attributes may be necessary to reflect the mental models used by soil scientists for mapping soils (Miller, 2014). Maynard and Johnson (2014) explored the role of raster pixel size compared with neighborhood extent for predicting soil properties in the Oregon Coast Range Mountains and concluded that a moderate pixel size (10 m) with a range of neighborhood extents was optimal. Manipulating the neighborhood extent around a moderate pixel size can incorporate multiscale information without losing potentially valuable detail by increasing the pixel size; however, optimizing the neighborhood size for any given soil class remains a formidable challenge.

Soil prediction models that use multiscale predictors can outperform models developed from only one analysis scale (Miller et al., 2015). Scale can easily be incorporated into topographic analyses because the analysis window can be modified to account for different shapes and sizes of neighborhoods. Conversely, reflectance data are more static with respect to scale and require other methods to aggregate information from neighboring pixels. In practice, multiscale information is commonly incorporated into DSM by manipulating the pixel size or analysis extent of topographic variables (Behrens et al., 2010b, Maynard and Johnson, 2014), whereas reflectance indices are more commonly modified by changing the pixel size (Vasques et al., 2012) or incorporating sensors with a different pixel size (Miller et al., 2015). An alternative approach to scaling covariate information is to aggregate the covariate data using existing soil map unit boundaries or to average covariate data within a focal neighborhood around the training points (e.g., a circular neighborhood) (Silva et al., 2016). Representing multiscale patterns and processes requires the integration of potentially large numbers of environmental covariates, which necessitates robust models for predictions.

Machine learning algorithms offer many benefits for predicting soil classes. For example, they are robust models that are resistant to overfitting and can accommodate both continuous and categorical predictor variables (Kuhn and Johnson, 2013; Olden et al., 2008). As such, they have become very popular in the DSM community. Heung et al. (2016) categorized machine learning approaches for predicting soil types into tree-based learners, logistic regression, logistic model trees, distance-based learners, artificial neural networks, and support vector machines. Brungard et al. (2015) further organized some commonly used machine learning algorithms into simple, moderate, and complex on the basis of model interpretability and the number of parameters required for model tuning. It is often helpful to compare multiple approaches to take advantage of the strengths of different methods (Brungard et al., 2015; Heung et al., 2016; Taghizadeh-Mehrjardi et al., 2015).

There is no standard method of reporting accuracy metrics in DSM, which complicates the comparisons of model performance between or among studies. For example, studies predicting soil classes have used K (Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015), entropy (Jafari et al., 2013; Kempen et al., 2009), soil richness (Jafari et al., 2013), percent correct (Nauman et al., 2014), and out-of-bag error (Rad et al., 2014) for overall assessments of model performance. Some studies also reported measures of accuracy for individual classes such as percent correctly classified (Hengl et al., 2007; Nauman et al., 2014), out-of-bag error (Rad et al., 2014), user's and producer's accuracy (Taghizadeh-Mehrjardi et al., 2015), or purity (Jafari et al., 2012). Furthermore, some studies report accuracy metrics for independent validation datasets (Häring et al., 2012; Hengl et al., 2007; Heung et al., 2016), whereas others do not use independent validation points and report performance metrics for cross-validation or other resampling techniques (Brungard et al., 2015; Jafari et al., 2013; Rad et al., 2014). Although relative differences between some metrics could

be inferred, lack of a common metric for reporting classification success limits comparisons between or among studies.

Accuracy assessments in classification models are generally based on exact matches of the measured and modeled points. However, soils with different names often have very similar characteristics, which is likely to lead to underestimation of model performance for predicting soil taxonomic groups. For this reason, some researchers have incorporated taxonomic distance into the evaluation routine for such models (Carré and Girard, 2002; Minasny and McBratney, 2007). This is important because direct matches of soil type may occur with second or third highest probabilities for a given point (Odgers et al., 2014). Although incorporating soil taxonomic distance into model assessments can provide more realistic evaluations of soil class predictions, it does not always improve model accuracy (Taghizadeh-Mehrjardi et al., 2015). Furthermore, this approach is not always used because taxonomic distances between soils are not readily available (Brungard et al., 2015) or an alternative method such as rank matching is used (Chaney et al., 2016; Odgers et al., 2014). Another method of accuracy assessment is to aggregate predicted values within a neighborhood surrounding the validation points (Nauman and Thompson, 2014). Although methods exist to modify classification accuracy metrics to improve interpretations and facilitate model improvement, the best measure of reality is an exact match for a given class.

The objective of this study was to create a series-level digital soil map in a semiarid landscape in southeastern Arizona without using existing soil map unit boundaries. This study evaluated the ability of point data to produce soil maps for comparison with an existing soil survey. More specifically, this study compared variable neighborhood sizes for summarizing covariate data around training point locations using eight different machine learning algorithms.

MATERIALS AND METHODS

Study Area

The focus of this study was a 78,569 ha site in Southeastern Arizona's Basin and Range Province in Cochise County, AZ (Fig. 1). It is situated at the border with Mexico and New Mexico. Elevations range from 1220 m in the valley bottom to over 2240 m in the adjacent Peloncillo Mountains. The majority of the study area comprises semidesert grassland with a transition to Madrean evergreen woodland in the Peloncillo Mountains (Brown and Lowe, 1994). Drainages converge at the San Bernardino National Wildlife Refuge before flowing south into Mexico. This area occupies the transition zone between the Sonoran and Chihuahuan Deserts, which differ in their annual precipitation regimes and dominant plant communities (Brown, 1994; Neilson, 1987). Common plant functional groups found in the semidesert grassland communities include grasses, forbs, shrubs, leaf succulents, and cacti (Brown, 1994). The climate is semiarid with 361 to 470 mm of mean annual precipitation and a mean annual temperature ranging from 13 to 18°C (PRISM Climate Group, 2012). The primary land use is cattle ranching.

There are a variety of soil orders within the study area, including Entisols, Aridisols, Vertisols, and Mollisols, formed in predominantly volcanic parent material with some granite, limestone, and associated basin fill (Soil Survey Staff, 2013). The published soil survey identified 48 unique named soil series represented in 36 soil map units in the study area (Supplemental Table S1; Soil Survey Staff, 2013). The San Bernardino Volcanic field occupies a considerable portion of the study area, resulting in landforms such as basalt flows and cinder cones formed during the Pleistocene (Biggs et al., 1999). Other common landforms in the area include alluvial fans and pediment fans. The soil survey covered 72,590 ha of the larger study area, leaving unmapped areas on national forest land.

Environmental Covariate Data

A total of 36 environmental covariates were derived from a digital elevation model (DEM) and Landsat reflectance (Table 1). Topographic variables are a critical source of information for soil prediction models (Brungard et al., 2015; Chaney et al., 2016; Jafari et al., 2013), so a suite of 12 variables was derived from a one-third arc-second National Elevation Dataset (http://nationalmap.gov/, accessed 16 Mar. 2017) DEM with 10-m spatial resolution using the SAGA graphical information system software (Conrad et al., 2015). Prior to analyses in SAGA, the DEM was preprocessed to a hydrologically correct surface using ArcGIS version 10.4 by filling sinks (Environmental Systems Research Institute, 2015). Two Landsat 8 Operational Land Imager scenes with standard terrain correction were acquired from the USGS LandsatLook Viewer (http://landsatlook.usgs.gov/viewer.html, accessed 16 Mar. 2017) that represented peak (1 Sept. 2013) and nonpeak (29 May 2013) vegetative conditions. Eight indices representing soil and geology were derived for both Landsat 8 scenes (Table 1). In addition, a 3-yr time series of normalized difference vegetation indices (NDVIs) from Landsat 5 Thematic Mapper representing all cloud-free scenes for Path 34, Row 38 from 2009 to 2011 (n = 31) was obtained from the USGS Earth Resources Observation and Science Center Science Processing Architecture on-demand interface (https://espa.cr.usgs.gov/, accessed 16 Mar. 2017). The time series represented a range of wet and dry conditions and was compressed with a principal component analysis using ArcGIS. Multitemporal vegetation indices can capture subtle differences in soil properties that are not discernable from single images (Maynard and Levi, 2017) and principal component analyses of NDVI time series are useful for deriving ecological units based upon vegetation dynamics in this region (Forzieri et al., 2011). The first seven principal components were used to represent vegetation dynamics for soil series predictions, as visual inspection suggested they were free of noise and represented landscape patterns.

Soil Profile Data

Soil pedon descriptions and field maps from initial soil survey efforts from 1994 to 1997 were collected from the Tucson, AZ, Soil Survey office of NRCS. This included scanned copies of topographic maps that had been annotated at the time of the initial soil survey and field data sheets representing soil profile descriptions and taxonomic information. Scanned topographic

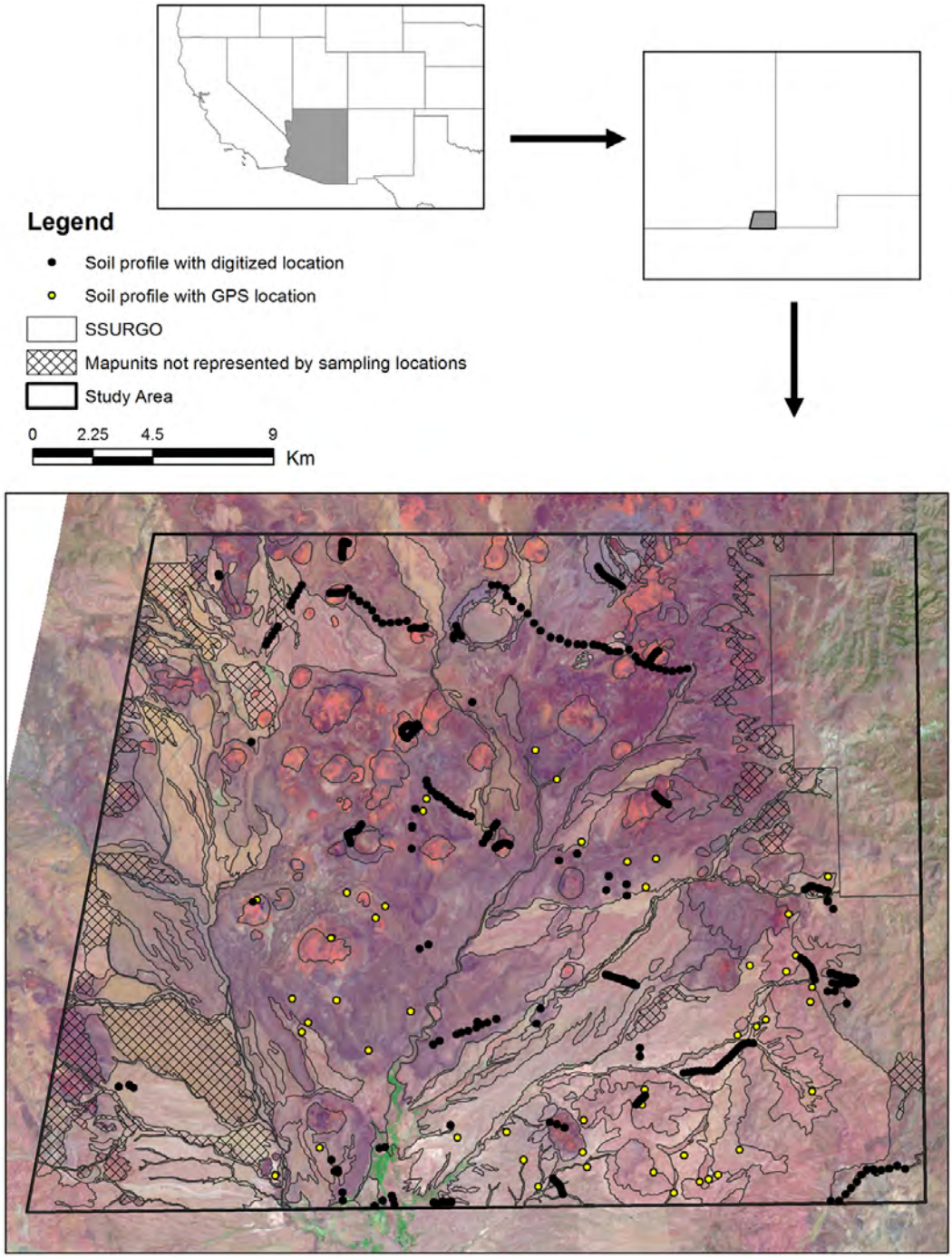

Fig. 1. Study area in southeastern Arizona showing 418 sample locations with soil series information and Soil Survey Geographic Database (SSURGO) map unit boundaries. Cross-hatched map units indicate that at least one named soil component was not present in the training data. The background is a Landsat 8 image from 29 May 2013, represented as a false color composite of Bands 7, 5, 3 in red, blue, and green at 25% transparency.

maps were orthorectified using ArcGIS. Soil profile descriptions that included GPS coordinates were converted into a shapefile. The remaining soil profile locations and sampling transect lines were hand-digitized as points and lines, respectively. Soil transect lines were broken into equal interval points based on the number of soil descriptions included in the matching transect documentation. Sampling intervals were comparable with the estimated distances reported in the soil transect documentation. The direction of the soil transects was not explicit for all transects, so annotations in the descriptions were used to determine the direction for the proper attribution of the assumed sampling locations with extensive guidance from a soil scientist who mapped much of the study area. Soil sampling points were assigned series names from profile data. Only sample locations with sufficient data to assign a soil series name were used in the modeling.

In addition to using existing soil pedon data, 50 new observations were made to represent the variability of the soils for a 19,342 ha portion of the study area. To capture the spatial patterns of unique soil–landscape features, an iterative principal component analysis (iPCA) data reduction routine was used to identify

Table 1. Environmental covariates derived from 10 m DEM and 30 m Landsat that was resampled to 10 m for modeling.†

Covariate	Source	Used in cLHS‡	Reference
Elevation	NED		
Slope	NED	X	
SAGA wetness index	NED		(Boehner et al., 2002; Freeman, 1991)
Topographic position index 100 pixel radius	NED	X	(Wilson and Gallant, 2000)
Topographic position index 250 pixel radius	NED		(Wilson and Gallant, 2000)
Topographic position index 500 pixel radius	NED		(Wilson and Gallant, 2000)
Topographic position index 1000 pixel radius	NED	X	(Wilson and Gallant, 2000)
Multiresolution index of river bottom flatness	NED	X	(Gallant and Dowling, 2003)
Multiresolution index of ridge top flatness	NED		(Gallant and Dowling, 2003)
Valley depth	NED		
Midslope position	NED		
Minimum curvature	NED		
Maximum curvature	NED		
NDVI time series_PCA_1	Landsat 5 TM		
NDVI time series_PCA_2	Landsat 5 TM		
NDVI time series_PCA_3	Landsat 5 TM		
NDVI time series_PCA_4	Landsat 5 TM		
NDVI time series_PCA_5	Landsat 5 TM		
NDVI time series_PCA_6	Landsat 5 TM		
NDVI time series_PCA_7	Landsat 5 TM		
May and September band ratio 3:1	Landsat 8 OLI	X (Sept.)	
May and September band ratio 3:2	Landsat 8 OLI		(Boettinger et al., 2008)
May and September band ratio 5:4	Landsat 8 OLI		
May and September band ratio 7:3	Landsat 8 OLI		(Boettinger et al., 2008)
May and September band ratio 7:5	Landsat 8 OLI	X (May)	(Boettinger et al., 2008)
May and September calcareous sediment index	Landsat 8 OLI		(Boettinger et al., 2008)
May and September gypsic index	Landsat 8 OLI		(Nield et al., 2007)
May and September natric index	Landsat 8 OLI		(Nield et al., 2007)

†NDVI, normalized difference vegetation index, PCA, principal component analysis; NED,National Elevation Dataset; Landsat 5 TM, Landsat Thematic Mapper sensor; Landsat 8 OLI, Operational Land Imager sensor.

‡ X denotes covariate data that were selected via iterative principal component analysis data reduction and used as input for a conditioned Latin hypercube sample (cLHS) design of 50 newly collected pedons.

the topographic and reflectance indices that explained the greatest amount of information in the covariate data by using custom code in R (Levi and Rasmussen, 2014; R Core Team, 2014). At the onset of this project, 27 of the 36 covariates described above were thought to be satisfactory for representing the landscape variability for the iPCA and subsequent sample design. After new samples were collected, the availability of additional soil pedon data from initial soil survey efforts (described above) became available, which expanded the study area. A total of 27 covariates were used for iPCA, including eight topographic variables: percent slope, SAGA wetness index, tangential curvature, multiresolution index of river bottom flatness, and topographic position index calculated with four different neighborhoods (100, 250, 500, and 1000 pixels). I derived 19 reflectance indices (nine from each Landsat 8 scene) and a difference in NDVI between September and May. The indices included the following: gypsic index, natric index, calcareous sediment index, NDVI, and the band ratios 3:1, 3:2, 5:4, 7:3, and 7:5. The final six covariates selected via the iPCA (Table 1) were applied to a conditioned Latin hypercube sample (cLHS) design to identify 50 sampling locations on ranches where access was available. The cLHS design is a stratified random technique that can represent the multivariate distributions of covariate data (Minasny and McBratney, 2006). Sample design was carried out using the 'clhs' package in R (Roudier, 2011).

Sampling of the cLHS points occurred between June 2014 and March 2015. Hand-dug soil pits were described and sampled according to the genetic soil horizon to a limiting layer (e.g., petrocalcic, bedrock) or as deep as possible according to National Cooperative Soil Survey standards. Samples were collected from the field for particle size analysis. Field descriptions and laboratory data were used to classify each pedon to the soil series level.

Modeling Soil Series

Models were developed for soil series classes that had at least two samples in the training data. Weighting the model training according to the frequency distribution of soil classes can improve the accuracy of minority classes; however, the overall accuracy may be reduced (Stum et al., 2010). Similar soils can be merged prior to modeling to reduce the number of classes and therefore create less imbalance in the training data (Rad et al., 2014); however, the goal of this work was to disaggregate the multi-component soil map units, so I wanted to predict as many unique classes as possible. Removing classes with small sample sizes can improve model performance by helping to balance the

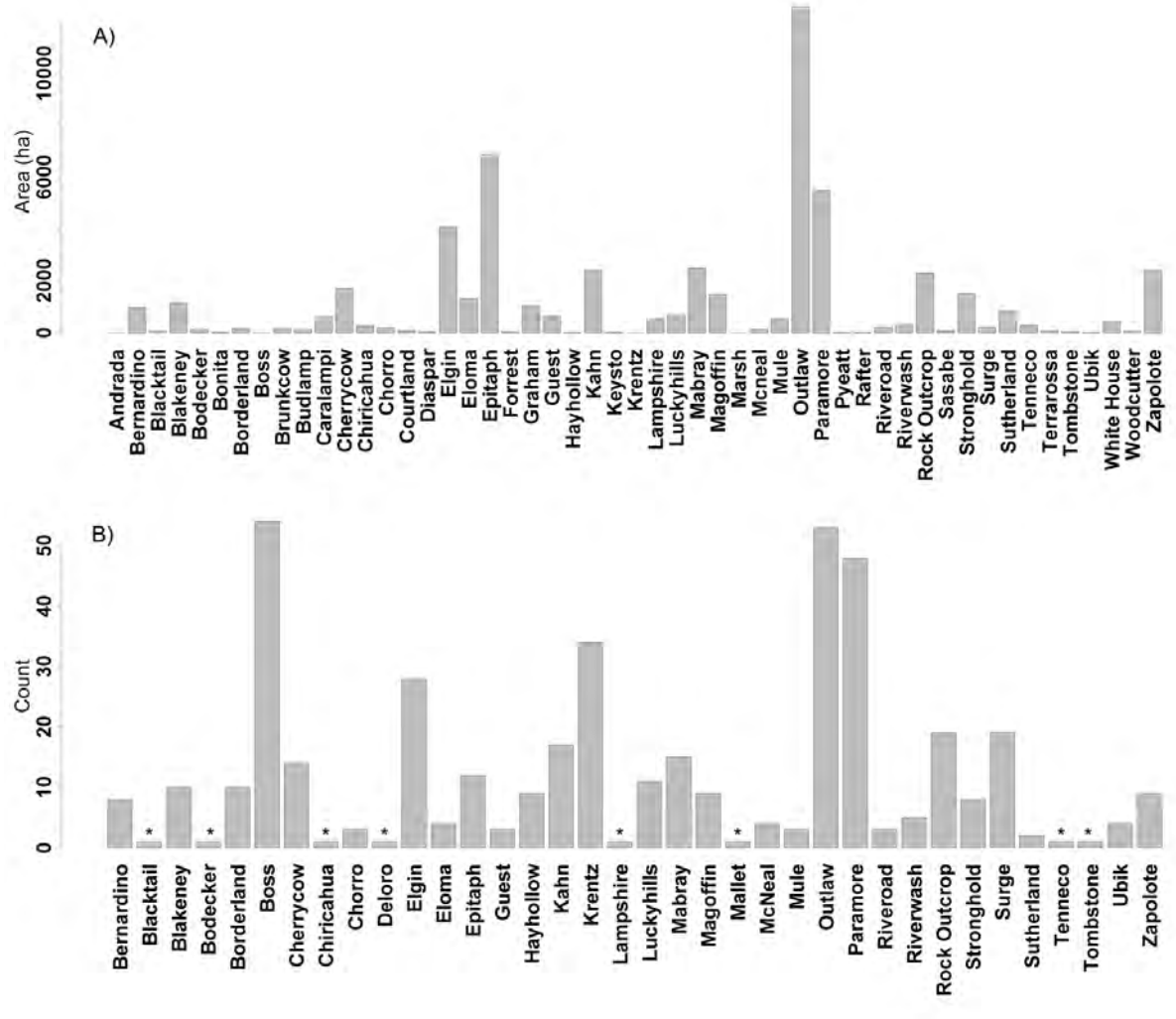

Fig. 2. Area of named soil series represented in published Soil Survey Geographic Database (SSURGO) map units within the study area (A) and soil series frequency of available pedon data (*n* = 426) (B). An asterisk indicates that the series was not included in prediction models because the sample size was one. The area of a named soil series does not include unnamed inclusions in published map units. The area of each soil series in the study area was approximated by multiplying the proportion of each component in a given map unit by the total mapped area of that map unit. For complex map units that did not explicitly state the proportion of each component, the named components were assumed to occupy 90% of the map unit, leaving 10% as unnamed inclusions. In these cases, the named components were assigned equal proportions of the 90% to calculate the area.

class distributions (Kovačević et al., 2010). In order to preserve the detailed soil series level information, I elected to remove minor classes with only one observation as the method of dealing with class imbalance. A total of 433 sample points had series-level classifications; however, after intersecting these with covariate data and removing samples with fewer than two training points in a given class, there were 418 sample points available for modeling. Those points represented 28 soil classes. Most of the samples represented dominant soil series and several series had fewer than five samples (Fig. 2). Soil map units that had at least one named component that was not represented in the 418 sample points were positioned along the edges of the study area and accounted for 8% of the mapped area (Fig. 1). Of the 418 points used for modeling, 53 had explicit GPS coordinates assigned to them and the remaining 365 points were assigned spatial locations using the hand-digitizing procedure described above. The 418 pedons represented the general area distribution of soil series in the published soil survey (Fig. 2).

Eight machine learning models were implemented using the 'caret' package (Kuhn, 2008) in R (R Core Team, 2014) and represented simple (*k*-nearest neighbors), moderate (classification trees and nearest shrunken centroids), and complex algorithms (bagged classification trees, random forests, linear support vector machines, radial basis support vector machines, and the decision tree and rule-based algorithm C5.0) according to Brungard et al. (2015). Detailed descriptions of all models are provided by Kuhn and Johnson (2013). For models that required selection of tuning parameters, the default search grid produced with the 'caret' package was used to select the parameters that produced the simplest model within one SE of the best model (James et al., 2014). The focal mean of each environmental covariate presented in Table 1 was generated for circular neighborhoods with radii of 0, 15, 30, 60, 90, 120, 150 and 180 m. The resulting covariates were extracted to soil sample locations and models were developed for each of the respective neighborhood sizes. A multiscale model including covariates from all neighborhoods (288 covariates) was

also tested. Recursive feature elimination (RFE), which is a backward selection algorithm, was applied to reduce the number of predictor variables prior to modeling for each neighborhood size (Guyon et al., 2002). This approach has been used in other DSM studies to identify the variables that are important for model development (Ballabio, 2009; Brungard et al., 2015). Recursive feature elimination was used using the random forest classifier with five repeated cross-validation routines with 10 folds each. Data were centered and scaled prior to RFE and subsequent modeling, and each model was run using the same seed for random number generation to ensure the models were comparable.

Model Validation

A leave-group-out cross-validation was used to assess model performance and compare results where 70% of points were used for training and 30% for validation (Kuhn and Johnson, 2013). The average value of the K statistic from the 100 iterations was used for overall model assessment and comparison between models. K provides a measure of overall model performance that is corrected for chance agreement (Foody, 2002). Values of K between 0 and 0.2 are considered slight, those between 0.2 and 0.4 are fair, those between 0.4 and 0.6 are moderate, those between 0.6 and 0.8 are substantial and those >0.8 are in near-perfect agreement (Landis and Koch, 1977). In this study, K was used as the preferred metric for overall model performance to allow comparisons to other studies that reported K. Individual class accuracies were evaluated via user's and producer's accuracies (Congalton, 1991). User's accuracy is the number of correctly classified pixels for a given class divided by the total number of pixels classified as that class (i.e., the reliability of the map). Producer's accuracy is the number of correctly classified pixels divided by the total number of reference pixels for the same class.

A confusion index (*CI*) was calculated for spatial assessment of model performance (Burrough et al., 1997):

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

where $\mu_{\max,i}$ is the probability value of the class with maximum probability at location i and $\mu_{(\max-1)i}$ is the second-largest probability at location i. The value of the confusion index ranges from 0 (low uncertainty) to 1 (high uncertainty). Qualitative evaluations of soil class predictions were made by visually comparing patterns in the modeled data and the published soil survey.

RESULTS

Covariate Selection

The number of variables retained from RFE varied slightly for the different neighborhood sizes (Table 2). Comparison of the top five variables selected from each neighborhood size illustrated the importance of the gypsic index and band ratio 7:5 from the May Landsat scene (representing dry conditions), and the SAGA wetness index, as these covariates were selected in all models. Elevation and the topographic position index (a 1000-pixel window) were important for smaller neighborhoods

Table 2. Top five variables selected via recursive feature elimination (RFE) using the random forest algorithm and the total number of variables selected for each circular neighborhood (see Table 1 for additional information about the variables).†

Neighborhood radius, m	Variable 1	Variable 2	Variable 3	Variable 4 ‡	Variable 5	Number of variables selected via RFE†
0	Elevation	TPI_1000	SAGA wetness index	May gypsic index	May band ratio 7:5	36
15	TPI_1000	May band ratio 7:5	May gypsic index	Elevation	SAGA wetness index	35
30	May band ratio 7:5	May gypsic index	SAGA wetness index	Elevation	TPI_1000	33
60	May band ratio 7:5	May gypsic index	SAGA wetness index	Elevation	TPI_1000	35
90	May band ratio 7:5	May gypsic index	SAGA wetness index	Elevation	Time series_PCA_4	27
120	May gypsic index	May band ratio 7:5	SAGA wetness index	MRVBF	Elevation	34
150	May gypsic index	May band ratio 7:5	SAGA wetness index	MRVBF	Time series_PCA_4	28
180	May gypsic index	May band ratio 7:5	SAGA wetness index	MRVBF	Time series_PCA_4	35
Multiscale	May 7:5_180 m	May gypsic index_180 m	May 7/5_150 m	May gypsic index_150 m	SAGA wetness index_90 m	246

†The original number of covariates from which the RFE was applied was 36 for each neighborhood size and 288 for the multiscale combination using all neighborhood sizes.
‡ TPI_1000, topographic position index with a 1000 pixel radius; MRVBF, multiresolution index of river bottom flatness; PCA, principal component analysis.

and gave way to the multiresolution index of river bottom flatness for the larger neighborhoods. The fourth principal component of the 3-yr NDVI time series was important for larger neighborhood sizes. For the multiscale combination of covariates, predictors representing three of the larger neighborhoods (180, 150, and 90 m) were selected for the top five covariates. Interestingly, both the band ratio of 7:5 and the gypsic index from the May (dry) Landsat scene at the largest neighborhoods appeared twice in the top five variables of the multi-scale model and only one topographic variable was in the top five (SAGA wetness index with a 90-m neighborhood).

Model Performance

A comparison of the 72 individual models showed a fair to moderate agreement for predicting soil series classes with K values ranging between 0.24 and 0.44 (Fig. 3). Model performance increased with neighborhood size for all eight models evaluated. The radial support vector machine and random forest models had the greatest value of K across all neighborhood sizes. Nearest

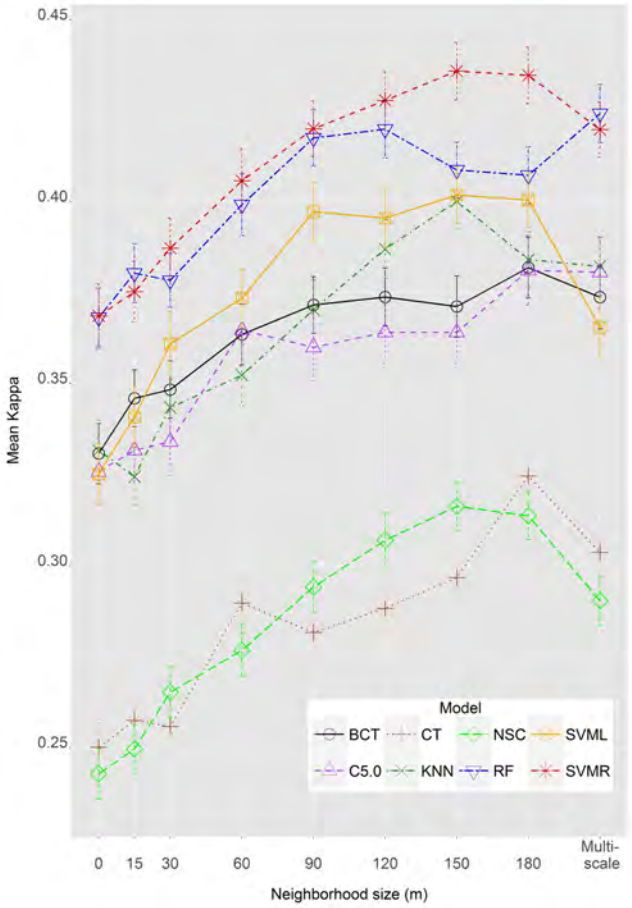

Fig. 3. Comparison of eight machine learning models and the effect of neighborhood size (the number indicates the radius) on the prediction of soil series in the San Bernardino watershed of southeastern Arizona. Multiscale represents the model with all neighborhood sizes. Each point represents the mean of 100 leave-group-out cross-validations, where 70% of the points were used for training and 30% for validation. Models include bagged classification trees (BCT), C5.0, classification tree (CT), nearest shrunken centroids (NSC), *k*-nearest neighbors (KNN), random forests (RF), radial support vector machines (SVMR), and linear support vector machines (SVML).

shrunken centroids and the random partition classification tree models had the poorest performance and all other models were intermediate. A neighborhood of 150 to 180 m emerged as a threshold beyond which model performance did not improve. Within each class of models, including covariates from all extracted neighborhoods (n = 288) did not improve performance over the largest neighborhood (n = 36), except for the random forest model. The model with overall highest K was the support vector machine model with a radial kernel and a 150-m neighborhood (SVMR.150; K = 0.44). The best performing multiscale model was the random forest model (RF.multiscale; K = 0.42). Spatial representation of the confusion index from the RF.multiscale model showed large portions of the study area with relatively high confusion between the top two predicted soil series classes for each 10-m raster pixel; however, localized zones of low confusion were also visible (Fig. 4).

Both producer's accuracy and user's accuracy were similar for the majority of predicted soil series for the SVMR.150 and RF.multi-scale models (Fig. 5, Supplemental Table S2 and Supplemental Table S3). However, the RF.multiscale model showed improved accuracy for several of the series with low numbers of samples. For example, six of the series had no correct classifications in the SVMR.150 model but did have correct classifications in the RF.multi-scale model. This demonstrates the RF.multi-scale model was better able to predict more soil classes than the SVMR.150.

Comparison of the user's and producer's accuracy for each class across all nine neighborhood sizes of the random forest model indicated that some classes were best modeled with a range of sizes (data not shown). Similar to Behrens et al. (2010b), some classes were better predicted with some attributes than others and had very similar accuracies across all neighborhood sizes, whereas others were more sensitive to the neighborhood size. In general, the larger neighborhoods (including RF.multiscale) produced greater accuracies for individual classes than smaller ones.

Spatial Prediction

The machine learning models captured the spatial variability of major soil types represented by the published soil map with varying degrees of similarity (Fig. 4 and Fig. 6). Output from the SVMR.150 model, which had the highest K, produced generalized patterns of soil types across the study area. Soils occupying small areas, such as narrow drainages, were not well represented in the generalized model because the averaged covariate values masked the fine-scale patterns. Generalized patterns of soil in upland landscapes with large coverage were well represented. In contrast, the best performing multiscale model (RF.multiscale) predicted a variety of both generalized and detailed soil features across the landscape. The SVMR.150 model predicted only 26 of the 28 unique soil series classes, in contrast to the RF.multiscale model, which predicted all 28. Models generated for smaller neighborhood sizes produced more complex maps of soil types, suggesting extensive heterogeneity across the study area (Fig. 6).

Because of the edge effects of covariate data, the final area of predicted soil types in this study area was smaller than the area

originally identified (Fig. 4). The most pronounced edge effects were on the western edge of the study area, which paralleled the edge of Landsat scenes. Variable boundaries of some Landsat scenes resulted in a reduction in usable Landsat data after neighborhood filtering, which caused the models to predict 'No Data' values. Additionally, some soil map units on the western edge of the study area (<1% of the study area) were very small and fell within the 'No Data' region of the final prediction maps.

Although the models produced a wide range of K values across neighborhood sizes, there were many similar patterns captured with them. For example, Fig. 6 illustrates the ability of random forest models derived from 0-, 90-, and 180-m and multiscale neighborhoods to delineate a surge ring surrounding a volcanic crater. The 0-m neighborhood showed the most complexity and predicted more individual soil classes than other models. In contrast, the larger neighborhood models produced more generalized maps of soil series and the RF.multiscale model retained some detailed information but heavily favored the general patterns of the larger neighborhoods. Maps of the corresponding confusion indices indicated high confusion for most of the area, with only localized areas of low confusion. The common patterns of prediction suggest that there may be utility in some sort of model averaging to produce more robust prediction maps of soil classes.

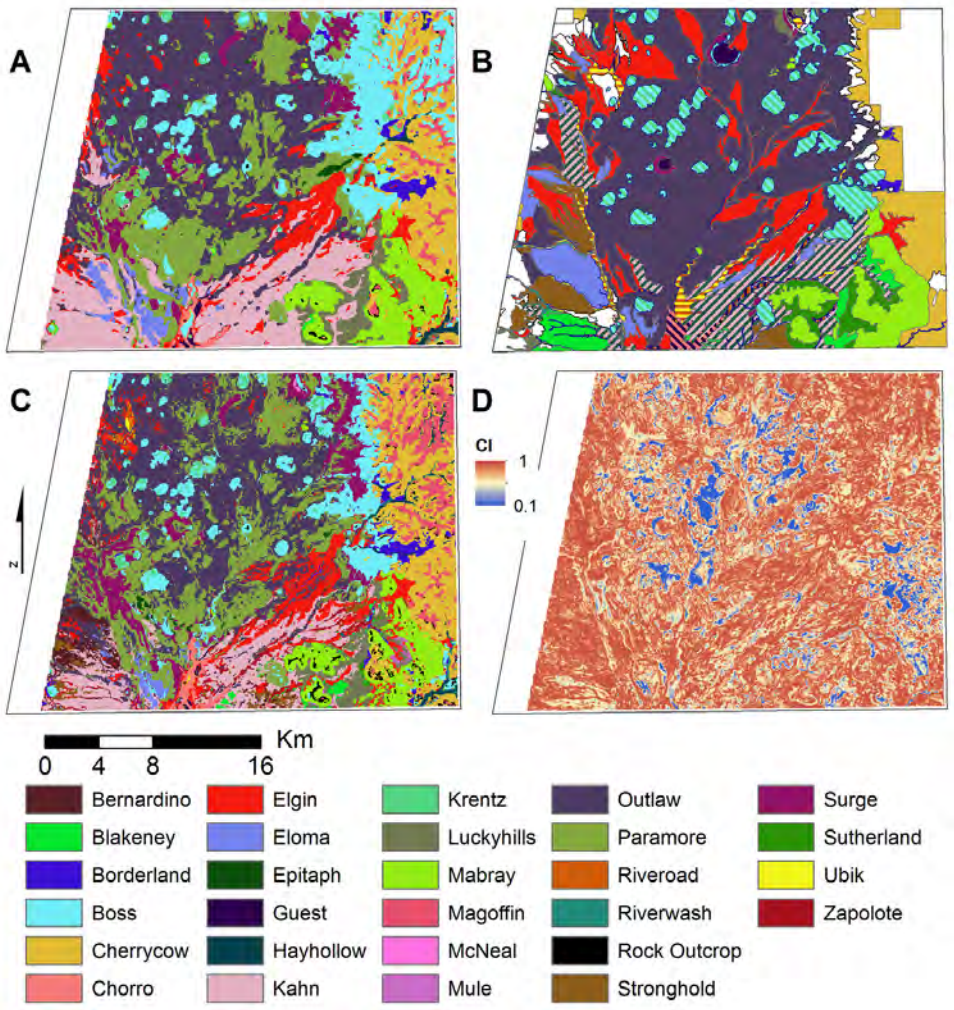

Fig. 4. Soil series map in the San Bernardino watershed of southeastern Arizona resulting from a radial support vector machine model using covariate data summarized with 150-m circular neighborhoods (A), a dominant component map of published Soil Survey Geographic Database (SSURGO) (B), a random forest model trained with multi-scale covariate data (C), and the confusion index (CI) for the random forest model (D). The legend for soil series applies to all prediction maps; SSURGO map units with the same percentage of dominant components are represented with appropriate shading. The black rectangle is the study area extent. White areas in A, B, and C represent no data values as a result of edge effects of covariate data; white areas in B indicate SSURGO map units with no components represented in the training data. Additionally, the large white area on the eastern edge of B is national forest land with no published SSURGO data available.

DISCUSSION

Utility of Covariates

Some interesting findings of this study were that the indices developed to identify specific features in other study areas effectively identified soil and geology features that differed from their original uses. For example, the gypsic index was developed to identify gypsic soil areas using short-wave infrared wavelengths (Landsat Bands 5 and 7) (Nield et al., 2007); however, this index was able to identify patterns of volcanic cinder cones and basalt flows that were not influenced by gypsum in this study area. The same short-wave infrared wavelengths are also important for distinguishing clay minerals and geologic formations (Yazdi et al., 2013). Similarly, the natric index was developed to identify natric soil areas sodium-rich using near- and short-wave infrared wavelengths (Bands 4 and 5) (Nield et al., 2007) but it also captured patterns of volcanic soils. Vegetation indices were originally developed to predict plant characteristics (e.g., vegetation cover, biomass, and phenology) but are also generally important variables for predicting soil features, including taxonomic classes (Odgers et al., 2014; Rad et al., 2014). Principal component four from the NDVI time series was important in the coarser models, confirming the results of other studies showing that multitemporal NDVI patterns capture soil property differences (Dobos et al., 2000; Levi et al., 2015; Maynard and Levi, 2017). The utility of vegetation indices for predicting soil properties is explained by the tight coupling of soil–vegetation–climate relationships in semiarid systems where soil texture influences vegetation dynamics (Maynard and Levi, 2017).

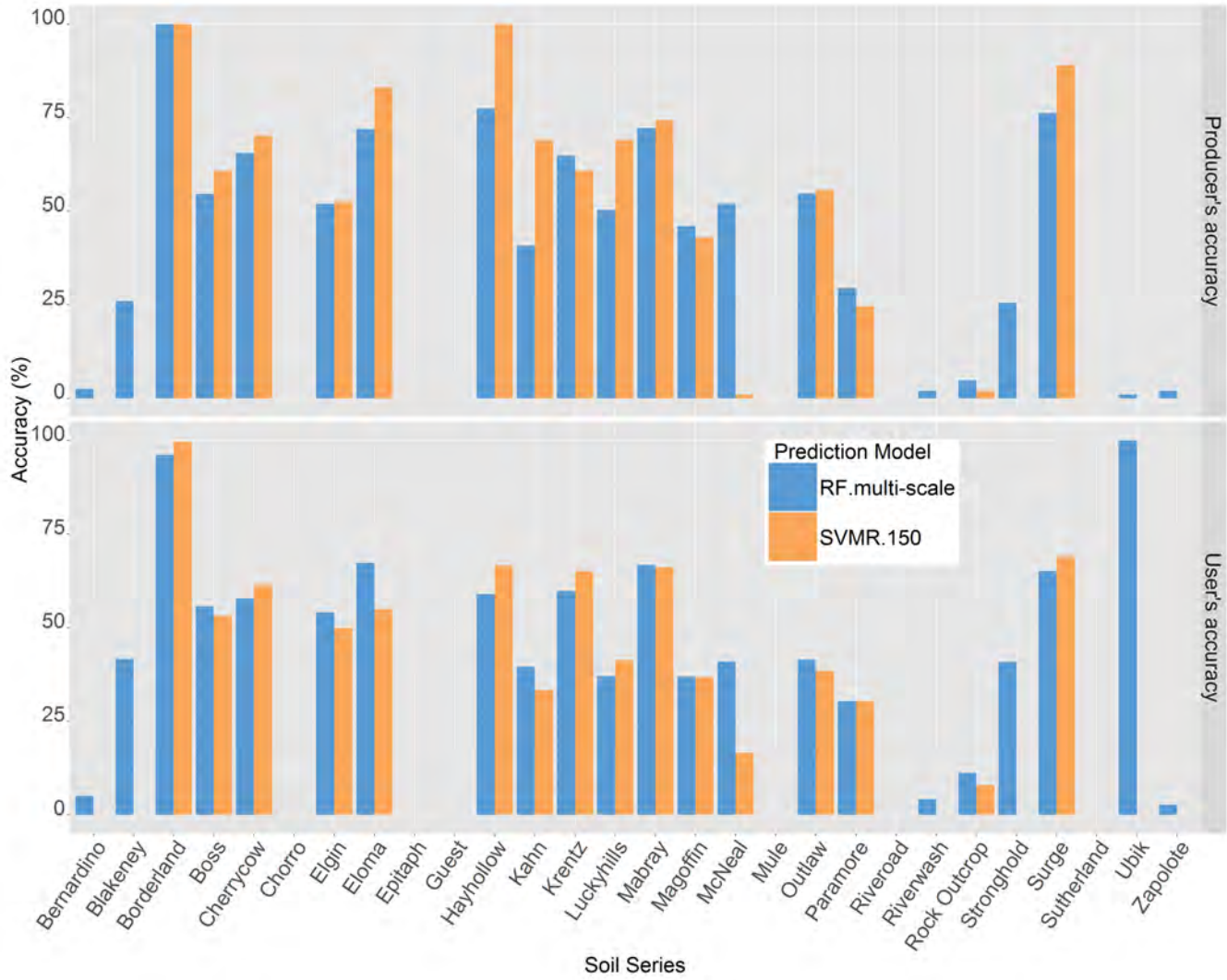

Fig. 5. User's and producer's accuracy for 28 soil series modeled with the support vector machine model with a radial kernel and a 150-m neighborhood (SVMR.150) (orange) and the random forest multiscale (RF.multiscale models) (blue) derived from 100 iterations of leave-group-out cross-validation.

Model Performance

Complex prediction models performed better than simple models, which corroborates the results of other studies that compared multiple machine learning approaches to modeling soil taxonomic classes (Brungard et al., 2015; Heung et al., 2016; Kovačević et al., 2010). My results indicate that random forest models were not significantly different (i.e., the 95% confidence intervals of the respective mean K overlapped) from the radial support vector machine models for seven of the nine neighborhood sizes (Fig. 3). Both random forest models (Chaney et al., 2016; Nauman et al., 2014; Rad et al., 2014) and support vector machine approaches (Hahn and Gloaguen, 2008; Lorenzetti et al., 2015) are commonly used to map soil taxonomic classes. Although random forest models have been shown to be robust classifiers, high model performance in this study may also be influenced by use of the random forest algorithm for covariate selection via RFE (Brungard et al., 2015). An interesting result is that seven of the eight models evaluated had a lower K for the multiscale model than the best performing neighborhood size in each respective suite of models. This pattern may be an artifact of the RFE procedure used to choose the important covariates for model development and future work should explore the performance of this approach compared with other available variable selection algorithms.

The ranges of K in this study were similar to those reported by Brungard et al. (2015) for soil taxonomic groups in similar landscapes in the western United States (0.1–0.59). Hengl et al. (2007) reported K values ranging from 0.33 to 0.54 for soil taxa across all of Iran, which overlapped the range of K in this study. Another study in Iran by Taghizadeh-Mehrjardi et al. (2015) showed that ranges of K for soil family were higher than those in this study (0.51–0.69) and increased with more general taxonomic groups up to a maximum of 0.84. K was lower than that reported by Heung et al. (2016) in southern British Columbia, although they used single-component soil surveys for creating sampling design and modeling, in contrast to my use of sample points distributed in multi-component surveys. Classification performance is impacted by the complexity of the landscape, the overall number of classes to be predicted, and the number of samples in each class and their spatial distribution (Brungard et al., 2015; Jafari et al., 2012).

The number of classes predicted in this study (28 soil series) is typical of DSM studies focused on taxonomic classification. For example, Brungard et al. (2015) predicted between three and five soil classes for separate study areas in the western United States, Taghizadeh-Mehrjardi et al. (2015) predicted five soil classes in an Iranian study area, Hengl et al. (2007) modeled 15 soil classes across the entire country of Iran, Silva et al. (2016) predicted four soil classes in Brazil, and Jafari et al. (2012) predicted 18 soil great groups in Iran. These studies focused on spatial predictions of higher levels of soil taxonomy than I did, which resulted in fewer unique soil classes. In contrast, Odgers et al. (2014) modeled 71 soil classes in Australia with 22.5% validation success for the most probable soil, and Nauman and Thompson (2014) predicted 56 soil series classes in West Virginia with 24% validation success. Model performance generally decreases with more detailed predictions of taxonomic classes (Heung et al., 2016, Taghizadeh-Mehrjardi et al., 2015). This is because the purity of the soil map units and the pedological diversity within a given map unit is inversely related, for models across the soil order, to subgroup hierarchy (Jafari et al., 2013). Modeling detailed taxonomic classes also increases the likelihood of class imbalance issues, which complicates model training.

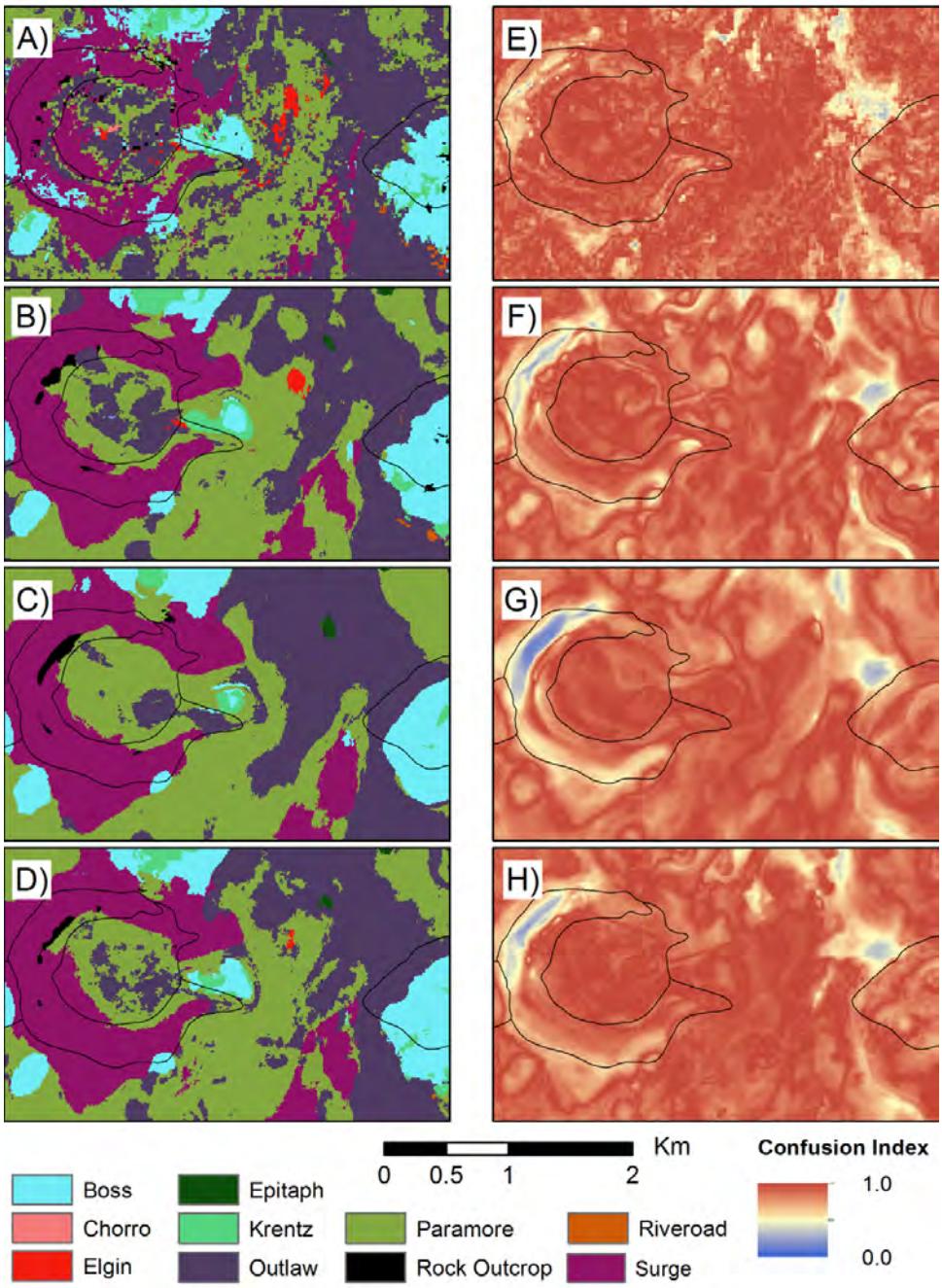

Fig. 6. Soil series predictions and respective confusion index maps from random forest models using unaltered covariates (A, E), focal means of covariates for 60 m (B,F) 180 m (C,G), and multi-scale (D,H) neighborhood sizes for a focus area with volcanic cinder cones and basalt flows.

The confusion index from the RF.multiscale model showed large portions of the study area with high confusion and smaller patches or zones with much lower confusion. This is a pattern typical of other studies that have reported high confusion indices or uncertainty values when predicting soil types (Brungard et al., 2015; Chaney et al., 2016; Nauman et al., 2014; Odgers et al., 2011, 2014). The most likely reason for the high confusion index values across large areas is that similar soils geographically associated with one another were misclassified as each another (Supplemental Tables S2 and Supplemental Table S3). Another possible explanation for the patterns of the confusion index could be problems associated with class imbalance. Some soil classes were only represented in small numbers in the available training data, which limited the ability of the prediction models to identify unique criteria for differentiating classes (Subburayalu and Slater, 2013). The classes with small numbers of training data were not represented with the same spatial extent as those classes with many more cases, which is likely to have led to problems with model performance.

Two general approaches have been suggested for overcoming class imbalance in soil prediction models: (i) increase observations for under-represented classes or (ii) reduce the number of classes to be modeled (Brungard et al., 2015). Perhaps the simplest solution for increasing the sample size of under-represented classes is to collect more field data; however, this is not always feasible because of time, cost, and logistical constraints. An alter-

native is to add 'synthetic' sample points based on probabilistic sample designs that use existing soil survey maps (Chaney et al., 2016; Heung et al., 2016; Odgers et al., 2014; Subburayalu et al., 2014). Soil surveys can also be mined for soil–landscape rules to predict soil class distributions in space.

A second approach for addressing class imbalance issues is to reduce the number of soil classes being predicted. This has been achieved by removing training data that represent classes with a small sample size in order to model only dominant soil classes (Kovačević et al., 2010; Subburayalu et al., 2014) and also by combining training data into similar groups on the basis of soil properties (Rad et al., 2014; Kempen et al., 2009) or taxonomic classification (Brungard et al., 2015). Reducing the number of soil classes being predicted can improve model performance but modeling only dominant soils generalizes the predictions and arguably provides limited information gain compared with a traditional soil survey. Incorporating a combination of these methods for addressing class imbalance issues with existing soil point data is likely to provide a more robust approach for improving soil mapping.

Scaling Pattern and Process Relationships

Soil patterns are a function of many processes occurring at different spatial scales. Covariate data from multiple sources are often integrated for DSM exercises because representation of the soil-forming factors requires many indices to reflect soil formation across a variety of spatial scales (Grunwald, 2009; Grunwald et al., 2011; McBratney et al., 2003). My results illustrate the importance of scale to prediction accuracy. Similar to Miller (2014), who calibrated topographic attributes to analyze scales used by soil scientists, model performance increased with the size of analysis scale up to some threshold that presumably represented an appropriate scale of landscape processes. Working at very coarse scales in France (1:250,000), Grinand et al. (2008) also found that model performance for soil classes increased with increasing the analysis scale. Although the overall model performance increased with neighborhood size in this study, it did result in poor representation of soil classes with low area/perimeter ratios and small spatial extent (e.g., drainages). This is similar to the results reported by Behrens et al. (2010b) in a forested site in Germany, which concluded that a multiscale approach provided minimal benefit for predicting small or elongated soil classes (e.g., drainages) because these soils were controlled by local-scale landscape processes and larger neighborhood sizes effectively masked their signal in covariate data. This explains why the model with the largest K (SVMR.150) had limited ability to predict soils in drainages, relative to the RF.multiscale model and other smaller neighborhood sizes. The more generalized prediction maps smoothed the covariate data and misrepresented the respective soil classes (Roecker and Thompson, 2010).

The most realistic model for preserving general and detailed soil class distributions was the RF.multiscale model (Fig. 4). It is common for different study areas to require unique sets of environmental covariates for optimal models. For example, Brungard et al. (2015) predicted soil taxonomic classes in three study areas across the western United States with the same types of models and covariates and found that optimal predictors differed by region. They also found that a range of pixel sizes were important for predicting soils in these regions. A study conducted in the Sonoran Desert by Nauman et al. (2014) also reported the importance of multiscale predictors for modeling soil types. Another consideration for optimizing model performance for predicting soil classes is that individual soil classes may require unique sets of multiscale predictors to produce the best model (Behrens et al., 2010b). Further research is needed to address this concept.

Choosing the correct pixel size and modeling scale are important and potentially challenging elements of DSM (Hengl, 2006; Malone et al., 2013). The use of a 10-m pixel size for covariates in this study follows the recommendations of Maynard and Johnson (2014), who suggested that moderate pixel size provides considerable flexibility for evaluating the influence of neighborhood summaries. It is possible that the results would have been different if soil series had been modeled at a different pixel size. Furthermore, changing pixel size and neighborhood size may prove useful for meeting use-specific needs of soil information across scales.

Deriving topographic covariates with different neighborhood extents is relatively easy and is commonly used for DSM (Behrens et al., 2010b; Maynard and Johnson, 2014; Miller, 2014) but focal summaries of reflectance indices for training soil prediction models is less common (Grinand et al., 2008; Vasques et al., 2012). For example, Nauman et al. (2014) used neighborhood SD for single bands as predictors for soil classes but did not compute mean values of bands or other indices. Vasques et al. (2012) applied smoothing functions to resample reflectance pixels to the larger pixel size for predicting soil C stocks. I argue that training soil prediction models with a neighborhood of covariate data around the sample points is important to link information in the covariates to the extent of similar soils on the ground and produce a more robust model of soil prediction. The optimal size and shape of the neighborhood is largely controlled by the soil-forming environment because soil–landscape features are closely tied to multiscale processes. Aggregating covariate data with multiple neighborhood sizes can help to link conceptual and quantitative models of soil pedogenesis to spatial predictions.

Spatial Prediction Limitations

It is important to note that this study area represents only a portion of the county soil survey and that some map unit descriptions reflect composition from samples outside the study area. Consequently, some soil series could not be modeled. For example, Riverwash is a miscellaneous unit that is easily discerned via aerial photography that was not sampled within the study area. Eight percent of the study area fell into the category of soil map units without representation by at least one component. In some cases, inadequate numbers of cases for a given soil series (i.e., <2) precluded those classes from entering the domain of potential classes for models to predict. This class imbalance can explain some differences in the patterns and composition of the soil prediction models compared with the published survey. Although obtain-

ing additional sample points to represent under-represented soil classes would have represented the study area better, further data collection was not possible because of logistical and financial constraints. Incorporating probabilistic approaches with existing soil point data and contextual soil map unit information would probably be a very useful improvement for modeling soil class variables.

Although the RF.multiscale model identified some of the patterns of soil features, it did not always assign them to the correct class. For example, some classes in drainages (e.g., Riverwash, Riveroad, Ubik) were differentiated from surrounding features but assigned an incorrect class attribute (Fig. 4). This is likely to reflect the low number of training data for these classes. In contrast, the SVMR.150 model did a poor job of identifying the patterns of these same cases because the 150-m smoothing of the covariate data masked these small features. The results suggest that incorporating multiscale predictors (e.g., neighborhood summaries of covariates) may be useful for capturing both dominant and rare soil–landscape features. Although the aggregation of training data can account for some spatial errors that may be present in legacy pedons, a post hoc comparison of model accuracy showed very little difference between legacy pedons and the cLHS pedons used in this study. Hence, the use of both legacy pedons and cLHS pedons in this modeling exercise appear to be acceptable, despite some spatial uncertainty in the point locations of the legacy pedons.

Misclassifications for both the SVMR.150 and RF.multiscale models were generally a result of predicting a similar soil within a common landform (Supplemental Tables S2 and Supplemental Table S3). For example, in both models, Boss, Krentz, and Paramore soils, which commonly occur on volcanic cinder cones, were often confused with one another. Similarly, the Cherrycow and Magoffin series occur on hills and mountains of volcanic parent materials and were confused with one another. Series in drainages that were misclassified were commonly confused with other series found in floodplain or alluvial fan settings. These trends mark the geographic association of these soils with one another in soil map units and are likely to reflect the development of the soil map units in the initial soil survey. These types of misclassification errors illustrate the source of overall classification errors and future work should focus on ways to differentiate geographically associated soils that have similar physical characteristics.

CONCLUSION

I have presented a method of using covariate data to train soil prediction models that uses concepts used in traditional soil surveys. Comparison of multiple models provides useful insights into the range of prediction accuracy that can be obtained. A threshold of a 150-m radius of aggregated covariate data emerged as the best scale to optimize overall model performance for this study area. There was, however, a tradeoff between overall model performance and individual class accuracy, where smaller neighborhoods had lower overall accuracy but predicted patterns of rare soils and larger neighborhoods had greater overall accuracy and predicted rare soils poorly. In contrast, the multiscale approach of integrating covariate data for soil prediction models produced a more realistic map with a combination of detailed and general soil–landscape patterns. Covariates selected in all models reflected an even mixture of both topographic and remotely sensed variables with differences in variable importance for different neighborhood sizes. Evaluating a range of neighborhood sizes for aggregating covariate data provides a method of accounting for multiscale processes that are important for predicting soil patterns without modifying the pixel size of final maps. Incorporating concepts from traditional soil surveys with DSM approaches can strengthen ties between the two and optimize the extraction of landscape information for predicting soil classes.

ACKNOWLEDGEMENTS

This work was supported by the USDA ARS Postdoctoral Research Associate Program and the National Cooperative Soil Survey Soil Survey Research Grant Program (Agreement # 6235-11210-007-31). I acknowledge W. Glenn, K. Kimbrough, B. McDonald, A. Magoffin, and T. Nevins for allowing me to access their ranches for sampling. I also thank B. Bestelmeyer, T. Nauman, S. Spiegal, D. Hirmas, and the two anonymous reviewers for valuable comments on previous drafts of the manuscript; B. Svetlik for assistance with the collection and interpretation of legacy soil survey data; and C. Brungard for assistance with R code.

SUPPLEMENTARY MATERIAL

The supplemental tables show detailed results of classification accuracy including soil map unit composition and the modeled distributions of each component for the RF.multiscale and SVMR.150 models. Comprehensive error matrices of the respective models also show the user's and producer's accuracy for each modeled soil component.

REFERENCES

Balkovic, J., Z. Rampasekova, V. Hutar, J. Sobocka, and R. Skalsky. 2013. Digital soil mapping from conventional field soil observations. Soil Water Res. 8:13–25.

Ballabio, C. 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151:338–350. doi:10.1016/j.geoderma.2009.04.022

Behrens, T., K. Schmidt, L. Ramirez-Lopez, J. Gallant, A.X. Zhu, and T. Scholten. 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213:578–588. doi:10.1016/j.geoderma.2013.07.031

Behrens, T., K. Schmidt, A.X. Zhu, and T. Scholten. 2010a. 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

Behrens, T., A.X. Zhu, K. Schmidt, and T. Scholten. 2010b. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155:175–185. doi:10.1016/j.geoderma.2009.07.010

Biggs, T.H., R.S. Leighty, S.J. Skotnicki and P.A. Pearthree. 1999. Geology and geomorphology of the San Bernardino Valley, southeastern Arizona. Arizona Geological Survey Open File Report, OFR-99-19. Arizona Geological Survey, Tucson, AZ.

Boehner, J., R. Koethe, O. Conrad, J. Gross, A. Ringeler, and T. Selige. 2002. Soil regionalisation by means of terrain analysis and process parameterisation. In: E. Micheli, F. Nachtergaele, L. Montanarella, editors, Soil classification 2001. Research Report 7, EUR 20398 EN. European Soil Bureau, Luxembourg City. p. 213–222.

Boettinger, J.L., R.D. Ramsey, J.M. Bodily, N.J. Cole, S. Kienast-Brown, S.J. Nield, et al. 2008. Landsat spectral data for digital soil mapping. In: A.E. Hartemink, A. McBratney, M.L. Mendonça-Santos, editors, Digital soil mapping with limited data. Springer-Verlag, Dordrecht, the Netherlands. p. 193–202.

Brown, D.E. 1994. Biotic communities: Southwestern United States and northwestern Mexico. University of Utah Press, Salt Lake City.

Brown, D.E., and C.H. Lowe. 1994. Biotic communities of the Southwest. University of Utah Press, Salt Lake City.

Brungard, C.W., J.L. Boettinger, M.C. Duniway, S.A. Wills, and T.C. Edwards, Jr. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240:68–83. doi:10.1016/j.geoderma.2014.09.019

Bui, E.N., B.L. Henderson, and K. Viergever. 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Modell. 191:431–446. doi:10.1016/j.ecolmodel.2005.05.021

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

Carré, F., and M.C. Girard. 2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110:241–263. doi:10.1016/S0016-7061(02)00233-1

Cavazzi, S., R. Corstanje, T. Mayr, J. Hannam, and R. Fealy. 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma 195–196:111–121. doi:10.1016/j.geoderma.2012.11.020

Chaney, N.W., E.F. Wood, A.B. McBratney, J.W. Hempel, T.W. Nauman, C.W. Brungard, et al. 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274:54–67. doi:10.1016/j.geoderma.2016.03.025

Congalton, R.G. 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

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, et al. 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8:1991–2007. doi:10.5194/gmd-8-1991-2015

Dobos, E., E. Micheli, M.F. Baumgardner, L. Biehl, and T. Helt. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97:367–391. doi:10.1016/S0016-7061(00)00046-X

Environmental Systems Research Institute. 2015. ArcGIS version 10.4. Environmental Systems Research Institute, Redlands, CA.

Foody, G.M. 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80:185–201. doi:10.1016/S0034-4257(01)00295-4

Forzieri, G., F. Castelli, and E.R. Vivoni. 2011. Vegetation dynamics within the North American monsoon region. J. Clim. 24:1763–1783. doi:10.1175/2010JCLI3847.1

Freeman, T.G. 1991. Calculating catchment area with divergent flow based on a regular grid. Comput. Geosci. 17:413–422. doi:10.1016/0098-3004(91)90048-I

Gallant, J.C., and T.I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39:1347. doi:10.1029/2002WR001426

Grinand, C., D. Arrouays, B. Laroche, and M.P. Martin. 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. 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195–207. doi:10.1016/j.geoderma.2009.06.003

Grunwald, S., J.A. Thompson, and J.L. Boettinger. 2011. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. Soil Sci. Soc. Am. J. 75:1201–1213. doi:10.2136/sssaj2011.0025

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

Hahn, C., and R. Gloaguen. 2008. Estimation of soil types by non linear analysis of remote sensing data. Nonlinear Process. Geophys. 15:115–126. doi:10.5194/npg-15-115-2008

Häring, T., E. Dietz, S. Osenstetter, T. Koschitzki, and B. Schroder. 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185–186:37–47. doi:10.1016/j.geoderma.2012.04.001

Hengl, T. 2006. Finding the right pixel size. Comput. Geosci. 32:1283–1298. doi:10.1016/j.cageo.2005.11.008

Hengl, T., N. Toomanian, H.I. Reuter, and M.J. Malakouti. 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

Heung, B., H.C. Ho, J. Zhang, A. Knudby, C.E. Bulmer, and M.G. Schmidt. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265:62–77. doi:10.1016/j.geoderma.2015.11.014

Hudson, B.D. 1992. The soil survey as a paradigm-based science. Soil Sci. Soc. Am. J. 56:836–841. doi:10.2136/sssaj1992.03615995005600030027x

Jafari, A., S. Ayoubi, H. Khademi, P.A. Finke, and N. Toomanian. 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 [erratum: 223: 19].

Jafari, A., P.A. Finke, J. Van de Wauw, S. Ayoubi, and H. Khademi. 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

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. An introduction to statistical learning: With applications in R. Springer, New York.

Kempen, B., D.J. Brus, G.B.M. Heuvelink, and J.J. Stoorvogel. 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151:311–326. doi:10.1016/j.geoderma.2009.04.023

Kovačević, M., B. Bajat, and B. Gajic. 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. doi:10.18637/jss.v028.i05

Kuhn, M., and K. Johnson. 2013. Applied predictive modeling. Springer, New York.

Landis, J.R., and G.G. Koch. 1977. Measurement of observer agreement for categorical data. Biometrics 33:159–174. doi:10.2307/2529310

Levi, M.R., and C. Rasmussen. 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219–220:46–57. doi:10.1016/j.geoderma.2013.12.013

Levi, M.R., M.G. Schaap, and C. Rasmussen. 2015. Application of spatial pedotransfer functions to understand soil modulation of vegetation response to climate. Vadose Zone J. 14. doi:10.2136/vzj2014.09.0126

Lorenzetti, R., R. Barbetti, M. Fantappie, G. L'Abate, and E.A.C. Costantini. 2015. Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. Geoderma 237–238:237–245. doi:10.1016/j.geoderma.2014.09.006

Malone, B.P., A.B. McBratney, and B. Minasny. 2013. Spatial scaling for digital soil mapping. Soil Sci. Soc. Am. J. 77:890–902. doi:10.2136/sssaj2012.0419

Maynard, J.J., and M.G. Johnson. 2014. Scale-dependency of LiDAR derived terrain attributes in quantitative soil–landscape modeling: Effects of grid resolution vs. neighborhood extent. Geoderma 230–231:29–40. doi:10.1016/j.geoderma.2014.03.021

Maynard, J.J., and M.R. Levi. 2017. Hyper-temporal remote sensing for digital soil mapping: Characterizing soil–vegetation response to climatic variability. Geoderma 285:94–109. doi:10.1016/j.geoderma.2016.09.024

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

Miller, B.A. 2014. Semantic calibration of digital terrain analysis scale. Cartogr. Geogr. Inf. Sci. 41:166–176. doi:10.1080/15230406.2014.883488

Miller, B.A., S. Koszinski, M. Wehrhan, and M. Sommer. 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239–240:97–106. doi:10.1016/j.geoderma.2014.09.018

Minasny, B., and A.B. McBratney. 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., and A.B. McBratney. 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

Nauman, T.W., and J.A. Thompson. 2014. Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. Geoderma 213:385–399. doi:10.1016/j.geoderma.2013.08.024

Nauman, T.W., J.A. Thompson, and C. Rasmussen. 2014. Semi-automated disaggregation of a conventional soil map using knowledge driven data mining and random forests in the Sonoran Desert, USA. Photogramm. Eng. Remote Sens. 80:353–366. doi:10.14358/PERS.80.4.353

Nield, S.J., J.L. Boettinger, and R.D. Ramsey. 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Sci. Soc. Am. J. 71:245–252. doi:10.2136/sssaj2006-0049

Neilson, R.P. 1987. Biotic regionalization and climatic controls in western North America. Vegetatio 70:135–147.

Nelson, M.A., T.F.A. Bishop, J. Triantafilis, and I.O.A. Odeh. 2011. An error budget for different sources of error in digital soil mapping. Eur. J. Soil Sci. 62:417–430. doi:10.1111/j.1365-2389.2011.01365.x

Odgers, N.P., A.B. McBratney, and B. Minasny. 2011. Bottom-up digital soil mapping. II. Soil series classes. Geoderma 163:30–37. doi:10.1016/j.geoderma.2011.03.013

Odgers, N.P., W. Sun, A.B. McBratney, B. Minasny, and D. Clifford. 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215:91–100. doi:10.1016/j.geoderma.2013.09.024

Olden, J.D., J.J. Lawler, and N.L. Poff. 2008. Machine learning methods without tears: A primer for ecologists. Q. Rev. Biol. 83:171–193. doi:10.1086/587826

PRISM Climate Group. 2012. PRISM climate data. Oregon State University. http://www.prism.oregonstate.edu/ (accessed 16 Mar. 2017).

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/ (accessed 16 Mar. 2017).

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

Roecker, S.M., and J.A. Thompson. 2010. Scale effects on terrain attribute calcluation and their use as environmental covariates for digital soil mapping. In: J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink, and S. Kienest-Brown, editors, Digital soil mapping: Bridging research, production, and environmental application. Springer-Verlag, Dordrecht, the Netherlands. p. 55–66.

Roudier, P. 2011. clhs: A R package for conditioned Latin hypercube sampling. R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/clhs/ (accessed 20 Mar. 2017).

Samuel-Rosa, A., G.B.M. Heuvelink, G.M. Vasques, and L.H.C. Anjos. 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244:214–227. doi:10.1016/j.geoderma.2014.12.017

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

Silva, S.H.G., M.D. de Menezes, P.R. Owens, and N. Curi. 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267:65–77. doi:10.1016/j.geoderma.2015.12.025

Soil Survey Staff. 1993. Soil survey manual. USDA Handb. 18. USDA Soil Conservation Service, Washington, DC.

Soil Survey Staff. 2013. Soil Survey Geographic (SSURGO) database for Cochise County, Arizona, Douglas-Tombstone part. USDA-NRCS., Available online at. http://soildatamart.nrcs.usda.gov (accessed 30 Aug. 2013).

Stum, A.K., J.L. Boettinger, M.A. White, and R.D. Ramsey. 2010. Random forests applied as a soil spatial predictive model in arid Utah. In: J.L. Boettinger, D.W. Howell, A.C. Moore, and S. Kienast-Brown, editors, Digital soil mapping: Bridging research, environmental application, and operation. Springer, Dordrecht, p. 179–190.

Subburayalu, S.K., I. Jenhani, and B.K. Slater. 2014. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. Geoderma 213:334–345. doi:10.1016/j.geoderma.2013.08.018

Subburayalu, S.K., and B.K. Slater. 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-Mehrjardi, R., K. Nabiollahi, B. Minasny, and J. Triantafilis. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253–254:67–77. doi:10.1016/j.geoderma.2015.04.008

Tobler, W.R. 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46:234–240. doi:10.2307/143141

Vasques, G.M., S. Grunwald, and D.B. Myers. 2012. Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. J. Geophys. Res. Biogeosci. 117. doi:10.1029/2012JG001982

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

Yazdi, M., M. Taheri, P. Navi, and N. Sadati. 2013. Landsat ETM plus imaging for mineral potential mapping: Application to Avaj area, Qazvin, Iran. Int. J. Remote Sens. 34:5778–5795. doi:10.1080/01431161.2013.797127