

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

2
3 *Colby W. Brungard^a Department of Plants, Soils and Climate, 4820 Old Main Hill, Utah State University Logan, UT,*
4 *84322, USA. Email: envsoilco@gmail.com. Corresponding author. Ph: +14357973404*

5 *Janis L. Boettinger^a Department of Plants, Soils and Climate, 4820 Old Main Hill, Utah State University Logan, UT,*
6 *84322, USA. Email: janis.boettinger@usu.edu.*

7 *Michael C. Duniway^b U.S. Geological Survey, Southwest Biological Science Center, 2290 SW Resource Blvd, Moab,*
8 *UT 84532, USA. Email: mduniway@usgs.gov.*

9 *Skye A. Wills^c National Soil Survey Center, Natural Resources Conservation Service – United States Department of*
10 *Agriculture. 100 Centennial Mall North, Lincoln, NE 68508, USA. Email: skye.wills@lin.usda.gov.*

11 *Thomas C. Edwards, Jr.^d U.S. Geological Survey, Utah Cooperative Fish and Wildlife Research Unit and Department*
12 *of Wildland Resources, Utah State University, Logan, UT 84322, USA. Email: t.edwards@usu.edu.*

13
14
15 **Abstract**

16
17 Mapping the spatial distribution of soil taxonomic classes is important for informing soil use and
18 management decisions. Digital soil mapping (DSM) can quantitatively predict the spatial distribution of
19 soil taxonomic classes. Key components of DSM are the method and the set of environmental covariates
20 used to predict soil classes. Machine learning is a general term for a broad set of statistical modeling
21 techniques. Many different machine learning models have been applied in the literature and there are
22 different approaches for selecting covariates for DSM. However, there is little guidance as to which, if
23 any, machine learning model and covariate set might be optimal for predicting soil classes across
24 different landscapes.

25 Our objective was to compare multiple machine learning models and covariate sets for predicting soil
26 taxonomic classes at three geographically distinct areas in the semi-arid western United States of
27 America (southern New Mexico, southwestern Utah, and northeastern Wyoming). All three areas were
28 the focus of digital soil mapping studies. Sampling sites at each study area were selected using
29 conditioned Latin hypercube sampling (cLHS). We compared models that had been used in other DSM
30 studies, including clustering algorithms, discriminant analysis, multinomial logistic regression, neural
31 networks, tree based methods, and support vector machine classifiers. Tested machine learning models
32 were divided into three groups based on model complexity: simple, moderate, and complex. We also

33 compared environmental covariates derived from digital elevation models and Landsat imagery that
34 were divided into three different sets: 1) covariates selected *a priori* by soil scientists familiar with each
35 area and used as input into cLHS, 2) the covariates in set 1 plus 113 additional covariates, and 3)
36 covariates selected using recursive feature elimination.

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

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

52 **Keywords:**

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

54

55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79

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

80 machines (Kovačević et al., 2010). Although machine learning models have been tested in different
81 landscapes around the world, it is rare for multiple models to be tested on the same landscape.

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

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

102 Model performance depends on the covariates used to represent soil-landscape relationships
103 and covariate selection is an important aspect of digital soil mapping (Vasques et al., 2012; Xiong et al.,

104 2014). Therefore, we also compared the influence of three groups of environmental covariates on
105 machine learning model performance in each of the three study areas: 1) covariates selected *a priori* by
106 soil scientists familiar with each area (expert knowledge; Zhu et al., 2001), 2) the covariates in set 1 plus
107 113 additional covariates derived from digital elevation models and Landsat imagery at several
108 resolutions that represented a large suite of potentially useful covariates, and 3) a subset of covariates
109 identified using recursive feature elimination (Guyon et al., 2002) from covariate sets 1 and 2.

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

115 116 **2. Methods**

117 118 *2.1. Study Areas*

119 120 *2.1.1 New Mexico (NM)*

121
122 The New Mexico (NM) study area is located on Otero Mesa in the northern reaches of the
123 Chihuahuan Desert, approximately 130 km northeast of El Paso, TX, USA. Centered at 105.6° W
124 longitude, 32.5° N latitude (Fig. 1), the area is approximately 190 km². The underlying geology is
125 primarily limestone and sandstone (Green and Jones, 1997). Soil parent material is primarily calcareous
126 alluvium but also includes eolian sands and residuum. Vegetation is a mix of shrublands (primarily
127 creosote bush [*Larrea tridentata*] and tar bush [*Florenxia cernua*]) and grasslands (primarily black grama
128 [*Boutaloua eriopoda*] and tobosa [*Pleuraphis mutica Buckley*]). Elevation ranges from 1430 to 1915 m.
129 The soil moisture regime is aridic bordering on ustic. Mean annual precipitation is 354 mm, the majority
130 of the precipitation arrives between June and December, and mean annual temperature is

131 approximately 15 °C (PRISM Climate Group, Oregon State University, <http://prism.oregonstate.edu/>,
132 accessed 4 March 2014).

133 134 2.1.2 Utah (UT)

135
136 The Utah (UT) study area is located in the eastern Great Basin physiographic province,
137 approximately 14 km southwest of Milford, UT, USA. Centered at 113° W longitude and 38° N latitude,
138 the area is approximately 300 km² and consists of mountainous terrain and associated alluvial fans
139 formed from a complex mix of limestone, dolomite, quartzite, basalt, quartz monzonite, quartz latite,
140 shale, sandstone, andesite, rhyolite, granite, and ash flows (Best et al., 1989). Elevation ranges from
141 1540 to 2100 m. Vegetation consists of shrubs (primarily Wyoming big sagebrush [*Artemisia tridentata*]
142 and black sagebrush [*Artemisia nova*]) and bunch grasses (Indian ricegrass [*Achnatherum hymenoides*])
143 at lower elevations, while trees (primarily Utah Juniper [*Juniperus osteosperma*] and Singleleaf Pinyon
144 [*Pinus monophylla*]) dominate higher elevations. The soil moisture regime is aridic bordering on xeric in
145 lower elevations and xeric in higher elevations. Mean annual temperature and precipitation for the
146 nearest weather station (Milford, UT) is 11°C and 200 mm, respectively, the majority of the precipitation
147 arrives in April and October (Western Regional Climate Center, 2013).

148 149 2.1.3 Wyoming (WY)

150
151 The Wyoming (WY) study area is located in the Powder River Basin of Wyoming, USA, part of the
152 Northern Rolling High Plains (United States Department of Agriculture, 2006), approximately 43 km
153 southwest from Gillette, WY. Centered at approximately 106° W longitude and 44° N latitude, the area is
154 approximately 296 km². Geology in the area consists of variegated mudstone, sandstone, conglomerate,
155 limestone, shale and coal (Cole and Boettinger, 2006; Green and Drouillard, 1994) . Topography is a mix
156 of bedrock-controlled, low rolling hills and badlands (locally known as the “Powder River Breaks”) a
157 system of steep, bedrock-controlled hills and gullies (gullies commonly > 6 m deep) with extremely high

158 rates of erosion and low vegetation cover (Cole, 2004). Vegetation is characterized by a mixture of mid-
159 stature cool season grasslands (bluebunch wheatgrass [*Pseudoroegneria spicata*] and needle-and-thread
160 [*Hesperostipa comata*]) and sagebrush shrublands (Wyoming big sagebrush [*Artemisia tridentata*])
161 (United States Department of Agriculture, 2006). Elevation ranges from 1220 and 1600 m. The soil
162 moisture regime is aridic bordering on ustic. Mean annual temperature and precipitation is 8°C and 310
163 mm, respectively, with the majority of the precipitation falling between April and October (Western
164 Regional Climate Center, 2013).

165 166 2.2 Sampling

167
168 Sampling locations for each study area were selected using conditioned Latin hypercube
169 sampling (cLHS) (Minasny and McBratney, 2006). Covariates used for input into cLHS were chosen by soil
170 scientists familiar with each study area and assumed to best represent soil-landscape relationships and
171 anticipated soil forming processes in each area (covariate set 1). The soil scientists who selected cLHS
172 input covariates for the NM study area had worked inside the study area and in similar landscapes for
173 approximately ten years. The soil scientist who selected cLHS input covariates for the UT study area had
174 visited the area, performed three months of field sampling in an adjacent area, and conducted a
175 literature review to identify important covariates in similar landscapes. The soil scientists who selected
176 cLHS input covariates for the WY area were Natural Resource Conservation Service (NRCS) soil scientists
177 who were conducting traditional soil surveys in similar landscapes around the study area.

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

183 included at the subgroup level. For each area, subgroup classes with only 1 observation were grouped
184 with the most similar subgroup class.

185 186 2.2.1 *New Mexico cLHS*

187
188 Covariates used for cLHS were derived from an October 2006 Landsat 5 TM image and a 5-m
189 Lidar digital elevation model (DEM). Imagery covariates from Landsat were band 5 (short wave infrared)
190 plus band 2 (green), band 5 minus band 2, and a normalized band 5/2 ratio ($[(\text{Band 5} - \text{Band 2}) / (\text{Band 5} + \text{Band 2})]$). Terrain attributes were aspect in degrees, elevation, slope, and a multipath wetness index
191 (Shi, 2013) calculated at four slope resolutions (5, 10, 25, 35 m) from the DEM. A categorical terrain
192 classification was also used. Imagery covariates were chosen for use in cLHS because they had been
193 shown to correlate with soil surface properties. Slope and the multipath wetness index, were chosen to
194 represent potential soil moisture distribution. Aspect and elevation were chosen to represent
195 microclimate and potential soil moisture (higher elevation, north-facing areas often have more potential
196 soil moisture than lower elevation, south-facing areas. The terrain classification consisted of seven
197 classes related to elevation and slope.

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

206 207 2.2.2 *Utah cLHS*

208

209 Covariates used for cLHS were derived from an atmospherically corrected (Chavez, 1996) July
210 31st 2000 Landsat 7 ETM+ image and a 10-m hydrologically correct DEM. A soil adjusted vegetation index
211 (SAVI) was derived from the imagery using an L value of 0.5 (Heute, 1988). Terrain attributes were slope,
212 inverse wetness index (Tarboton, 2013) and transformed aspect (a measure of northness vs. southness).
213 Land cover and geologic type were also used. Land cover type was obtained from the Southwest
214 Regional Gap Analysis Program (Lowry et al., 2007). Geology was obtained from a United States
215 Geological Survey 1:50,000 geology map (Best et al., 1989). Land cover and SAVI were chosen because it
216 was anticipated that vegetation type and density was correlated with soil properties such as soil depth.
217 Geologic type was chosen because the highly complex geology in this area was anticipated to exert a
218 strong control on potential pedogenesis. Terrain covariates were chosen to represent microclimate,
219 because microclimate heavily influences soil moisture, which in turn influences pedogenesis.

220 Three hundred locations were visited. Soil pedons were excavated, described, and classified to
221 family level. Subgroup classes were extracted from family names. Three soil observations were excluded
222 from modeling as they were located in highly disturbed areas. This resulted in 297 soil observations in
223 15 subgroup classes (Fig. 3, Table 1).

224 225 *2.2.3 Wyoming cLHS*

226
227 Covariates used for cLHS were derived from a Landsat 5 TM image and a 2-m Lidar DEM.
228 Imagery covariates were Normalized Difference Vegetation Index (NDVI) and band ratios 5/2 and 5/7.
229 Terrain derivatives were topographic wetness index, topographic position index, stream power index
230 (Wilson and Gallant, 2000) and distance to the nearest road. All covariates for cLHS, except distance to
231 the nearest road, were selected using the Optimum Index Factor (OIF). OIF identifies the combination of
232 input covariates that maximize variability, with the lowest correlation among covariates (Kienast-Brown
233 and Boettinger, 2010). Distance to the nearest road was included for a vegetation sampling project not
234 directly related to soil mapping.

235 Similar to the NM study area, cLHS was used to select hierarchical nested sets of 150, 100, and
236 50 potential sampling sites from 200 original sampling sites. Fifty-seven soil pedon observations were
237 made: the set of 50 nested cLHS samples plus an additional seven pedon observations (Fig. 4). Each soil
238 pedon was excavated, described, and assigned to a soil series. Subgroup classes were extracted from
239 each series using official soil series descriptions (<https://soilseries.sc.egov.usda.gov/osdname.asp>). This
240 resulted in 5 subgroup classes (Table 1).

241 242 *2.3 Additional covariates*

243
244 Additional terrain covariates were created from a 5-m Lidar derived DEM for the NM study area,
245 a 5-m auto-correlated DEM (Utah Automated Geographic Reference Center, 2013) for the UT study area
246 and resampling the 2-m WY Lidar DEM to 5-m. Terrain covariates were created in R (R Core Team, 2012)
247 with the RSAGA package (Brenning, 2008). For each area the following terrain covariates were created:
248 slope, total curvature, plan and profile curvature, SAGA wetness index, catchment area, catchment
249 slope, modified catchment area, convergence index, morphometric protection index (Yokoyama et al.,
250 2002), multi-resolution index of valley bottom flatness and multi-resolution index of ridge top flatness
251 (Gallant and Dowling, 2003), topographic position index, and terrain ruggedness index. Definitions of
252 individual terrain covariates can be found in Wilson and Gallant (2000) and Hengl and Reuter (2008).

253 Estimated potential direct, diffuse, total, and the duration of incoming solar radiation of the
254 approximate growing season in each area were also calculated. All potential incoming solar radiation
255 was calculated for clear sky and standard atmosphere conditions, and represent potential solar radiation
256 in the absence of clouds or significant amounts of atmospheric aerosols. All terrain and potential solar
257 radiation covariates were calculated at 5, 10, 30, 50, and 100 m cell sizes. Digital elevation models with
258 10, 30, 50, and 100 m cell sizes were created from the 5-m DEMs by averaging over blocks of cells at
259 these resolutions. The morphometric protection index calculated at 100-m cell size was not used

260 because at this resolution there was no variance in the covariate. This resulted in 89 terrain covariates
261 for each area.

262 For each area, we selected Landsat 5 TM imagery from 2 different dates. Each image pair
263 consisted of an image acquired during a season of peak vegetation growth and a season of dormant
264 vegetation. Each image was atmospherically corrected using the “Cost without Tau” method (Chavez,
265 1996) in the R Landsat package (Goslee, 2011). From each image the following covariates were created:
266 normalized band ratios 5/2, 5/7, 3/1, and 1/7; NDVI; six bands of the tasseled cap transformation (Crist
267 and Kauth, 1986); and greenness above bare soil (GRABS) index (Jensen, 2005). This resulted in 24
268 imagery covariates for each area. Total additional terrain and imagery covariates for each area were 113
269 (covariate set 2).

270 These covariates represent a wide range of topographic and spectral derivatives commonly used
271 for DSM in the western USA (Boettinger, 2010), but these additional covariates are not exhaustive of the
272 potentially available covariates. For example, in other DSM studies, Heung et al. (2014) included
273 distance to the nearest stream/river and relative hydrological slope position. Behrens et al. (2010) used
274 elevation differences from the center pixel of a DEM as predictor covariates. Xiong et al. (2012) used
275 covariates such as LANDFIRE (Landscape fire and resource management tools project) vegetation maps
276 and geospatial land cover maps as vegetation related covariates. Poggio et al., (2013) used multi-
277 temporal MODIS (Moderate Resolution Imaging Spectroradiometer) vegetation and drought indices.
278 Taylor et al. (2013) used potential evapotranspiration from ASTER (Advanced Spaceborne Thermal
279 Emission and Reflection Radiometer) imagery. Although a wide range of potential covariates exist, we
280 chose to incorporate the specific terrain and imagery covariates in covariate set 1+2, because they were
281 easily calculated with the available software with which we were familiar, and because we anticipated
282 these covariates to adequately characterize soil distribution in these areas. While relatively coarse-
283 resolution (3rd order soil survey; Soil Survey Division Staff, 1993) soil maps were available for the NM

284 and WY study areas (the UT area was previously unmapped), we did not include existing soil maps in
285 covariate set 1+2 for these areas in an effort to keep all covariate sets as consistent as possible.
286 Geological maps were not included in covariate set 1+2, because only a single geological unit was
287 mapped in the NM and WY areas.

288 289 *2.4 Covariate selection*

290
291 Recursive feature elimination (Guyon et al., 2002; Kuhn and Johnson, 2013) was used to identify
292 an optimal subset of covariates from the set of all available covariates (covariate set 1+2) for each area
293 (Fig. 5). Recursive feature elimination identifies optimal subsets (lowest misclassification error) of
294 predictor covariates by constructing a classification model with all predictor covariates, ranking each
295 predictor covariate, eliminating the covariate(s) with the lowest importance, and repeating this
296 procedure until a predefined threshold is reached or only one predictor covariate remains. Xiong et al.
297 (2012) used recursive feature elimination to identify important predictors for digital soil mapping of soil
298 carbon in Florida.

299 For each study area, random forests (Liaw and Wiener, 2002; parameters $mtry = \text{default}$ and
300 $ntree = 1000$) was used to calculate covariate importance, as random forests is not highly sensitive to
301 non-informative predictors (Kuhn and Johnson, 2013). Random forests identifies important covariates
302 by generating multiple classification trees (a forest) using bootstrap sampling, randomly scrambling the
303 covariates in each bootstrap sample and reclassifying the bootstrap sample. The misclassification error
304 of the bootstrap sample (termed the “out-of-bag” error) using the scrambled covariate is compared to
305 the misclassification error using the original covariate and the percent difference is used as a measure of
306 covariate importance (Peters et al., 2007). Important covariates will have a large increase in “out-of-
307 bag” error. For each area, the optimal subset of covariates was identified as the subset of covariates
308 with the minimum OOB error (Fig. 5).

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

313 314 *2.4 Modeling*

315 All modeling was performed using the caret package (Kuhn et al., 2013) in R (R Core Team,
316 2012). We tested eleven classification models for each area (Table 2). Each model was chosen based on
317 a review of machine learning methods used in other published DSM literature. Selected machine
318 learning models represented several broad classes of machine learning techniques and included
319 multinomial logistic regression, tree based classifiers, neural networks, support vector machines, and
320 clustering methods. An accessible explanation of all tested models can be found in Kuhn and Johnson
321 (2013) and James et al. (2014). Tested models were divided into three groups based on model
322 complexity: simple, moderate, and complex (Table 2). Models were assigned to one of the three groups
323 based on the interpretability and number of parameters of each model. Complex models (e.g., support
324 vector machines and neural networks) were difficult to interpret (i.e., black-box models) and had many
325 parameters. Simple models were interpretable and had few parameters, while medium complexity
326 models were between simple and complex models.

327
328 The goal of machine learning is to find a useful approximation of the function that underlies the
329 predictive relationship between input covariates and desired outcomes (Hastie et al., 2001). In this study
330 input covariates were derived from DEM's and Landsat imagery and the desired outcomes were
331 subgroup classes. Each type of model (e.g., support vector machines, neural networks) has specific and
332 different required parameters (referred to as tuning parameters) that control how the relationship
333 between input covariates and outcomes is defined. These parameters must be optimized to generate
334 the best "fit" possible between covariates and outcomes.

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

345 For each required model parameter (the number of required model parameters ranged
346 between 0 and 2) ten potential candidate values were defined. This resulted in an $n \times 10$ matrix of
347 potential model tuning parameters, where n = the number of required parameters. Models were tuned
348 using each set of parameters, and the average Kappa (Section 2.5) was calculated over the 100 repeated
349 training/test splits. Optimal parameters were chosen using the one-standard-error rule (James et al.,
350 2014), where the simplest (smallest) tuning values within one standard error of the tuning parameters
351 which produced the maximum kappa, were selected as optimal (Kuhn, 2008). For those models that did
352 not require tuning parameters (bagged classification tree, linear discriminant analysis) no optimization
353 was possible.

354 Each model was applied to three sets of covariates for each area: the soil scientist selected
355 covariates used as input into cLHS (covariate set 1), the covariates in set 1 plus the 113 additional terrain
356 and imagery covariates that we created (covariate sets 1 + 2), and those covariates that were selected
357 by recursive feature elimination from all available covariates (covariate set 3). Because some models
358 required covariates to have similar ranges (e.g., K-nearest neighbors), all environmental covariates were

359 centered and scaled to have mean = 0 and standard deviation = 1 before use. Multinomial logistic
 360 regression and linear discriminant analysis could not be fit using covariate set 1+2.

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

367 368 *2.5 Model accuracy comparison*

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

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

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

380 where r = number of taxonomic classes, n = number of observations in the test set, F_{ij} is the
 381 probability estimate that observation n_i belongs to class r_j , and E_{ij} is an indicator covariate such that $E_{ij} =$
 382 1 if n_i was the subgroup class and 0 otherwise. Brier scores range between zero and two, with lower

383 scores indicating better model performance. A Brier score of 1.25 indicates that each taxonomic class
 384 was predicted with the same probability. Brier scores for both linear and radial support vector machines
 385 were not calculated, because support vector machines require more than three observations per class
 386 to calculate probability estimates and several subgroup classes in each area had three or less
 387 observations (Table 1).

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

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

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

397 Where μ_{max} is the probability value of the class with the maximum probability at each cell and
 398 μ_{max-1} is the second largest probability value at the same cell. The confusion index ranges between 0 and
 399 1; high CI values indicate greater uncertainty in subgroup class predictions.

400 401 **3. Results** 402

403 Models built using covariate set 3 had the highest κ for all three study areas (Figs. 6, 8, & 10).
 404 The model with the highest average κ for the NM study area ($\kappa = 0.32 \pm 0.09$) was support vector
 405 machines using a radial basis function (SVMR; Fig. 6); however, t-tests indicated no significant difference
 406 in κ existed between SVMR and random forests (RF). Random forests (RF) had the highest average κ for
 407 both the UT ($\kappa = 0.19 \pm 0.06$; Fig. 8) and the WY study areas ($\kappa = 0.53 \pm 0.14$; Fig. 10). Kappa values for

408 the WY study area represent moderate agreement between observed and predicted subgroup classes,
409 while κ for the NM and UT study areas represent low agreement between observed and predicted
410 subgroup classes. The models with the highest κ also had the highest percent correctly classified (PCC)
411 for each area; PCC was $47 \pm 0.07\%$, $43 \pm 0.04\%$, and $72 \pm 0.08\%$, for the NM, UT, and WY study areas,
412 respectively.

413 Models constructed using covariate set 3 had the lowest Brier scores for the NM and WY study
414 areas (Figs. 7 & 11). The lowest Brier score for the UT area was obtained using covariate set 1+2 (Fig. 9),
415 but t-tests indicated no significant difference existed between models with the lowest Brier scores from
416 covariate set 1+2, and covariate sets 1 ($t = 5.34$, $df = 198$, $p\text{-value} = 2.537e-07$) and 3 ($t = -2.52$, $df = 198$,
417 $p\text{-value} = 0.0126$) for this area. Random forests (RF) was the model with the lowest average Brier score
418 for the NM ($BS = 0.70 \pm 0.05$; Fig. 7), UT (0.70 ± 0.01 ; Fig. 9) and WY study areas (0.46 ± 0.08 ; Fig. 11).

419 Average individual subgroup class producer's accuracy ranged between 0 and 86 percent (Table
420 1). The number of optimal covariates as determined by recursive feature elimination for each study area
421 ranged between six and ten and included terrain derivatives at multiple cell sizes as well as several
422 Landsat derivatives (Table 3). Spatial predictions using the model identified as the most accurate for
423 each area generally met expected soil-landform patterns (Figs. 12A, 13A, & 14A). Confusion index values
424 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.
425 13B), and 0.04 and 0.98 for the WY study area (Fig. 14B).

426

427 **4. Discussion**

428

429 *4.1 Model performance*

430

431 Random forests (RF) models using covariates selected by recursive feature elimination
432 (covariate set 3) were consistently the most accurate, or was among the most accurate, classifiers (had
433 the highest κ and lowest Brier score), between study areas and between covariate sets within each

434 study area (Figs. 5-10). Although, single-hidden-layer neural networks (NNET), multilayer-perceptron
435 neural networks (MLP), and nearest shrunken centroids (NSC) had slightly lower average Brier scores
436 than random forests (RF) for the UT study area (Fig. 9) the differences were minimal. The consistency of
437 random forests (RF) and covariate set 3 for producing the most accurate subgroup classifications across
438 all study areas is likely because random forests was used in the recursive feature elimination procedure,
439 which optimized covariates for subgroup class prediction (Section 2.4).

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

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

451 As the model with the highest classification accuracy for each study area is of most interest for
452 predicting soil subgroup classes we restrict further discussion to random forests models using covariate
453 set 3 when discussing differences in classification accuracy between study areas. Differences in
454 classification accuracy between study areas can be partially attributed to the number of soil subgroup
455 classes and the frequency distribution (the balance of observations between subgroup classes) of soil
456 pedon observations. The UT study area was the least accurately modeled, had the most soil subgroup
457 classes ($n = 15$), and the most skewed frequency distribution of soil pedon observations between

458 subgroup classes. Two subgroup classes for the UT study area contained approximately 70% of the total
459 soil pedon observations (Table 1). In contrast, the WY study area, the most accurately classified study
460 area, had the fewest soil subgroup classes ($n = 5$) and a somewhat more balanced soil pedon
461 observation distribution frequency. The classification accuracy, number of soil subgroup classes ($n = 10$)
462 and soil pedon observation distribution frequency for the NM study area was between those of the UT
463 and WY study areas. This suggests that overall classification accuracy will be highest when there are
464 many soil observations, few soil classes, and the frequency distribution of soil observations between
465 classes is approximately equal.

466 The frequency distribution of soil pedon observations heavily influenced individual subgroup
467 class accuracies (Table 1). In general, classes with lower sampling frequencies were modeled less
468 accurately. This finding is consistent with data presented by others (Barthold et al., 2013; Hengl et al.,
469 2007; Kim et al., 2012; Marchetti et al., 2011; Stum et al., 2010; Taghizadeh-Mehrjard et al., 2012) and is
470 likely because there are simply not enough observations to separate such classes in feature space.

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

476 The frequency distribution of soil pedon observations between subgroup types in a study area is
477 likely a result of the sampling strategy used to select sites. Conditioned Latin hypercube sampling is a
478 sampling method designed to identify sampling sites which represent the multivariate distribution of
479 input environmental covariates and assumes that the input environmental covariates are related to the
480 covariate of interest (Minasny and McBratney, 2006). Environmental covariates used as input to cLHS
481 for each study area were selected to represent broad soil-landscape relationships. Our results suggest

482 that in complex landscapes where likely many different soil types exist, such input environmental
483 covariates result in adequate sampling of the most frequent soil types, but not of rare soil types (e.g.,
484 the UT study area).

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

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

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

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

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

527 Modeling separate sub-areas might also decrease the number of taxonomic classes by reducing
528 the area and thus the number of soil types considered in a model. For example, it is likely that the
529 number of subgroup classes in one model would have been fewer had the UT study area been

530 segregated into uplands and alluvial fan sub-areas. Although such an approach would increase the
531 number of required models in proportion to the number of chosen sub-areas, this is theoretically
532 appealing as different pedo-geomorphic sub-areas are likely to have different relationships between
533 subgroup classes and environmental covariates (McBratney et al., 2003).

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

544 545 4.2 Covariate set comparison

546
547 Surprisingly, models using all available covariates (covariate set 1+2) were as accurate, or
548 slightly more accurate (higher κ , lower Brier scores), than models using the covariates selected by soil
549 scientists (covariate set 1) for each area (Figs. 6-10). As covariate set 1 was selected by soil scientists
550 anticipating how soil-landscape relationships would be best represented for modeling, the fact that this
551 covariate set did not result in the most accurate models suggests that soil scientists may be unable to *a*
552 *priori* identify optimal covariates for predicting taxonomic classes. In hindsight, this is not entirely
553 surprising given the complexity of soil taxonomic classes and the disparate kinds of knowledge needed
554 to predict these relationships *a priori*. Soil taxonomic classes represent multiple soil forming factors
555 operating over long periods of time (likely decades to millennia) at several scales. Thus choosing optimal

556 predictive covariates for modeling requires knowing both 1) how, and the scale at which, multiple soil
557 forming factors vary across the landscape to produce soil taxonomic classes and 2) how those factors
558 are best distinguished using spectral and topographic covariates. Being able to do both requires
559 extensive familiarity with the local landscape and an understanding of terrain modeling and remote
560 sensing. This suggests a pressing need for further investigation into relationships between specific
561 environmental covariates and soil forming processes.

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

577 The optimal covariates for the UT study area were related to topographic shading (diffuse
578 insolation), slope, slope position, and terrain ruggedness (Table 3). The UT area was highly variable in
579 topographic relief. This local topography strongly influences soil erosion and deposition as well as the

580 amount of incoming solar radiation, which in turn influences soil distribution (Beaudette and O'Geen,
581 2009). As the UT area had the greatest geologic complexity between the three study areas, it is
582 surprising that covariates related to geology (Landsat band ratios 5/2, 5/7) were not among those
583 identified as optimal. This may be because the influence of local topography exerted a stronger effect on
584 soil development than did the relatively larger scale influence of geology. In semi-arid steeply sloping
585 uplands and mountainous erosional landscapes, covariates related to soil erosion/deposition processes
586 and solar radiation may be the most useful for predicting soil distribution.

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

598 599 *4.3 Spatial predictions*

600 Spatial predictions of subgroup classes using the most accurate model visually correspond to
601 expected soil-landscape relationships for each study area (Figs. 12A, 13A, & 14A). For the NM and WY
602 study areas spatial predictions generally agree with published soil surveys (data not shown) although
603 our predictions show much finer spatial detail. For the NM study area (Fig. 12A), soils with a bedrock
604 contact (Lithic Ustic Haplocalcids) were predicted on steeply sloping uplands. Calcic Petrocalcids
605

606 (subsurface cemented CaCO_3) were predicted on older, stable alluvial surfaces. Ustic Haplocambids
607 (little soil development) were predicted on what are likely more active and recent geomorphic surfaces.
608 Petronodic Ustic Haplocalcids (subsurface CaCO_3 concretions, possibly approaching cementation) were
609 predicted on landforms intermediate between where Calcic Petrocalcids and Ustic Haplocambids were
610 predicted. Ustic Haplocalcids (subsurface CaCO_3 accumulation) were predicted to occur in an
611 intermingled pattern with Calcic Petrocalcids and Ustic Haplocambids, but may be over-predicted on
612 steeply sloping uplands. For the WY study area (Fig. 14A), both Ustic Torriorthents (minimal
613 development) and Badlands (steep hills and gullies) were predicted on steeply sloping, dissected
614 landforms near stream channels where active erosion may be occurring. Ustic Haplargids (subsurface
615 clay accumulation) were predicted on flatter, more stable upland surfaces that likely had enough time
616 for clay to form and/or translocate in the subsoil.

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

627 Spatial prediction uncertainty was generally lowest (lowest confusion index values) in relatively
628 low relief alluvial channels and run-in areas in the NM (Fig. 12B) and UT study areas (Fig. 13B), and in
629 lower relief portions of the WY area (Fig. 14B). This is likely because low relief areas had comparatively

630 low covariate complexity and suggests that soil taxonomic class prediction will be least uncertain in
631 relatively low relief areas.

632 633 **5. Conclusions**

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

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

650
651 **Acknowledgements:** This research was supported in large part by the USDI Bureau of Land Management
652 (BLM) via Cooperative Ecosystem Study Unit (CESU) Agreement number L09AC15757 and the Utah
653 Agricultural Experiment Station (UAES), Utah State University. Approved as UAES journal paper number
654 8655. We would like to thank Jeremiah Armentrout, Brook Fannesbeck, and the Fort Bliss military base
655 for providing the NM dataset. Brook Fannesbeck, Mike Leno, Zamir Libohova, Shawn Nield and Amanda

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

660

661

662 **6. References**

- 663
- 664 Barthold, F.K., Wiesmeier, M., Breuer, L., Frede, H.-G., Wu, J., Blank, F.B., 2013. Land use and climate
665 control the spatial distribution of soil types in the grasslands of Inner Mongolia. *J. Arid Environ.*
666 88, 194–205. doi:10.1016/j.jaridenv.2012.08.004
- 667
- 668 Beaudette, D.E., O’Geen, A.T., 2009. Quantifying the aspect effect: an application of solar radiation
669 modeling for soil survey. *Soil Sci. Soc. Am. J.* 73, 1345–1352. doi:10.2136/sssaj2008.0229
- 670
- 671 Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil
672 mapping using artificial neural networks. *J. Plant Nutr. Soil Sci.* 168, 21–33.
673 doi:10.1002/jpln.200421414
- 674
- 675 Behrens, T., Scholten, T., 2006. A Comparison of Data-Mining Techniques in Predictive Soil Mapping, in:
676 Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), *Digital Soil Mapping: An Introductory*
677 *Perspective*. Elsevier, Amsterdam, pp. 353 – 617.
- 678
- 679 Behrens, T., Schmidt, K., Zhu, A.X., Scholten, T., 2010. The ConMap approach for terrain-based digital soil
680 mapping. *Eur. J. Soil Sci.* 61, 133–143. doi:10.1111/j.1365-2389.2009.01205.x
- 681
- 682 Best, M.G., Lemmon, D.M., Morris, H.T., 1989. Geologic map of the Milford quadrangle and east half of
683 the Frisco quadrangle, Beaver county, Utah. *Miscellaneous Investigations Series Map I-1904*.
684 U.S. Geological Survey. Reston.
- 685

- 686 Bestelmeyer, B.T., Ward, J.P., Havstad, K.M., 2006. Soil-geomorphic heterogeneity governs patchy
687 vegetation dynamics at an arid ecotone. *Ecology* 87, 963–973. doi:10.1890/0012-
688 9658(2006)87[963:SHGPVD]2.0.CO;2
689
- 690 Boer, M., DelBarrio, G., Puigdefabregas, J., 1996. Mapping soil depth classes in dry Mediterranean areas
691 using terrain attributes derived from a digital elevation model. *Geoderma* 72, 99–118.
692 doi:10.1016/0016-7061(96)00024-9
693
- 694 Boettinger, J.L., 2010. Environmental covariates for digital soil mapping in the western USA, in:
695 Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), *Digital Soil*
696 *Mapping: Bridging Research, Environmental Application, and Operation*. Springer, Dordrecht,
697 pp. 17–27.
698
- 699 Boruvka, L., Pavlu, L., Vasat, R., Penizek, V., Drabek, O., 2008. Delineating acidified soils in the jizera
700 mountains region using fuzzy classification, in: Hartemink, A.E., McBratney, A., Mendonça-
701 Santos, M. de L. (Eds.), *Digital Soil Mapping with Limited Data*. Springer, Netherlands, pp. 303–
702 309.
703
- 704 Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide
705 susceptibility analysis with generalized additive models, in: *SAGA – Seconds Out (= Hamburger*
706 *Beitraege zur Physischen Geographie und Landschaftsoekologie, Vol. 19)*. J. Boehner, T.
707 Blaschke, L. Montanarella, pp. 23–32.
708

- 709 Brier, G.W., 1950. Verification of forecast expressed in terms of probability. *Mon. Weather Rev.* 78, 1–3.
710 doi:[http://dx.doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
711
- 712 Brungard, C.B., Boettinger, J.L., 2012. Spatial prediction of biological soil crust classes; value added DSM
713 from soil survey, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), *Digital Soil Assessments and*
714 *Beyond Proceedings of the 5th Global Workshop on Digital Soil Mapping*. CRC Press, Sydney, pp.
715 57–60.
716
- 717 Bui, E.N., Moran, C.J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example
718 from the Murray-Darling basin of Australia. *Geoderma* 111, 21–44. doi:10.1016/S0016-
719 7061(02)00238-0
720
- 721 Burrough, P.A., Gaans, P.F.M. van, Hootsmans, R., 1997. Continuous classification in soil survey: spatial
722 correlation, confusion and boundaries. *Geoderma* 77, 115 – 135.
723 doi:[http://dx.doi.org/10.1016/S0016-7061\(97\)00018-9](http://dx.doi.org/10.1016/S0016-7061(97)00018-9)
724
- 725 Campling, P., Gobin, A., Feyen, J., 2002. Logistic Modeling to spatially predict the probability of soil
726 drainage classes. *Soil Sci. Soc. Am. J.* 66, 1390–1401.
727
- 728 Chavez Jr, P.S., 1996. Image-Based Atmospheric Corrections - Revisited and Improved. *Photogramm.*
729 *Eng. Remote Sens.* 62, 1025–1036.
730
- 731 Cole, N.J., 2004. A pedogenic understanding raster classification model for mapping soils, Powder River
732 Basin, Wyoming (MS Thesis). Utah State University, Logan, USA.

- 733
- 734 Cole, N.J., Boettinger, J.L., 2006. Pedogenic understanding raster classification methodology for mapping
735 soils, Powder River Basin, Wyoming, USA, in: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.),
736 Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam, pp. 377–619.
- 737
- 738 Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data.
739 Remote Sens. Environ. 37, 35–46. doi:10.1016/0034-4257(91)90048-B
- 740
- 741 Congalton, R.G., Green, K., 1998. Assessing the Accuracy of Remotely Sensed Data: Principles and
742 Practices. CRC/Taylor & Francis, Boca Raton.
- 743
- 744 Crist, E., Kauth, R., 1986. The tasseled cap de-mystified. Photogramm. Eng. Remote Sens. 52, 81–86.
- 745
- 746 Duniway, M.C., Herrick, J.E., Monger, H.C., 2010. Spatial and temporal variability of plant-available water
747 in calcium carbonate-cemented soils and consequences for arid ecosystem resilience. Oecologia
748 163, 215–226. doi:10.1007/s00442-009-1530-7
- 749
- 750 Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping
751 depositional areas. Water Resour. Res. 39. doi:10.1029/2002WR001426
- 752
- 753 Gile, L.H., Hawley, J.W., Grossman, R.B., 1981. Memoir 39—Soils and geomorphology in the Basin and
754 Range area of southern New Mexico: Guidebook to the Desert Project. New Mexico Bureau of
755 Mines and Mineral Resources, Socorro.
- 756

- 757 Goslee, S.C., 2011. Analyzing remote sensing data in R: the landsat package. *J. Stat. Softw.* 43, 1–25.
758
- 759 Green, G.N., Drouillard, P.H., 1994. The Digital Geologic Map of Wyoming in ARC/INFO Format. U.S.
760 Geological Survey Open-File Report 94-0425. [http://geo-nsdi.er.usgs.gov/metadata/open-](http://geo-nsdi.er.usgs.gov/metadata/open-file/94-425/)
761 [file/94-425/](http://geo-nsdi.er.usgs.gov/metadata/open-file/94-425/) (last accessed: 7/5/2014).
762
- 763 Green, G.N., Jones, G.E., 1997. The Digital Geologic Map of New Mexico in ARC/INFO Format. U.S.
764 Geological Survey Open File Report 97-0052. [http://pubs.usgs.gov/of/1997/ofr-97-](http://pubs.usgs.gov/of/1997/ofr-97-0052/new_mex.htm)
765 [0052/new_mex.htm](http://pubs.usgs.gov/of/1997/ofr-97-0052/new_mex.htm) (last accessed: 7/5/2014).
766
- 767 Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an
768 existing soil map: sampling intensity, validation procedures, and integration of spatial context.
769 *Geoderma* 143, 180–190. doi:10.1016/j.geoderma.2007.11.004
770
- 771 Grunwald, S., 2006. Environmental Soil-Landscape Modeling: Geographic Information Technologies and
772 Pedometrics. CRC/Taylor & Francis, Boca Raton.
773
- 774 Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support
775 vector machines. *Mach. Learn.* 46, 389–422. doi:10.1023/A:1012487302797
776
- 777 Hastie, T., Tibshirani, R., Friedman, J.H., 2001. *The Elements of Statistical Learning: Data Mining,*
778 *Inference, and Prediction.* Springer, New York.
779

- 780 Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties
781 using decision trees. *Geoderma* 124, 383–398. doi:10.1016/j.geoderma.2004.06.007
782
- 783 Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. Methods to interpolate soil categorical
784 variables from profile observations: lessons from Iran. *Geoderma* 140, 417–427.
785 doi:10.1016/j.geoderma.2007.04.022
786
- 787 Hengl, T., Reuter, H.I., 2008. *Geomorphometry. Concepts, Software, Applications. Developments in Soil*
788 *Science*. Elsevier, Amsterdam.
789
- 790 Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale:
791 A Random Forest approach. *Geoderma* 214–215, 141–154.
792 doi:10.1016/j.geoderma.2013.09.016
793
- 794 Heute, A.R., 1988. A Soil-Adjusted Vegetation Index (SAVI). *Remote Sens. Environ.* 25, 295–309.
795
- 796 Inzana, J., Kusky, T., Higgs, G., Tucker, R., 2003. Supervised classifications of Landsat TM band ratio
797 images and Landsat. *J. Afr. Earth Sci.* 37, 59–72. doi:10.1016/S0899-5362(03)00071-X
798
- 799 Jafari, A., Finke, P.A., Van de Wauw, J., Ayoubi, S., Khademi, H., 2012. Spatial prediction of USDA- great
800 soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict
801 diagnostic horizons and soil types. *Eur. J. Soil Sci.* 63, 284–298. doi:10.1111/j.1365-
802 2389.2012.01425.x
803

- 804 Jafari, A., Ayoubi, S., Khademi, H., Finke, P.A., Toomanian, N., 2013. Selection of a taxonomic level for
805 soil mapping using diversity and map purity indices: a case study from an Iranian arid region.
806 *Geomorphology* 201, 86–97. doi:10.1016/j.geomorph.2013.06.010
807
- 808 James, G., Witten, D., Hastie, T., Tibshirani, R., 2014. *An Introduction to Statistical Learning: with*
809 *Applications in R*. Springer, New York.
810
- 811 Jensen, J.R., 2005. *Introductory Digital Image Processing: A Remote Sensing Perspective*. Pearson
812 Prentice Hall, Upper Saddle River.
813
- 814 Johansson, U., König, R., Niklasson, L., 2010. Genetic rule extraction optimizing brier score, in: Pelikan,
815 M., Branke, J. (Eds.), *GECCO*. ACM, pp. 1007–1014.
816 http://bada.hb.se/bitstream/2320/6795/1/Gecco2010_GREX_Optimizing_Brier_Score.pdf
817
- 818 Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012. Soil type mapping using the generalised linear
819 geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma* 189, 540–553.
820 doi:10.1016/j.geoderma.2012.05.028
821
- 822 Kienast-Brown, S., Boettinger, J.L., 2010. Applying the optimum index factor to multiple data types in soil
823 survey, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.),
824 *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Springer,
825 Dordrecht, pp. 385–398.
826

- 827 Kim, J., Grunwald, S., Rivero, R.G., Robbins, R., 2012. Multi-scale modeling of soil series using remote
828 sensing in a wetland ecosystem. *Soil Sci. Soc. Am. J.* 76, 2327–2341. doi:10.2136/sssaj2012.0043
829
- 830 Kovačević, M., Bajat, B., Gajić, B., 2010. Soil type classification and estimation of soil properties using
831 support vector machines. *Geoderma* 154, 340–347. doi:10.1016/j.geoderma.2009.11.005
832
- 833 Kuhn, M., 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26.
834
- 835 Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., A. Engelhardt. 2012. caret: Classification and
836 Regression Training. R package version 5.17-7. <http://CRAN.R-project.org/package=caret>
837
- 838 Kuhn, M., Johnson, K., 2013. *Applied Predictive Modeling*. Springer, New York.
839
- 840 Kuhn, M., 2014. A short introduction to the caret package. [http://cran.r-](http://cran.r-project.org/web/packages/caret/vignettes/caret.pdf)
841 [project.org/web/packages/caret/vignettes/caret.pdf](http://cran.r-project.org/web/packages/caret/vignettes/caret.pdf) (last accessed: 7/5/2014).
842
- 843 Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News* 2, 18–22.
844
- 845 Liu, J., Pattey, E., Nolin, M.C., Miller, J.R., Ka, O., 2008. Mapping within-field soil drainage using remote
846 sensing, DEM and apparent soil electrical conductivity. *Geoderma* 143, 261 – 272.
847 doi:<http://dx.doi.org/10.1016/j.geoderma.2007.11.011>
848
- 849 Lowry, J., Ramsey, R.D., Thomas, K., Schrupp, D., Sajwaj, T., Kirby, J., Waller, E., Schrader, S., Falzarano,
850 S., Langs, L., Manis, G., Wallace, C., Schulz, K., Comer, P., Pohs, K., Rieth, W., Velasquez, C., Wolk,

- 851 B., Kepner, W., Boykin, K., O'Brien, L., Bradford, D., Thompson, B., Prior-Magee, J., 2007.
852 Mapping moderate-scale land-cover over very large geographic areas within a collaborative
853 framework: a case study of the Southwest Regional Gap Analysis Project (SWReGAP). *Remote*
854 *Sens. Environ.* 108, 59 – 73. doi:<http://dx.doi.org/10.1016/j.rse.2006.11.008>
855
- 856 Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil
857 property rasters from legacy soil maps and from point data. *Geoderma* 232–234, 34–44.
858 doi:[10.1016/j.geoderma.2014.04.033](https://doi.org/10.1016/j.geoderma.2014.04.033)
859
- 860 Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., Francaviglia, R., 2011. Simulation of soil types in
861 Teramo province (Central Italy) with terrain parameters and remote sensing data. *CATENA* 85,
862 267–273. doi:[10.1016/j.catena.2011.01.012](https://doi.org/10.1016/j.catena.2011.01.012)
863
- 864 McBratney, A.B., Mendonça Santos, M.D.L., Minasny, B., 2003. On digital soil mapping. *Geoderma* 117,
865 3–52. doi:[10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
866
- 867 Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence
868 of ancillary information. *Comput. Geosci.* 32, 1378–1388. doi:[10.1016/j.cageo.2005.12.009](https://doi.org/10.1016/j.cageo.2005.12.009)
869
- 870 Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital
871 mapping of soil classes. *Geoderma* 142, 285–293. doi:[10.1016/j.geoderma.2007.08.022](https://doi.org/10.1016/j.geoderma.2007.08.022)
872
- 873 Moonjun, R., Farshad, A., Shrestha, D.P., Vaiphasa, C., 2010. Artificial Neural Network and Decision Tree
874 in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand, in: Boettinger, J.L., Howell,

- 875 D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging
876 Research, Environmental Application, and Operation, Springer, Dordrecht, pp. 151–164.
877
- 878 Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. I. Soil layer classes.
879 Geoderma 163, 38–44. doi:10.1016/j.geoderma.2011.03.014
880
- 881 Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.B., Bogaert, P., 2014.
882 Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the
883 loess derived soils of northern Iran. Geoderma 232–234, 97–106.
884 doi:10.1016/j.geoderma.2014.04.036
885
- 886 Peters, J., Baets, B.D., Verhoest, N.E.C., Samson, R., Degroeve, S., Becker, P.D., Huybrechts, W., 2007.
887 Random forests as a tool for ecohydrological distribution modelling. Ecol. Model. 207, 304–318.
888 doi:10.1016/j.ecolmodel.2007.05.011
889
- 890 Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their
891 uncertainty with a large number of satellite-derived covariates. Geoderma 209–210, 1–14.
892 doi:10.1016/j.geoderma.2013.05.029
893
- 894 R Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical
895 Computing, Vienna, Austria. <http://www.R-project.org/>.
896

- 897 Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D. (Eds.), 2003. Field Book for
898 Describing and Sampling Soils, Version 2.0. Natural Resources Conservation Service. National
899 Soil Survey Center, Lincoln.
900
- 901 Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type
902 prediction in a desert landscape. *Ecol. Model.* 181, 1–15. doi:10.1016/j.ecolmodel.2004.06.036
903
- 904 Shi, X., Long, R., Dekett, R., Philippe, J., 2009. Integrating different types of knowledge for digital soil
905 mapping. *Soil Sci. Soc. Am. J.* 73, 1682. doi:10.2136/sssaj2007.0158
906
- 907 Shi, X., 2013. ArcSIE user's guide. http://www.arcsie.com/Download/ArcSIE_UsersGuide_130319.pdf
908 (last accessed: 7/5/2014).
909
- 910 Smith, C.A.S., Daneshfar, B., Frank, G., 2012. Use of weights of evidence statistics to define inference
911 rules to disaggregate soil survey maps, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.),
912 Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil
913 Mapping. CRC Press, Sydney, pp. 215–220.
914
- 915 Soil Survey Staff, 1999. Soil taxonomy: a basic system of soil classification for making and interpreting
916 soil surveys, 2nd ed. U.S. Department of Agriculture Handbook 436. Natural Resources
917 Conservation Service, Lincoln.
918
- 919 Stum, A.K., Boettinger, J.L., White, M.A., Ramsey, R.D., 2010. Random Forests Applied as a Soil Spatial
920 Predictive Model in Arid Utah, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E.,

- 921 Kienast-Brown, S. (Eds.), *Digital Soil Mapping: Bridging Research, Environmental Application, and*
922 *Operation*, Springer, Dordrecht, pp. 179–190.
- 923
- 924 Subburayalu, S.K., Slater, B.K., 2013. Soil series mapping by knowledge discovery from an Ohio county
925 soil map. *Soil Sci. Soc. Am. J.* 77, 1254–1268. doi:10.2136/sssaj2012.0321
- 926
- 927 Taghizadeh-Mehrjard, R., Minasny, B., McBratney, A.B., Triantafilis, J., Sarmadian, F., Toomanian, N.,
928 2012. Digital soil mapping of soil classes using decision trees in central Iran, in: Minasny, B.,
929 Malone, B.P., McBratney, A. (Eds.), *Digital Soil Assessments and Beyond: Proceedings of the 5th*
930 *Global Workshop on Digital Soil Mapping*. CRC Press, Sydney, pp. 197–202.
- 931
- 932 Tarboton, D., 2013. *Terrain Analysis Using Digital Elevation Models (TauDEM)*.
933 <http://hydrology.usu.edu/taudem/taudem5/index.html> (last accessed: 7/5/2014).
- 934
- 935 Taylor, J.A., Jacob, F., Galleguillos, M., Prévot, L., Guix, N., Lagacherie, P., 2013. The utility of remotely-
936 sensed vegetative and terrain covariates at different spatial resolutions in modelling soil and
937 watertable depth (for digital soil mapping). *Geoderma* 193–194, 83–93.
938 doi:10.1016/j.geoderma.2012.09.009
- 939
- 940 Tesfa, T.K., Tarboton, D.G., Chandler, D.G., McNamara, J.P., 2009. Modeling soil depth from topographic
941 and land cover attributes. *Water Resour. Res.* 45, W10438. doi:10.1029/2008WR007474
- 942
- 943 Triantafilis, J., Earl, N.Y., Gibbs, I.D., 2012. Digital soil-class mapping across the Edgeroi district usings
944 numerical clustering and gamma-ray spectrometry data, in: Minasny, B., Malone, B.P.,

- 945 McBratney, A. (Eds.), *Digital Soil Assessments and Beyond: Proceedings of the 5th Global*
946 *Workshop on Digital Soil Mapping*. CRC Press, Sydney, pp. 187–191.
947
- 948 United States Department of Agriculture, 2006. *Land Resource Regions and Major Land Resource Areas*
949 *of the United States, the Caribbean, and the Pacific Basin*. [ftp://ftp-](ftp://ftp-fc.sc.egov.usda.gov/NSSC/Ag_Handbook_296/Handbook_296_low.pdf)
950 [fc.sc.egov.usda.gov/NSSC/Ag_Handbook_296/Handbook_296_low.pdf](ftp://ftp-fc.sc.egov.usda.gov/NSSC/Ag_Handbook_296/Handbook_296_low.pdf) (last accessed:
951 7/5/2014).
952
- 953 Utah Automated Geographic Reference Center, 2013. *5 Meter Auto-Correlated Elevation Models*.
954 <http://gis.utah.gov/data/elevation-terrain-data/5-meter-auto-correlated-elevation-models/>
955 (last accessed: 7/5/2014).
956
- 957 Van Zijl, G.M., le Roux, P.A.L., Smith, A.B., 2012. Rapid soil mapping under restrictive conditions in Tete,
958 Mozambique, in: Minasny, B., Malone, B.P., McBratney, A. (Eds.), *Digital Soil Assessments and*
959 *Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping*. CRC Press, Sydney, pp.
960 335–339.
961
- 962 Vasques, G.M., Grunwald, S., Myers, D.B., 2012. Associations between soil carbon and ecological
963 landscape variables at escalating spatial scales in Florida, USA. *Landsc. Ecol.* 27, 355–367.
964 doi:10.1007/s10980-011-9702-3
965
- 966 Webster, R., Welham, S.J., Potts, J.M., Oliver, M.A., 2006. Estimating the spatial scales of regionalized
967 variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood.
968 *Comput. Geosci.* 32, 1320–1333. doi:10.1016/j.cageo.2005.12.002

- 969
- 970 Western Regional Climate Center, 2013. Cooperative Climatological Data Summaries.
971 <http://www.wrcc.dri.edu/climatedata/climsum/> (last accessed: 7/5/2014).
972
- 973 Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. John Wiley & Sons, New
974 York.
- 975
- 976 Witten, I.H., Frank, E., Hall, M.A., 2011. Data mining: Practical Machine Learning Tools and Techniques.
977 Morgan Kaufmann, Burlington.
- 978
- 979 Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2012. Which soil,
980 environmental and anthropogenic covariates for soil carbon models in Florida are needed? In:
981 Minasny, B., Malone, B.P., McBratney, A. (Eds.), Digital Soil Assessments and Beyond:
982 Proceedings of the 5th Global Workshop on Digital Soil Mapping. CRC Press, Sydney, pp. 335–
983 339.
- 984
- 985 Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental
986 soil-landscape modeling of soil organic carbon. *Environ. Model. Softw.* 57, 202–215.
987 doi:10.1016/j.envsoft.2014.03.004
- 988
- 989 Yokoyama, R., Shirasawa, M., Pike, R.J., 2002. Visualizing topography by openness: a new application of
990 image processing to digital elevation models. *Photogramm. Eng. Remote Sens.* 68, 257–266.
991

992 Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil mapping using GIS, expert knowledge,
993 and fuzzy logic. *Soil Sci. Soc. Am. J.* 65, 1463–1472.
994
995
996

997 **Figure Captions:**

998

999 Fig. 1. Study area locations in western USA.

1000

1001 Fig. 2. Spatial distribution of pedon observation locations in the NM study area overlain on google
1002 physical map. Total number of pedon observations was 103.

1003

1004 Fig. 3. Spatial distribution of pedon observation locations in the UT study area overlain on google
1005 physical map. Total number of pedon observations was 297.

1006

1007 Fig. 4. Spatial distribution of pedon observation locations in the WY study area overlain on google
1008 physical map. Total number of pedon observations was 57.

1009

1010 Fig. 5. Optimal covariate subset selection using recursive feature elimination. Out-of-bag (OOB) error is
1011 random forests misclassification error. Random forests models were begun with the total available
1012 covariates and the least important covariate was iteratively removed. Optimal covariate subsets were
1013 selected as those covariates that returned the lowest OOB error and which had the fewest covariates.
1014 Arrows indicate optimal covariate subset.

1015

1016 Fig. 6. Average κ for the NM study area. Model with highest κ is the most accurate classifier. Error bars
1017 are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification
1018 Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA),
1019 Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer-Perceptron
1020 Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines
1021 (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1022

1023 Fig. 7. Average Brier scores for the NM study area. Model with lowest Brier score is the most accurate
1024 classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows:
1025 Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant
1026 Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR),
1027 Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support
1028 Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1029

1030 Fig. 8. Average κ for the UT study area. Model with highest κ is the most accurate classifier. Error bars
1031 are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification
1032 Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA),
1033 Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron
1034 Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines
1035 (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1036

1037 Fig. 9. Average Brier scores for the UT study area. Model with lowest Brier score is the most accurate
1038 classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows:
1039 Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant
1040 Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR),
1041 Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support
1042 Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1043

1044 Fig. 10. Average κ for the WY study area. Model with highest κ is the most accurate classifier. Error bars
1045 are 95% confidence intervals from cross validation. Abbreviations are as follows: Bagged Classification
1046 Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA),
1047 Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR), Multilayer Perceptron
1048 Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support Vector Machines
1049 (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1050
1051 Fig. 11. Average Brier scores for the WY study area. Model with lowest Brier score is the most accurate
1052 classifier. Error bars are 95% confidence intervals from cross validation. Abbreviations are as follows:
1053 Bagged Classification Tree (BCT), Classification Tree (CT), K Nearest Neighbors (KNN), Linear Discriminant
1054 Analysis (LDA), Linear Support Vector Machines (SVML), Multinomial Logistic Regression (MLR),
1055 Multilayer Perceptron Neural Network (MLP), Nearest Shrunken Centroids (NSC), Radial-Basis Support
1056 Vector Machines (SVMR), Random Forests (RF), Single-Hidden-Layer Neural Networks (NNET).

1057
1058 Fig. 12. Spatial predictions of subgroup classes (A), and the confusion index (B) for the NM study area
1059 using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown
1060 (5 of 10 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions;
1061 values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.

1062
1063 Fig. 13. Spatial predictions of subgroup classes (A), and the confusion index (B) for the UT study area
1064 using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown
1065 (5 of 15 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions;
1066 values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.

1067
1068 Fig. 14. Spatial predictions of subgroup classes (A), and the confusion index (B) for the WY study area
1069 using random forests (RF) and covariate set 3. Only predicted subgroups visible at this scale are shown
1070 (3 of 5 subgroups). Confusion index values near zero indicate low uncertainty in spatial predictions;
1071 values near one indicate high uncertainty in spatial predictions. Both images are overlain on a hillshade.