Modeling Rare Endemic Shrub Habitat in the Uinta Basin Using Soil, Spectral, and Topographic Data

Julie B. Baker  
NRCS Sonora Soil Survey Office  
19777 Greenley Rd.  
Sonora, CA 95370

Brook B. Fonnesbeck  
Plants, Soils, and Climate Dep.  
Utah State Univ.  
4820 Old Main Hill  
Logan, UT 84322

Janis L. Boettinger*  
Plants, Soils, and Climate Dep.  
Utah State Univ.  
4820 Old Main Hill  
Logan, UT 84322

Conserving rare plants is important to biodiversity. Shrubby reed-mustard ([Schoenocrambe suffrutescens (Rollins) S.L. Welsh and Chatterly] (SRM), a shrub endemic to the Uinta Basin, faces habitat loss due to energy development of fossil fuels. Little is known about the soil and site properties required for successful establishment, growth, and survival of SRM. Our goal was to create random forests models to identify potential SRM habitat and to verify habitat suitability in the field with easily measured properties. The soil properties (SP) model indicated that CaCO$_3$ equivalent, silt, and dry color value predicted the presence or absence of SRM with 10% out-of-bag (OOB) error. Soil properties were correlated with Landsat 5 Thematic Mapper spectral data and a digital elevation model in a Soil–Spectral Correlation (SSC) model. Yellowness, 3/2 normalized difference ratio (NDR), and 3/1 NDR were most strongly correlated with CaCO$_3$ equivalent and silt, and they predicted SRM habitat with 28% OOB error. Guided by SSC model output, 250 additional field presence or absence points were used to train the Spectral–Topographic (ST) model, which was validated by an independent set of SRM plant locations. The ST model using area solar radiation, 7/2 NDR, 5/2 NDR, and the normalized difference vegetation index gave an OOB error of 23%. The ST model can be used to identify potential habitat across a large area. Once identified, easily measured soil and site data from the SP model can verify SRM habitat suitability. These models can help land managers locate and check key soil properties at potential SRM habitat sites.

Core Ideas

• Soil and site properties predict rare plant distribution.
• Rare plant habitat can be predicted by random forests modeling of data.
• CaCO$_3$ equivalent, silt, and dry color value predict SRM presence or absence.
• New rare plant locations identified by random forests modeling.
• Models help land managers locate and identify potential habitat.

Preserving populations of rare plants is important to maintain biodiversity, which increases genetic diversity, provides possible sources of food and medicinal plants, and provides ecosystem services such as protection of soil and water resources, nutrient cycling, and pollutant adsorption or uptake (Australian Department of the Environment, 1993). A species may be rare for a variety of reasons, including low tolerance to environmental change, preference for a very specific habitat, limited dispersal capabilities, etc. Endemic plants are often characterized by few populations concentrated in a narrow geographic distribution. Soil and geologic properties have been successfully used to explain the distribution of some rare endemic plants. For example, endemic species grow in soils formed on serpentinite with low Ca/Mg ratios and high levels of heavy metals (Cr, Co, Ni) inherited from the ultramafic parent material (Alexander et al., 2007; Lazarus et al., 2011; Proctor, 1971; van der Ent and Reeves, 2015).

Rare, endemic plants in Utah and the surrounding states frequently occur in unusual edaphic environments that can have distinct plant communities. In many
cases, rare plants occur in soils that are considered physically harsh for most common desert plants. For example, *Cycladenia humilis* var. *jonesii* (Eastw.) S.L. Welsh & N.D. Atwood occurs on rocky (80 to almost 100% volume of rock fragments), shallow soils (<50 cm deep) formed in shale or mudstone in the northern Colorado Plateau of Utah. Reported to be an obligate gypsumophile (Welsh et al., 1987), populations of *C. humilis* var. *jonesii* were shown to occur in both gysiferous and non-gysiferous soils (Boettinger and Sipes, 1998). Similarly, *Penstemon debilis* O’Kane & J.L. Anderson occurred only on shallow, rocky soils (<50 cm deep) formed in shale on steep slopes subject to soil creep in the eastern Colorado Plateau of Colorado (McMullen, 1998). *Arctomecon californica* Torr. & Frém. and *Eriogonum corystous* var. *nilesii* Reveel occurred in very highly calcareous soils that were either shallow to a layer cemented with CaCO₃ or had >50% volume of cemented carbonate nodules in the Mojave Desert of southern Nevada (MacMahon et al., 2008).

*Schoenocrambe suffrutescens* (shrubby reed-mustard, SRM), is a federally listed endangered, low-growing (10–25 cm), perennial shrub endemic to Duchesne and Uintah counties within the Uinta Basin, Utah (Fig. 1). Shrubby reed-mustard faces habitat loss and fragmentation due to development and expansion of natural gas and oil extraction, including construction of roads and well pads. Although shrubby reed-mustard has long been observed in particular semi-barren shale strata of the Green River formation (Eocene), little is known about the ideal range of soil and landscape characteristics associated with its successful establishment, growth, and survival. Shrubby reed-mustard populations have been observed to occur on the Evacuation Creek member of the Green River Formation, the uppermost member (Cashion, 1959). Both the Evacuation Creek member and the Parachute Creek member directly below it contain oil shale strata rich in oil and natural gas deposits that are currently under pressure for natural resource recovery. Because of changes in nomenclature and difficulties in distinguishing the Evacuation and Parachute Creek members (now both assigned to an “upper member” of the Green River formation [Cashion and Donnell, 1974; Weiss et al., 1990; Remy, 1992]), geologic mapping alone is not a sufficient criterion for identification of current and potential SRM habitat.

Species distribution models are a range of models used to gather insight about ecological questions, including distribution, niche or habitat characterization, and resource selection, and to predict spatial distributions across landscapes of varying scales (Elith and Leathwick, 2009). Observations of species occurrence are combined with environmental data to predict occurrence or abundance. With the advent of improved algorithms, computing power, and environmental and spatial data, predictive applications of species distribution models have become more common in recent years and are often utilized in conservation and management decisions (Guillera-Arroita et al., 2015; Guisan et al., 2013). Distributions can be predicted by interpolation within the measured spatial, temporal, or ecological framework or extrapolation outside the known framework, for example climate change scenarios under different environmental conditions or invasive species colonization of new spatial extents. The models are generally improved by the addition of absence data (Guillera-Arroita et al., 2015), which may also aid in understanding the underlying environmental factors affecting species distributions.

Statistical models such as random forests, support vector machines, and logistic regression have been useful for ecological characterizations of species distributions with high degrees of accuracy (Lawrence et al., 2006). Random forests models have been applied to modeling the presence and absence of invasive plants and rare lichens and outperform other models (classification trees, logistic regression, linear discriminant analysis) in several measures of accuracy (Cutler et al., 2007). In addition, random forests models have been applied to predict soil classes in semi-arid environments and perform as well as or better than other model types in all cases (Brungard et al., 2015). Random forests, developed for ecological applications by Breiman and Cutler (Breiman, 2001; Cutler et al., 2007; Breiman and Cutler, 2009), is a machine learning algorithm using decision trees to identify or predict classes of data. In random forests, many computationally light trees are grown using a subset of variables at each split, with the strongest variable selected to split the data, rather than one tree using all the available data. A random subset with replacement (bootstrap) is used to train the model, and the remaining data are used to validate the model. An out-of-bag (OOB) error is calculated as the number of misclassified cases divided by the total number of cases (Stum et al., 2010). The number of variables selected at each split is typically the square root of the total number of samples (\(\sqrt{n}\)) and, on average, one-third of cases are not selected for each bootstrap sample—left “out of the bag” (Breiman, 2001). Random forests is considered to be doubly random, because of the random bootstrap sample and the random subset of variables selected at each split. Because hundreds to thousands of weak trees are grown for each analysis, patterns in the data may be discerned that would not be apparent

![Fig. 1. Shrubby reed-mustard in bloom, occurring in a shale bed (white) of the Green River Formation, Johnson Draw population, Uinta Basin, Utah. A sandstone bed (red) is visible in the upper left of the photo.](image-url)
with a few computationally robust classifications. Random forests models have been shown to be highly accurate and applicable to ecological questions because of their ability to determine variable importance, model complex interactions among predictor variables, and deal with missing values (Cutler et al., 2007).

As outlined in the 2010 SRM recovery plan (US Fish and Wildlife Service, 2010), research was needed to characterize soil and landscape properties for identifying potentially suitable habitat for SRM conservation and possible reintroduction. The aims of this study were to quantify environmental variables that relate to species occurrence, correlate these soil and site environmental variables with spectral remotely sensed and topographic data, and then model potential species habitat. In applications involving rare or endangered plants, potential habitat may be a more useful tool than distribution alone for conservation and management decisions, especially where loss of habitat is an issue. In this study, we developed three models that use soil, spectral, and topographic data to predict the potential habitat of SRM in the Uinta Basin. The first objective was to use a statistical model, specifically random forests (Breiman, 2001), to link SRM presence or absence to soil and site properties as simple field metrics to characterize potential habitat. The second objective was to use random forests to classify SRM presence or absence using remotely sensed spectral imagery and topography from a digital elevation model (DEM) and then select variables that can be correlated to soil and site characteristics. The third objective was to use random forests to spatially predict potential SRM habitat in the Uinta Basin based on spectral and topographic data. Our ultimate goal was to create a final map product to facilitate the identification of potential SRM habitat and to verify the suitability of potential habitat in a field setting with easily measured soil and site properties.

MATERIALS AND METHODS

Natural History of Shubby Reed-Mustard

*Schoenocrambe suffrutescens* was first discovered in 1935 by Edward Graham (Graham, 1937). The species was described as *Thelypodium suffrutescens* by Reed Rollins and renamed *Glaucocarpum suffrutescens* in 1938 (Rollins, 1938). Initial population counts revealed fewer than 3000 individuals in nine populations (including three populations of fewer than 30 plants) in the southern Uinta Basin area, and *Glaucocarpum suffrutescens* was listed as an endangered species on 6 Oct. 1987 under the Endangered Species Act of 1973 (US Fish and Wildlife Service, 2010). The genus name was changed from *Glaucocarpum* to *Schoenocrambe* in 1985 (Welsh and Chatterley, 1985) and is the scientific name accepted by the US Fish and Wildlife Service (2010). In light of recent molecular and morphological data, *Hesperidanthus* has been proposed as the genus name for several *Schoenocrambe* species (Al-Shehbaz, 2010), but this change is not yet widely accepted (US Fish and Wildlife Service, 2010).

Due to small populations and limited individual plant numbers, a recovery plan was approved in 1994 and again in 2010 (US Fish and Wildlife Service, 2010). Habitat fragmentation from energy development is considered to be the main threat to SRM on public lands, although habitat destruction from building stone mining was formerly a concern on federal lands and continues to be a concern for populations on private lands. In addition, herbivory by grazing livestock and feral horses is a possible source of SRM mortality that has not been studied (US Fish and Wildlife Service, 2010). According to the recovery plan, establishment or discovery of five populations of 2000 or more individuals each is required for downlisting to threatened, while 10 populations of 2000 or more individuals each is required for delisting. Partial population surveys conducted in 2004 and 2005 (Glisson, 2004, 2005) indicate that SRM numbers are limited to about 3000 individuals in seven populations in three main areas, and a plan to update SRM species biology and distribution was outlined in the 2010 recovery plan (US Fish and Wildlife Service, 2010). Recent research indicates that although SRM is capable of self-pollination, seed set is lower in individuals that are self-pollinated compared with cross-pollination with other individuals, and SRM may be pollinator limited (Lewis and Schupp, 2014). Although the Bureau of Land Management (BLM) implements a buffer zone of 91 m (300 ft) for soil surface disturbances near known SRM locations, the proliferation of roads and well pads in the study area probably creates additional disturbances due to dust generation. Several studies have documented the negative effects of dust deposition on vegetation health near unpaved roads (Auerbach et al., 1997; Myers-Smith et al., 2006; Lewis, 2013), and dust may also have indirect effects on pollinators (Lewis, 2013).

Study Area Description

The study area is in the Uinta Basin of northeastern Utah (Fig. 2). The surface geology of the Uinta Basin consists mainly of late Jurassic to Paleogene sedimentary strata and valley fill generated by uplift of the Uinta Mountains. The Green River forma-
tion (Eocene) represents a period of deposition in ancestral shallow lakes, in basins created during the Laramide orogeny, over what are now parts of Utah, Wyoming, and Colorado (Hinzte, 1993). The Eocene epoch (56–33 Ma) is generally believed to have had a warmer and wetter climate than today (Frantz et al., 2014), resulting in episodic periods of lush plant growth recorded in the sedimentary strata as deposits of oil, natural gas, and coal. Fluctuations in shorelines and water levels of the inland lakes resulted in episodic deposition of sandstones (near shore), shales and limestones (open water), and evaporites (tidal flats and sabkhas), with layers rich in fossil flora and organic deposits indicating a deltaic or estuarine system.

We investigated three known populations of SRM on public lands in the Uinta Basin occurring at Badland Cliffs (UTM 12N 570760E 4410912N), Big Pack Mountain (UTM 12N 611457E 4413344N), and Johnson Draw (UTM 12N 611302E 4401966N; Fig. 3). These populations represent two of the three main areas (the Big Pack Mountain and Johnson Draw populations are considered to be within the same area) where SRM populations are known to occur; the third area, Gray Knolls, is within the Uintah and Ouray Indian Reservation. The soil temperature regime is mesic (8 to <15°C) at all three population sites. Badland Cliffs and Johnson Draw soils were classified as Ustic Aridic (semi-desert) moisture regimes based on the presence of *Pinus edulis* Engelm. (two-needle pinyon) and *Juniperus osteosperma* (Torr.) Little (Utah juniper) vegetation; Big Pack Mountain soils were classified as a Typic Aridic (desert) moisture regime based on the lack of pinyon or juniper tree vegetation (NRCS, 2014).

**Field and Laboratory Characterization of Soils**

Field characterization and soil sampling were conducted in the Uinta Basin in October 2010, June 2011, and May 2012, on three known populations of SRM occurring at Big Pack Mountain, Johnson Draw, and Badland Cliffs (Fig. 3). Soil profile descriptions, vegetation, and site landscape characteristics were recorded at each sampling site. A total of 49 soil pedons (25 in SRM areas, 24 in non-SRM areas) were described and sampled by morphologic horizon in small, hand-dug excavations (Schoeneberger et al., 2012). Excavations dug to the 100-cm depth, or to bedrock if shallower, were at least 50 cm away from the canopy of a central plant within each plot, disturbing an area <1 m². Central plants were SRM for SRM plots and typically *Artemisia nova* A. Nelson (black sage) for non-SRM plots, although non-SRM plots also included *Pinus edulis*, *Atriplex confertifolia* (Torr. & Frém.) S. Watson (shadscale saltbush), *Eriogonum corymbosum* Benth. var. *corymbosum* (crispleaf buckwheat), and *Ephedra viridis* Coville (green Mormon tea). Of the 49 soils excavated, all 25 of the sampled SRM sites were on shale parent material (c.g., Fig. 1). Seventeen of the non-SRM sites...
were located on sandstone or an interbedded or mixed parent material containing sandstone; seven non-SRM sites were located on shale parent material. Soil sampling location coordinates at SRM presence and absence sites were taken with handheld Garmin 62s GPS units (Garmin International) with observed spatial accuracy in the field of about 3 m. Spatial locations of points were imported into R (R Core Team, 2013) and projected using the rgdal package (Bivand et al., 2013) to create point shapefiles.

Soil samples by morphologic horizon were characterized in the laboratory for rock fragment content and fine gravels. Soils were then sieved to obtain the <2-mm fraction and analyzed for air-dry water content (105°C) and chemical properties. Soils ground to <0.25-mm fraction and 0.002- to 0.05-mm fraction (Jackson, 2005). Water extracts (1:1 soil/water) were analyzed for pH. After pretreating samples with NaOAc to remove carbonates, particle size analysis was performed by the pipette method and sieving of the sand fraction (Soil Survey Staff, 2004), followed by settling to separate the silt fraction and 0.002- to 0.05-mm fraction (Jackson, 2005).

Random Forests Modeling

Random forests was chosen because of the use of OOB classification to validate the data set, as well as the ability to easily determine variable importance in the classification, facilitating variable elimination. For this study, variable elimination was an important criterion in developing the final predictive model because our objective was to create a final map product to facilitate the identification of potential SRM habitat in a field setting. To achieve this goal, minimal variable inputs and variables that are quickly and easily measured in the field were required. Partial dependence plots for each modeling effort were generated to aid in variable elimination; a subset of variables was eliminated by examining the inflection point of the plot of mean decrease in variable elimination. For this study, variable elimination was an important criterion in developing the final predictive model because our objective was to create a final map product to facilitate the identification of potential SRM habitat in a field setting. To achieve this goal, minimal variable inputs and variables that are quickly and easily measured in the field were required. Partial dependence plots for each modeling effort were generated to aid in variable elimination; a subset of variables was eliminated by examining the inflection point of the plot of mean decrease in accuracy, and the remaining variables were manually eliminated stepwise to obtain the best model fit.

Soil and Site Data

The soil and site point data set consisted of the 49 sampled (25 presence, 24 absence) sites of SRM. Soil property data, from surface horizons only, and site data were incorporated into the model because our objective was to use soil chemical, physical, and geomorphic properties that can be visually, spectrally, or geomorphically represented and correlated with remotely sensed spectral and topographic data. Summary statistics of the soil and site data for SRM presence and absence locations are given in Table 1. Soil data, with model variable codes in parentheses, analyzed from the <2-mm fraction consisted of CaCO$_3$ equivalent (CaCO$_3$Eq), pH (pH), clay content (Clay), silt content (Silt), gravel percentage (GRV), dry and moist Munsell color value (Dvalue and Mvalue) and chroma (Dchroma and Mchroma; Munsell Color, 2009), and depth to lithic contact (LDepth).

<table>
<thead>
<tr>
<th>Soil/site variable</th>
<th>Model code</th>
<th>SRM presence</th>
<th>SRM absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaCO$_3$ equivalent, g kg$^{-1}$†</td>
<td>CaCO3Eq</td>
<td>473</td>
<td>75</td>
</tr>
<tr>
<td>pH</td>
<td>pH</td>
<td>8.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Clay, %</td>
<td>Clay</td>
<td>19.8</td>
<td>4.8</td>
</tr>
<tr>
<td>Silt, %</td>
<td>Silt</td>
<td>57.6</td>
<td>9.8</td>
</tr>
<tr>
<td>Gravel, %</td>
<td>GRV</td>
<td>24.2</td>
<td>23.0</td>
</tr>
<tr>
<td>Dry value</td>
<td>Dvalue</td>
<td>7.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Moist value</td>
<td>Mvalue</td>
<td>5.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Dry chroma</td>
<td>DChroma</td>
<td>2.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Moist chroma</td>
<td>MChroma</td>
<td>2.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Soil depth, cm</td>
<td>LDepth</td>
<td>22.2</td>
<td>9.8</td>
</tr>
<tr>
<td>Northeastness</td>
<td>TAAspect1</td>
<td>0.367</td>
<td>0.257</td>
</tr>
<tr>
<td>Eastness</td>
<td>TAAspect2</td>
<td>0.543</td>
<td>0.382</td>
</tr>
<tr>
<td>Landscape position†</td>
<td>LFCode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface fragments, %</td>
<td>SurfFrag</td>
<td>51.1</td>
<td>28.4</td>
</tr>
<tr>
<td>Elevation, m</td>
<td>Elev</td>
<td>1828</td>
<td>186</td>
</tr>
<tr>
<td>Slope, %</td>
<td>Slope</td>
<td>17</td>
<td>14</td>
</tr>
</tbody>
</table>

† Class data; mean and SD not calculated.

<table>
<thead>
<tr>
<th>Soil/site variable</th>
<th>Model code</th>
<th>SRM presence</th>
<th>SRM absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil-depth, cm</td>
<td>LDepth</td>
<td>22.2</td>
<td>9.8</td>
</tr>
<tr>
<td>Northeastness</td>
<td>TAAspect1</td>
<td>0.367</td>
<td>0.257</td>
</tr>
<tr>
<td>Eastness</td>
<td>TAAspect2</td>
<td>0.543</td>
<td>0.382</td>
</tr>
<tr>
<td>Landscape position†</td>
<td>LFCode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface fragments, %</td>
<td>SurfFrag</td>
<td>51.1</td>
<td>28.4</td>
</tr>
<tr>
<td>Elevation, m</td>
<td>Elev</td>
<td>1828</td>
<td>186</td>
</tr>
<tr>
<td>Slope, %</td>
<td>Slope</td>
<td>17</td>
<td>14</td>
</tr>
</tbody>
</table>

Random forests was chosen because of the use of OOB classification to validate the data set, as well as the ability to easily determine variable importance in the classification, facilitating variable elimination. For this study, variable elimination was an important criterion in developing the final predictive model because our objective was to create a final map product to facilitate the identification of potential SRM habitat in a field setting. To achieve this goal, minimal variable inputs and variables that are quickly and easily measured in the field were required. Partial dependence plots for each modeling effort were generated to aid in variable elimination; a subset of variables was eliminated by examining the inflection point of the plot of mean decrease in accuracy, and the remaining variables were manually eliminated stepwise to obtain the best model fit.

Soil and Site Data

The soil and site point data set consisted of the 49 sampled (25 presence, 24 absence) sites of SRM. Soil property data, from surface horizons only, and site data were incorporated into the model because our objective was to use soil chemical, physical, and geomorphic properties that can be visually, spectrally, or geomorphically represented and correlated with remotely sensed spectral and topographic data. Summary statistics of the soil and site data for SRM presence and absence locations are given in Table 1. Soil data, with model variable codes in parentheses, analyzed from the <2-mm fraction consisted of CaCO$_3$ equivalent (CaCO$_3$Eq), pH (pH), clay content (Clay), silt content (Silt), gravel percentage (GRV), dry and moist Munsell color value (Dvalue and Mvalue) and chroma (Dchroma and Mchroma; Munsell Color, 2009), and depth to lithic contact (LDepth).

Site data included in the model consisted of two forms of transformed aspect (TAAspect1 and TAAspect2), surface fragments (SurfFrag), elevation (Elev), slope (Slope), and hillslope position (LFCode). Aspect measured in the field was transformed using the Roberts and Cooper (1989) aspect value transformations for radiation index, which relates to northeastness (TAAspect1):

$$\frac{1-\cos\left(\frac{\pi}{180}(\text{aspect}-30)\right)}{2}$$

and eastness (TAAspect2):

$$\frac{1-\sin\left(\frac{\pi}{180}\text{aspect}\right)}{2}$$

The variable LFCode was created by assigning a category to the landscape position of the site based on a modified hillslope model (Rohe, 1960) to create landform codes: 1, low-lying flats, steppes, and structural benches; 2, toeslopes and footslopes; 3, sideslopes and backslopes; 4, shoulders; 5, summits.

Additional data consisting of 250 SRM presence and absence points were collected in May 2013 by describing transects in the study areas. Using the initial predictive raster generated from soil and site data in the soil properties model, transects were generated in ArcMap that covered both presence and absence sites within the three population areas. These data were collected to use as training data for the final model, which used only remotely sensed spectral and topographic data to predict SRM presence or absence. A data set consisting of 1699 known SRM presence locations in the study area was obtained from the BLM for validation of the Spectral–Topographic model.
Topographic Data

Three DEMs from the National Elevation Dataset, with resolution of 9.327 m, were obtained from the Utah Automated Geographic Reference Center (AGRC) State Geographic Information Database (SGID) website (gis.utah.gov/data/) in September 2011. The DEMs were obtained in .dem format, then converted to raster format and mosaicked together to cover the study area. The DEM was pit-filled in ESRI ArcMap software (Environmental Systems Research Institute, 2010). An area solar radiation map was generated in ArcMap, using the spring equinox layer as a reasonable average representation of total annual incoming solar radiation. The pit-filled DEM was used to calculate several common surface derivatives using the Soil Inference Engine add-in for ArcGIS (ArcSIE; Shi, 2013) using a 10-m square neighborhood with the Shi algorithm (Shi et al., 2007). A slope map in gradient percentage was calculated along with the overall curvature, the general shape of the landscape (Curvature); profile curvature or vertical landscape shape (ProfCurvature); planform curvature or horizontal landscape shape (PlanCurvature), and tangential curvature, the shape of the landscape in the direction that is perpendicular to the surface at the given location (TanCurvature). Negative values of curvatures indicate convex shapes and positive values indicate concave shapes. Variation in neighborhood size can be used to select for the scale of the desired terrain attribute; microtopographical features are more evident with smaller neighborhood sizes (one nearest cell in each direction, or 10 m, is a neighborhood of nine), whereas larger neighborhood sizes (three nearest cells, or 30 m, is a neighborhood of 27) create a smoothing effect that emphasizes broad landscape features (Roecker and Thompson, 2010).

Spectral Data

Landsat 5 Thematic Mapper imagery was obtained from the USGS Global Visualization Viewer (GloVis) website (glovis.usgs.gov, Earth Resources Observation and Science Center [EROS]; Fig. 3). Imagery from Path 37, Row 32, was acquired by the Landsat 5 Thematic Mapper on 1 Sept. 2011. The imagery included the normal visible (1, 2, and 3), near-infrared (4), shortwave infrared (5 and 7), and thermal (6) bands. All bands were at 30-m resolution except the thermal band, which was at 90-m resolution.

The Landsat imagery was re-projected to the NAD83 UTM Zone 12N GRS 1980 projected coordinate system, atmospherically corrected (Chavez, 1996), clipped to an area containing the study sites, and resampled to match the resolution (9.327 m) and cell centers of the DEM using ERDAS Imagine image processing software (Leica Geosystems, 2011) and R software (R Core Team, 2013) using the rgdal (Bivand et al., 2013) and raster (Hijmans and van Etten, 2012) packages. The DEM was clipped to the Landsat imagery in R, and the imagery was analyzed using ERDAS Imagine. Landsat 5 Bands 1 through 5 and 7, several normalized band ratios, and the normalized difference vegetation index (NDVI), were used in the random forests models (Table 2). Tasseled cap transformations were also calculated.

Table 2. Summary of spectral topographic variables, their abbreviations, and generalized physical representations of each spectral variable used in the Soil–Spectral Correlation and Spectral–Topographic models.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Data represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat 5 Thematic Mapper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band 1</td>
<td>Band1</td>
<td>water, surface color</td>
</tr>
<tr>
<td>Band 2</td>
<td>Band2</td>
<td>water, vegetation, surface color</td>
</tr>
<tr>
<td>Band 3</td>
<td>Band3</td>
<td>vegetation, surface color</td>
</tr>
<tr>
<td>Band 4</td>
<td>Band4</td>
<td>vegetation</td>
</tr>
<tr>
<td>Band 5</td>
<td>Band5</td>
<td>soil and parent material mineralogy</td>
</tr>
<tr>
<td>Band 7</td>
<td>Band6</td>
<td>soil and parent material mineralogy</td>
</tr>
<tr>
<td>Simple 3/2</td>
<td>SR3over2</td>
<td>carbonates, mineralogy</td>
</tr>
<tr>
<td>Normalized 3/1</td>
<td>NormSand1</td>
<td>geology</td>
</tr>
<tr>
<td>Normalized 3/2</td>
<td>NormSand2</td>
<td>geology, carbonates</td>
</tr>
<tr>
<td>Normalized 5/2</td>
<td>NormSand7</td>
<td>geology, vegetation</td>
</tr>
<tr>
<td>Normalized 5/7</td>
<td>NormSand7</td>
<td>soil and vegetation moisture content, surface salts</td>
</tr>
<tr>
<td>Normalized 7/2</td>
<td>NormSand2</td>
<td>vegetation</td>
</tr>
<tr>
<td>Normalized difference vegetation index</td>
<td>NDVI</td>
<td>vegetation</td>
</tr>
<tr>
<td>Tasseled cap transformations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightness</td>
<td>Brightness</td>
<td>surface color, vegetation, carbonates</td>
</tr>
<tr>
<td>Greenness</td>
<td>Greenness</td>
<td>vegetation</td>
</tr>
<tr>
<td>Wetness</td>
<td>Wetness</td>
<td>soil moisture</td>
</tr>
<tr>
<td>Yellowness</td>
<td>Yellowness</td>
<td>carbonate content, geology</td>
</tr>
<tr>
<td>Fifth</td>
<td>Fifth</td>
<td></td>
</tr>
<tr>
<td>Sixth</td>
<td>Sixth</td>
<td></td>
</tr>
<tr>
<td>Topographic data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area solar radiation</td>
<td>AreaSol</td>
<td>soil moisture, depth, aspect effects</td>
</tr>
<tr>
<td>Total curvature</td>
<td>Curvature</td>
<td>soil moisture, depth, water shedding</td>
</tr>
<tr>
<td>Planform curvature</td>
<td>PlanCurvature</td>
<td>soil moisture, depth, water shedding</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>ProCurvature</td>
<td>soil moisture, depth, water shedding</td>
</tr>
<tr>
<td>Tangential curvature</td>
<td>TanCurvature</td>
<td>soil moisture, depth, water shedding</td>
</tr>
</tbody>
</table>
Objective 1: Soil Properties Model

Soil and site data (Table 1) for the 49 SRM presence and absence sites were analyzed for classification of presence or absence using random forests in R (Liaw and Wiener, 2002). Initially, 30,000 trees were fitted. However, plotting classification error vs. number of iterations indicated that 10,000 trees were sufficient to stabilize the prediction and accurately classify the data with minimal variation in classification accuracy from randomness, and were used in all classifications throughout the remaining analysis. The top 10 variables were selected based on the inflection point at about 10% in mean decrease in accuracy in the variable importance plot of all soil and site data (Fig. 4). These top variables were manually removed stepwise to obtain the best model fit (highest model accuracy with the fewest number of variables).

Objective 2: Soil–Spectral Correlation Model

Soil and site data were correlated to spectral and digital topographic data in the Soil–Spectral Correlation (SSC) model. Using R software, Landsat Bands 1 through 5 and 7 were combined into a Raster Stack object, and common band ratios were calculated from the Raster Stack object. Band ratios were also put into a Raster Stack and included a simple Band 3/Band 2 ratio, a Band 3/Band 2 normalized difference ratio [NDR, (Band 3 – Band 2)/(Band 3 + Band 2)], Bands 3/1 NDR, Bands 7/2 NDR, an NDVI (Bands 4/3 NDR), and Bands 5/2 NDR. A tasseled cap transformation (Kauth and Thomas, 1976), a global vegetation index that accounts for soil brightness and moisture content (Crist and Kauth, 1986; Jensen, 2005), was calculated in R from the imagery, and six bands were generated: brightness, greenness, wetness, yellowness, fifth, and sixth. These bands were included in the model to test their relationship to measured soil properties. In particular, we hypothesized that brightness and yellowness might be positively correlated with CaCO$_3$ equivalent and lithology type at SRM population sites. The six tasseled cap bands were then stacked together into a Raster Stack object. All Raster Stack objects, including the six Landsat 5 Thematic Mapper bands, the six calculated indices, the tasseled cap bands, and the DEM-derived rasters, were stacked into a Raster Brick object.

The presence or absence points were then converted to a SpatialPointsDataFrame and edited so that each 30-m pixel was represented by only one point. This eliminated six points (four presence, two absence) to total 43 points. The edited points were used to sample through the Raster Brick to obtain values of the bands, ratios, and tasseled cap bands at each point. Correlation of the predictor variables (soil and site data) was checked for collinearity. The soil and site variables selected were correlated to the remotely sensed imagery and the derived indices and band ratios. The sampled values were then put into a data frame and analyzed using random forests classification. Variables above about 10% in mean decrease of accuracy (Fig. 5) were then manually eliminated stepwise to obtain the subset of the fewest variables with the highest model accuracy. The final spectral variables were combined into a Raster Stack object and used in the random forests model to create a predictive raster layer of SRM spatial extent. Final classification error was calculated by extracting predicted SRM presence or absence from the classified raster and comparing with the original 43 presence or absence soil data points.

The predictive raster generated by the SSC model was then used to inform collection of presence or absence locations to be
used as the training data set for the final Spectral–Topographic model. Because the original 49 soil sampling locations were clustered in known SRM populations, transects were designed to cover a larger area within the general study area. Six digitally generated transects, varying from 1 to 2 km in length and covering both SRM presence and absence locations as predicted by the SSC model, defined the field sampling locations for the ST model training data. The correlation of soil and site data to spectral and topographic data also eliminated the necessity of physically testing the soil to determine SRM habitat suitability, a time-consuming and labor-intensive step.

Objective 3: Spectral–Topographic Model

A similar process to the development of the SSC model was performed to generate the Spectral–Topographic (ST) model. Data consisting of 250 SRM presence and absence observation points, collected in the field by following transects through SRM presence and absence sites, as predicted by the SSC model, were used to train the model. A data set consisting of 1699 known SRM presence locations in the study area was obtained from the BLM for validation of the ST model. In this model, only spectral and topographic data (Landsat bands and associated derivatives and DEM derivatives) were used in the random forests classification to generate a final predictive raster. Presence or absence points were converted to a SpatialPointsDataFrame, edited so that each point represented only one pixel, and analyzed using random forests classification. Variables with a mean decrease of accuracy of greater than about 30% in the variable importance plot (Fig. 6) were analyzed by manual stepwise elimination to determine the fewest variables with the highest model accuracy. The remaining spectral and topographic variables were combined into a Raster Stack object, and model validation was performed by sampling through the final predictive raster with the 1699 known SRM locations and calculating the percentage of points correctly predicted.

Assessment of Model Accuracy

An imbalance in numbers of presence and absence classes in the predictor data can influence the bootstrap samples, biasing predictions toward the majority class (Chen et al., 2004). Approximately equal proportions of presence and absence data were used as training data in the random forests models to avoid bias in predictions and model fits. Measures of prediction accuracy between presence and absence classes, model specificity (percentage of absence points correctly classified) and model sensitivity (percentage of presence points correctly classified), were included to quantify any imbalances in model predictions. Other measures of model accuracy used to compare classifications among models included the kappa statistic (Monserud and Leemans, 1992), a goodness of model fit measure that compares the proportion of agreement of map cells after chance agreement (random relocation of all map cells), also an important consideration for imbalanced classes. Kappa can vary from −1 (no agreement) to 1 (perfect agreement; Prasad et al., 2006), and negative values indicate less agreement than expected by chance. Kappa values are often interpreted as >0.8 representing good agreement, 0.4 to 0.8 representing moderate agreement, and <0.4 representing poor agreement (Congalton and Green, 1998). The area under the receiver operating char-

![Variable importance plot](image)

**Fig. 5.** Variable importance plot showing all spectral data used in shrubby reed-mustard site classification in the Soil–Spectral Correlation model. Variables to the right of the dashed line (>10% mean decrease in accuracy) were selected and reduced to a subset by stepwise elimination to give the best model fit with the fewest variables. The top four variables were yellowness, Bands 3 and 2 normalized difference ratio (NDR) (norm3and2), Bands 3 and 1 NDR (norm3and1), and Bands 7 and 2 NDR (norm7and2).
acteristic curve, or AUC (DeLong et al., 1988), a measure of accuracy that places equal emphasis on both classes and takes into account any correlation in the data, was calculated in R as an additional measure of model classification accuracy.

RESULTS AND DISCUSSION

Soil Properties Model Results

Using a random forests classification scheme, all field and laboratory soil and site data were ranked in order of predictive value for determining SRM presence or absence (SP model) in a variable importance plot (Fig. 4), with an OOB error of 16.3%. Classification with only the top five variables (CaCO₃ equivalent, silt, dry color value, pH, and northeastness) improved the OOB error to 10.2%. Partial dependence plots for these variables were also generated (Fig. 7), showing both the range in values of each variable and the effect of the variable on prediction probability. Partial dependence, the dependence of the probability of presence on a predictor variable, after averaging out the effects of the other variables (Cutler et al., 2007), is nonlinear for the five most important predictor variables. For CaCO₃ equivalent, silt, dry color value, and pH, the probability of SRM presence prediction increased as the value of

Fig. 6. Variable importance plot showing all spectral and topographic data used in shrubby reed-mustard site classification in the Spectral-Topographic model. Variables to the right of the dashed line (>30% mean decrease in accuracy) were selected and reduced to a subset by stepwise elimination to give the best model fit with the fewest variables. The top six variables were area solar radiation, Bands 7 and 2 normalized difference ratio (NDR) (norm7and2), Bands 5 and 2 NDR (norm5and2), Bands 3 and 2 NDR (norm3and2), normalized difference vegetation index (NDVI), and greenness.

Fig. 7. Partial dependence plots of the top five variables used in shrubby reed-mustard site classification in the soil properties model, including CaCO₃ equivalent (10 g kg⁻¹), silt content (%), dry color value (DValue), pH, and northeastness (TAspect1). The y axis is (logit of probability of presence)/2.
each variable increased. In the field, measured CaCO₃ equivalent values above 450 g kg⁻¹, silt content above 50%, dry color value above 7, and pH above 8.5 are all indicators of increased probability of site prediction as SRM habitat. For northeast-southwest, values between 0.2 and 0.7, or north-facing slopes, had a higher probability of predicting SRM presence. The dependence of the probability of SRM occurrence on CaCO₃ equivalent, silt, and pH levels off at values at the high end of the range of each predictor variable, indicating that although SRM presence is positively correlated to these variables, no further predictive value is gained once these variables reach a certain value due to the nonlinear nature of the relationship.

**Soil–Spectral Correlation Model Results**

In the SSC model, the random forests classification determined which digital layers could be related to soil and site characteristics. These digital layers were used to predict the spatial extent of SRM habitat, and a variable importance plot was generated (Fig. 5). Classification of all Landsat-derived and DEM-derived layers had an OOB error of 30.2%.

In the SSC model, the soil and site data were also correlated to spectral data to determine which bands or derivatives might be most useful as proxies for field data. In particular, the yellowness tasseled cap band was positively correlated to both CaCO₃ equivalent (0.58) and silt (0.56). The 3/2 NDR and 3/1 NDR were also negatively correlated to CaCO₃ equivalent (−0.57 and −0.69, respectively) and silt (−0.56 and −0.61, respectively). Using only the top four variables (yellowness, 3/2 NDR, 3/1 NDR, and 7/2 NDR) in the random forests model slightly improved the OOB error to 27.9%. High values of yellowness indicated a high probability of predicting SRM presence, whereas low values of 3/2 NDR, 3/1 NDR, and 7/2 NDR indicated a high probability of predicting SRM presence. The dependence of the probability of occurrence on these predictor variables is also nonlinear in partial dependence plots and drops off sharply before leveling off (Fig. 8). These indexes and band ratios typically indicate differences in surface color and geology (Jensen, 2005) and provide further support for observations that SRM habitat is linked to lithology. Other bands and band ratios that are indicative of vegetation cover and type (NDVI, greenness, Band 3, Band 5) were not as useful in the sparsely vegetated study area or were negatively correlated to SRM presence, indicating that SRM typically grows in low-vegetation areas.

A predictive raster indicating SRM presence or absence and spatial extent was generated by sampling the spectral data at the locations of the original soil descriptions (43 of 49 original points were used as training data due to the close proximity of some points). Although not intended to be used in the final stepwise prediction process for determining SRM presence or absence, the SSC model output was a necessary step to determine point sampling locations in both predicted presence and absence sites for the ST model.

In this study, a neighborhood size of 27 (three nearest cells in each direction or 30 m) was too large to capture the variation in slope curvature, so we used a neighborhood size of nine (one nearest cell or 10 m), with a mean decrease in accuracy result of 8% for overall slope curvature. Because SRM is often located on small, localized shale beds (on the order of a few meters thick) between thicker sandstone beds (e.g., Fig. 1), a high-resolution DEM (such as lidar, unavailable in this study) may contribute to model accuracy where microtopography or microclimate effects are of concern.

**Spectral–Topographic Model Results**

A final model was created with the objective of using only spectral and topographic data to predict the spatial extent of SRM habitat (ST model). In this model, the 250 SRM presence or absence transect points were used as training data across the spatial extent and validated with a data set obtained from the
BLM of 1699 known SRM locations. A variable importance plot was generated with all the remotely sensed data variables (Fig. 6), with an OOB error of 22%. Manual stepwise elimination of the top 10 variables to obtain the best model fit with the fewest variables identified six variables (area solar radiation, 7/2 NDR, 5/2 NDR, 3/2 NDR, greenness, and NDVI) with a similar OOB error of 23.2%. The partial dependence plots of the ST model variables also indicated a nonlinear relationship of each predictor variable to the probability of presence prediction (plots not shown). Previous modeling research identified solar radiation as an important predictor of rare plant habitat (Wiser et al., 1998). The final predictive raster, showing SRM presence or absence classification, and the locations of all the data used in the model, is given in Fig. 9. Validation of the ST model, performed with 1699 known SRM locations, gave an accuracy of 69.8%. However, many of the points incorrectly classified were in close proximity (1 cell or 10 m) to predicted SRM presence locations (Fig. 10), so actual accuracy could be higher given the uncertainty associated with resampling the Landsat data to DEM resolution (9,327 m). The difference in scales between SRM plots (50 cm) and Landsat data (30 m) may have also resulted in errors in the final predictive raster. The use of spectral data with finer spatial resolution could aid in improving model results; resampled Landsat 8 imagery in conjunction with lidar DEM (not currently available for the study area) would more closely match the soil and site sampling plots.

**Assessment of Model Accuracy Results**

In the SP model, which used soil and site data to classify potential SRM habitat, model sensitivity (percentage of absence points correctly classified) was lower than model specificity (percentage of presence points correctly classified; Table 3). In the SSC and ST models, which classified SRM presence or absence...
Table 3. Summary of model accuracy and classification metrics. Model statistics are given for all variables (all) and the top variables as determined by stepwise elimination of the 10 variables with the highest values in the variable importance plot for the soil properties (SP model), Soil–Spectral Correlation (SSC) model, and Spectral–Topographic (ST) model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Kappa</th>
<th>AUC†</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>all</td>
<td>83.7</td>
<td>75.0</td>
<td>92.0</td>
<td>0.672</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>top 5</td>
<td>89.8</td>
<td>87.5</td>
<td>92.0</td>
<td>0.796</td>
<td>0.918</td>
</tr>
<tr>
<td>SSC</td>
<td>all</td>
<td>69.8</td>
<td>72.7</td>
<td>66.7</td>
<td>0.394</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>top 4</td>
<td>72.1</td>
<td>72.7</td>
<td>71.4</td>
<td>0.442</td>
<td>0.729</td>
</tr>
<tr>
<td>ST</td>
<td>all</td>
<td>78.0</td>
<td>79.0</td>
<td>76.8</td>
<td>0.556</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>top 6</td>
<td>76.8</td>
<td>81.2</td>
<td>71.4</td>
<td>0.528</td>
<td>0.837</td>
</tr>
</tbody>
</table>

† Area under the receiver operating characteristic curve.

Potential Use of Shrubby Reed-Mustard Modeling for Land Management

As a stepwise system for locating potential SRM habitat, these models can provide a tool for land managers to locate, identify, and check key soil properties at potential SRM habitat sites. First, the ST model can be used to narrow potential habitat within the study area (Objective 3). The ST model incorporates only spectral and topographic data and facilitates the rapid examination of large areas of land when time, funding, and/or accessibility are of concern. Many of the areas where SRM populations are currently known or could possibly occur are in remote, rugged areas that are difficult and time consuming to access. A model that uses existing Landsat and DEM data as predictors of SRM presence with >75% accuracy, such as the ST model, can help focus time, manpower, and monetary efforts into areas with greater potential for success in conservation efforts. Many previously unknown SRM individuals were found in the spatial extent predicted by the SSC model during training data collection for the ST model in spring 2013. Similarly, Williams et al. (2009) successfully used random forests models to predict new rare plant occurrences in northern California. Although SRM was not found in all locations in the spatial extent predicted by the model, it is a rare plant with low population numbers, and these areas may indicate good potential habitat for restoration and revegetation. The successful interpolation of SRM niche spaces within the spatial modeling framework suggests that soil and site effects on SRM species distribution are well understood in an ecological context and accurately depicted by random forests models.

ACKNOWLEDGMENTS

Leanna Reynolds Hayes, Vance W. Almquist, Lauren S. Kelly, Joel Drake, and Jeremiah D. Armentrout assisted with field data collection and laboratory analyses. Matthew Lewis and Eugene Schupp shared their knowledge and expertise about the reproductive biology and natural history of shrubby reed-mustard. Colby W. Brungard created the final high-resolution figures (Fig. 3–10). Financial and logistical support was provided by the US Fish and Wildlife Service (Agreement no. 601819G319), the Utah Agricultural Experiment Station (UAES), and the Utah State University Ecology Center. Approved as UAES Journal Paper no. 8840.

REFERENCES

vegetation, Oxford Univ. Press, New York.
Leica Geosystems. 2011. ERDAS Imagine, Version 11. Leica Geosystems, Atlanta, GA.
Shi, X. 2013. ArcSIE 10.0.519 for Arc GIS 10.0. Dartmouth College, Hanover, NH.