5-2010

Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah

Alexander Knell Stum
Utah State University

RANDOM FORESTS APPLIED AS A SOIL SPATIAL
PREDICTIVE MODEL IN ARID UTAH

by

Alexander Knell Stum

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Soil Science

Approved:

Dr. Janis L. Boettinger Major Professor	Dr. R. Douglas Ramsey Committee Member
Dr. Michael White Committee Member	Dr. Byron R. Burnham Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2010

ABSTRACT

Random Forests Applied as a Soil Spatial

Predictive Model in Arid Utah

by

Alexander Knell Stum, Master of Science

Utah State University, 2010

Major Professor: Janis L. Boettinger
Department: Plant, Soils, and Climate

Initial soil surveys are incomplete for large tracts of public land in the western USA. Digital soil mapping offers a quantitative approach as an alternative to traditional soil mapping. I sought to predict soil classes across an arid to semiarid watershed of western Utah by applying random forests (RF) and using environmental covariates derived from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and digital elevation models (DEM). Random forests are similar to classification and regression trees (CART). However, RF is doubly random. Many (e.g., 500) weak trees are grown (trained) independently because each tree is trained with a new randomly selected bootstrap sample, and a random subset of variables is used to split each node. To train and validate the RF trees, 561 soil descriptions were made in the field. An additional 111 points were added by case-based reasoning using aerial photo interpretation. As RF makes classification decisions from the mode of many independently grown trees, model uncertainty can be derived. The overall out of the bag (OOB) error was lower without weighting of classes; weighting increased the overall OOB error and the resulting output did not reflect soil-landscape relationships observed in the field. The final RF model had an OOB error of 55.2% and predicted soils on landforms consistent with soil-landscape relationships. The OOB error for individual classes typically decreased with increasing class size. In addition to the final classification, I determined the second and third most likely classification, model confidence, and the hypothetical extent of individual classes. Pixels that had high possibility of belonging to multiple soil classes were aggregated using a minimum confidence value based on limiting soil features, which is an effective and objective method of determining membership in soil map unit associations and complexes mapped at the 1:24,000 scale. Variables derived from both DEM and Landsat 7 ETM+ sources were important for predicting soil classes based on Gini and standard measures of variable importance and OOB errors from groves grown with exclusively DEM- or Landsat-derived data. Random forests was a powerful predictor of soil classes and produced outputs that facilitated further understanding of soil-landscape relationships.

(147 pages)

ACKNOWLEDGMENTS

I want to thank the Utah Agricultural Experiment Station and the Bureau of Land Management (BLM) of Utah for funding this research. And special thanks to Lisa Bryant with the BLM for her direct support in the field and her desire to apply this kind of research in the field.

Kent Sutcliffe and Tom Simper with the Natural Resources Conservation Service (NRCS) gave invaluable technical assistance, teaching me all the local plants, including me on field reviews, sharing knowledge of soil-landscape relationships, and much more.

My days in the field would have been at times unbearable without the company of my friends and undergraduate assistants Morgan Koenig and Colby Brungard. There is no way I could have canvassed so much area and remained so focused without them. I especially want to thank Colby whose desire to learn inspired me further.

I simply marvel at Janis's aptitude to learn and deeply understand what each of her graduate students have done. My research has been no exception. I have learned so much from her in and out of the academic setting. Thank you for your continual support these many years.

This task could not have been completed without the gentle (sometimes urgent) persuasions of my mother (Rebecca Stum) and also my wife, Marci. Marci has sacrificed and given more than anyone else to get me here. I love you and thank you.

Alexander K. Stum

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

Knowledge of soil systems is necessary to understand our world's natural systems, geomorphology, hydrology, ecology, and climatology (Lookingbill and Urban, 2004). The incorporation of topographic and remotely sensed (RS) data into the study of soil systems has increased our ability to predict the spatial distribution of soils across the landscape (McBratney et al., 2003; Scull et al. 2003).

Typically, soils are represented on a thematic map made up of polygons representing individual map units (Soil Survey Division Staff, 1993; Scull et al., 2003; USDA-NRCS, 2009). Each map unit represents the generalized distribution of soils in the landscape as an association, complex, consociation, or as undifferentiated (Moran and Bui, 2002). Soils or landscape features (e.g., rock outcrop) that are known to occur within a map unit are referred to as components. Components in an association can be delineated at the scale of mapping, whereas components of a complex cannot be delineated at the scale of mapping (Soil Survey Division Staff, 1993).

Traditional soil maps illustrate conceptual models of soil distribution on the landscape. When a soil scientist draws a line on the map he/she is predicting that certain soils are likely to be found within the delineated polygon. The parameters of this conceptual model are complicated, and are related to the individual experience or tacit knowledge of the soil scientist, and are, therefore, subjective (Hudson, 1992).

Traditional soil survey methods, while thoroughly reviewed, are not assessed for accuracy. Many soil predictive models also produce uncertainty maps which can focus future field activities and give the user further information about the map. Ongoing research in digital soil mapping has demonstrated that reasonably accurate soil maps can be produced using quantitative predictive models. Digital soil mapping may also expedite soil survey (Lagacherie and Holmes, 1997; Dobos et al., 2000; Zhu, 2000; Moran and Bui, 2002; McBratney et al., 2003; Scull et al., 2003; Cole, 2004; Shi et al., 2004; Henderson et al., 2005; Saunders, 2005; Scull et al., 2005; Cole and Boettinger, 2007; Saunders and Boettinger, 2007; Brungard, 2009).

Soil Formation – A Theoretical Framework

Hans Jenny (1941) presented an elegant function to explain the current state of a soil: S = f(c, o, r, p, t). Simply said, soil (S) is a function of five environmental factors: climate (c), organisms (o), relief (r), parent material (p), and time (t). While this function has proven difficult, if not impossible, to solve, it has set forth a theoretical framework whereby soil formation can be studied. To simplify the function and better understand soil formation, studies have focused on identifying transitions in soil properties along sequences of soils related to one environmental factor; such as climate (climosequences), time (chronosequences), relief (toposequences), etc. (Jenny, 1980; Birkeland, 1999).

Digital Soil Mapping – A Spatial Framework

Spatial data analysis seeks to elucidate some pattern or process that occurs in space, perhaps allowing us to make predictions where no observations have been made (Bailey and Gatrell, 1995). Based on Jenny's soil forming factors (1941), McBratney et al. (2003) proposed an empirical formulation to quantitatively find correlations between spatially explicit data and the soil. They considered seven factors, or environmental covariates, in their model, referred to as "scorpan." The five soil forming factors from Jenny (1941) are still present as covariates: ‘c’ climate; ‘o’ organisms, vegetation, fauna, and/or human activity; ‘r’ topography and landscape attributes; ‘p’ parent material, lithology; and ‘a’ time or age.

Two additional scorpan covariates are specifically directed towards spatial predictive models: ‘s’ soil or soil properties, and ‘n’ space, spatial position, or relative position. There are two general forms of the scorpan model (McBratney et al., 2003):

$$S_c = f(s,c,o,r,p,a,n) \quad or \quad S_a = f(s,c,o,r,p,a,n)$$

where S_c is soil class and S_a is a soil attribute or property.

McBratney et al. (2002) demonstrated that some soil properties (e.g., saturated hydraulic conductivity) may be predicted from other soil properties (e.g. sand content) using quantitative functions, known as pedotransfer functions. Where a sufficient number of soil property observations are available, pedotransfer functions can be incorporated into spatial models to predict soil class or other soil attributes. Also, soil maps representing soil classes can help predict soil attributes. The general form of a soil spatial prediction function is

$$S(x,y,z,t) = f(Q)$$

where S is a soil located at coordinates x,y,z, for a period of time, t, and is a function of predictor variable(s) Q.

The values of predictor variables (independent variables) must be known at the points where the soil class or attributes are to be predicted.

Objective

The Bureau of Land Management (BLM) administers 258 million surface acres, mostly in the western United States (BLM, 2006). The BLM must make appropriate management decisions related to grazing allotments, recreational activities, fire restoration, mine reclamation, chaining, hydrologic studies, wildlife monitoring and much more. The BLM land managers need to know the spatial distribution of soils to support these management decisions.

For rangeland planning, the BLM normally needs third order soil survey maps at the 1:24,000 scale. Large tracts of BLM land have no soil data, or only have fourth order or fifth order soil maps which are completed at a very coarse spatial resolution and present insufficient detail for many land management activities (Table 1).

Table 1. Orders of soil mapping (Soil Survey Division Staff, 1993).

Mapping level	Minimum-size delineation [ha]	Appropriate map scale
1st Order – experimental plots, building sites	1 or less	1:15,840 or larger
2nd Order – agriculture/urban planning	0.6 to 4	1:12,000 to 1:31,680
3rd Order – range planning	1.6 to 16	1:20,000 to 1:63,360
4th Order – general soil information	16 to 252	1:63,360 to 1:250,000
5th Order – regional planning	252 to 4,000	1:250,000 to 1:1,000,000 or smaller

The Natural Resources Conservation Service (NRCS) recently completed the soil mapping of privately owned lands in central Beaver County, Utah. The BLM manages 440,648 ha (1.14 million ac) of Beaver County, or 68.8% of the county's area. Because of the high cost of traditional methods of soil mapping and the remoteness of much of the county, the BLM has been interested in facilitating the investigation of alternative soil mapping techniques. A 47,000-ha watershed northwest of Milford, Utah, was selected as a trial area to implement soil spatial predictive models to create a soil map. The objective of this study was to apply a soil spatial predictive model to create a soil map at the 1:24,000 scale with topographic and remotely sensed data as environmental covariates (soil-forming factors). I hypothesized that these spatially explicit data layers can be successfully incorporated into quantitative models (i.e. random forests) to predict soil types across the landscape and generate estimates of prediction uncertainty.

LITERATURE REVIEW

The following is a review of pertinent literature of scorpan environmental covariates and of spatial prediction functions and models.

Environmental Covariates

Most environmental covariates can be represented by remotely sensed spectral data or derivatives from digital elevation models. Satellite imagery and aerial photography are remotely measured properties of the land surface itself, be it soil, water, geology, human infrastructure or various combinations of these. Electromagnetic (EM) radiation is reflected, absorbed, or emitted as a function of the physical and chemical properties of that surface (Goetz, 1989; Rees, 2001).

The topographic surface (x, y, z) can be represented in several formats, such as an isarithmic map (contour map), triangular irregular network (TIN), raster digital elevation model (DEM), and others (DeMers, 2000). The raster digital elevation model is the representation of a point data set as a raster surface. Each grid cell value is the predicted or interpolated elevation at the center or corner of the cell. Digital elevation models are the more commonly used in geographic information systems (GIS) environments because primary derivations, such as slope and aspect, can be easily calculated (Chaplot et al., 2006). Very useful secondary derivations, such as compound topographic index (CTI), specific catchment area, and stream networks can also be derived.

Soil Properties (s) & Parent Material (p)

Soil is a combination of inorganic solids, organic matter, gases, and soil water constituents. Each one of these influences the way in which electromagnetic radiation interacts with the soil surface. Teasing out specific soil properties from this surface signal can be difficult.

Gomez et al. (2008) modeled soil organic carbon by multivariate regression of Hyperion hyperspectral satellite imagery (242 bands in the visible and near infrared, 400-2500 nm). Anderson and Croft (2009) reviewed literature related to soil surface roughness and soil moisture studies. They found promising applications of optical remote sensing to measure albedo and bidirectional reflection with regard to soil surface roughness and physical structure. They also explored active microwave systems (e.g. ALOS, RADARSAT-2, and Terra SAR-X), which take advantage of the difference in dielectric constants between water and soil to determine soil surface moisture.

The largest constituent of the soils in this study area is the inorganic solids, which are primary and secondary minerals derived from the weathering of parent material (local geology) (Birkeland, 1999). These minerals may impart a unique spectral signature that can be used to identify the mineralogy of the soil surface (Goetz, 1989; Irons et al., 1989; McBratney et al., 2003). This information can be related to other soil attributes or classes.

El Rakaiby et al. (1994) measured *in situ* reflectance of geology members on the Sinai Peninsula. The spectral signatures gathered in the field with handheld radiometers were analyzed and used to evaluate Landsat 7 ETM+ satellite imagery taken over the region. They determined that Landsat 7 ETM+ band ratios were effective in distinguishing different geological units (Table 2). The use of band ratios also mitigates the influence of shadows, as absolute brightness values of the satellite image are not being used but rather the relative brightness of one band compared to another (Goetz, 1989; Jensen, 2005).

Bodily (2005) used the normalized difference ratio of Landsat 7 ETM+ bands 5 and 2 to identify limestone outcroppings (Table 2). His findings were consistent with radiometric lab measurements of the spectral profile of limestone and dolomite, where limestone and dolomite have greater reflectance in band 5 relative to band 2 while andesite and other igneous materials have greater reflectance in band 2 relative to band 5 (NASA, 2008). Cole (2004) and Saunders (2005) incorporated simple Landsat band ratios (3/2, 3/7, and 5/7) with DEM-derived data to predict soil types with knowledge-based and decision tree classifications, respectively.

Nield et al. (2007) successfully used normalized difference band ratios with Landsat 7 ETM+ bands 5 and 7 and bands 5 and 4 to identify gypsic and natric soil areas, respectively, in an arid area of central Utah. Bands 5 and 7 appeared to be correlated with the occurrence of near-surface secondary gypsum, whereas bands 5 and 4 were most likely correlated with the co-occurrence of Fe-bearing desert varnish on surface rocks fragments near natric soil area training sites.

Climate (c)

Data layers that explicitly represent climatic variables at an appropriate scale (1:24000) are usually not readily available. For example, PRISM data has a spatial resolution of approximately 4km (USDA-NRCS, 2000), which is too coarse for many 3rd order (or lower order) soil surveys (Table 1).

Table 2. Landsat 7 ETM+ spectral and spatial resolution of each band (Jensen, 2005).

Band	Spectral Resolution [µm]	Spatial Resolution [m] at Nadir
1	0.450 – 0.515	30 x 30
2	0.525 – 0.605	30 x 30
3	0.630 – 0.690	30 x 30
4	0.750 – 0.900	30 x 30
5	1.55 – 1.75	30 x 30
6	10.40 – 12.50	60 x 60
7	2.08 – 2.35	30 x 30
8	0.52 – 0.90	15 x 15

Organisms (o)

Vegetation cover affects the absorbance and reflectance response of the land surface, and thus can be quantified using satellite imagery. Chlorophyll and other pigments of plants absorb light in the visible spectrum (0.35-0.70 µm) for photosynthesis (Jensen, 2005). This absorption feature is most pronounced around the blue (0.45-0.52 µm) and red (0.63-0.69 µm) portions of the visible spectrum. The spongy mesophyll layer of the leaf transmits or reflects 90-95% of radiant energy from 0.7 to 1.2 µm in the near infrared (NIR) portion of the spectrum. Therefore, the simple ratio of the measured reflectance (ρ) of NIR to Red is greatest in areas of leafy vegetation. Normalized difference vegetation index (NDVI) has most commonly been used to represent vegetation in digital soil mapping (McBratney et al., 2003): $\frac{\rho_{NIR} - \rho_{\mathrm{Re}d}}{\rho_{NIR} + \rho_{\mathrm{Re}d}} = NDVI$

Relief (r)

Topography is the most commonly used soil factor in soil genesis studies related to soil catena (Birkeland, 1999; McBratney et al., 2003). Often, topography is the soil forming factor that expresses the most variation at the field scale and, therefore, must be addressed. Soil depth and soil production vary with slope and curvature (Heimsath et al., 1997). Aspect is an important control of soil microclimate and vegetation (Jenny, 1980). Gessler et al. (2000) demonstrated that compound topographic index (CTI), also referred to as the steady-state wetness index, is related to soil properties, such as A horizon thickness: $CTI = \ln \frac{A_s}{\tan \beta}$

where A_s is the specific catchment area (area [m^2] per unit width [m]) and β is the slope angle.

Time (a)

Time is the most difficult soil-forming factor to represent explicitly (McBratney et al., 2003; Noller, 2010) and at best only relative time can be assumed without performing complicated and expensive dating procedures, e.g. luminescence, cosmogenic isotope dating, etc. Also, ages from these procedures would need to be interpolated or used to assume the age of entire surfaces. Specific geomorphic surfaces and positions may approximate relative age (Noller, 2010). Scull et al. (2005) represented time implicitly with Landsat imagery which can detect desert varnish.

Spatial Position (n)

Considering a pixel value in the context with its neighbors, or its spatial context, can enhance classification (Moran and Bui, 2002). Humans observe patterns, tonal differences, edges, etc. simultaneously. Generally, computer software can only consider each pixel individually and often cannot see the forest for the trees and/or the space between the trees. To add spatial relationships to the data set, various texture transformations can be performed. A simple first-order statistical example is the use of a low pass filter. A pixel is assigned the average value of the neighboring pixel values. Variance and entropy are other first-order statistics in the spatial domain (Jensen, 2005). Similarly, Saunders (2005) buffered individual sample points to capture the range of characteristics within a map unit.

Relative position can be related to specific soil forming factors, such as proximity to a steep mountain slope or categorical distinctions. When strongly contrasting pedogeomorphic or geologic regimes exist, McBratney et al. (2003) suggested that stratifying the study area into smaller physiographic regions may simplify the modeling process. Similar to a climosequence, where all factors are said to be constant except climate, stratification isolates regions where one or more factors are generally the same or similar. This allows the modeler to focus on the specific relationships between the soil and the environmental factors within each physiographic region (Di Paolo and Hall, 1983; Birkeland, 1999). Zhu (2000) stratified his study area by geology type before modeling the distribution of soil series using artificial neural networks. Scull et al. (2005) stratified their study area into two distinctive physiographic regions, basin and mountain regions, improving the overall accuracy of their prediction with decision tree analysis (DTA).

Soil Spatial Prediction Functions

Initially, geographic information system (GIS) technologies for soil mapping focused on stratifying the area or creating topographic data layers derived from digital elevation models (e.g., slope and aspect), which assisted the soil scientist in line placement on soil maps (Di Paolo and Hall, 1983; Amen and Foster, 1987). Since then, several workers have applied statistical models to predict both soil types and properties across the landscape (McBratney et al., 2003). Much of the current research in soil predictive models revolves around the method(s) by which the data are analyzed and the uncertainty of resulting predictions. Faster computers and more efficient software continually allow us to explore new techniques of data analysis.

Both McBratney et al. (2003) and Scull et al. (2003) extensively reviewed digital (predictive) soil mapping studies. Scull et al. (2003) generalized predictive soil mapping approaches into four categories: geostatistical methods (e.g., kriging), statistical methods (e.g., generalized linear models), decision tree analysis (e.g., classification and regression tree analysis), and expert systems (e.g., SoLIM [Shi et al., 2004]). Generally, geostatistical and statistical methods are used in predicting soil attributes (S_a), where the predicted outputs are continuous values (Scull et al., 2003). Logistic regression has been used to predict the presence or absence of soil features, such as an E horizon, and fuzzy logic has been used to predict soil classes (Odeh et al., 1992; Gessler et al., 1995).

Decision tree analysis and expert systems have usually been used to predict discrete soil classes (S_c).

The pedogenic understanding raster-based classification (PURC) method was developed by Cole (2004), where conceptual models of the soil-landscape relationships were explicitly defined. Spatially explicit topographic and remotely sensed data represented the soil forming factors. Initially, Cole produced unsupervised and supervised classifications of the individual data layers to identify patterns within the data. The exploration and analysis of these data layers also guided future sampling, by allowing a soil scientist to observe which of these patterns may be meaningful. Rules for classifying data to represent conceptual models were created to predict soil distribution on the landscape.

The Soil-Land Inference Model (SoLIM) also incorporates expert knowledge from soil scientists (Zhu, 2000; Shi et al., 2004). Originally, Shi et al. (2004) had the soil scientist explicitly define the rules. But, they found that this approach can be complicated as it is difficult to explicitly make quantitative rules from tacit knowledge. Another complication arises from the assumption of independence between variables. This interplay between soil forming factors complicates the process of defining rules for each individual variable. To overcome this, Shi et al. (2004) took a case-based reasoning (CBR) approach. Similar to supervised classification, the soil scientist selects points and/or polygons within the GIS as training sites.

The significant difference between expert systems and decision trees is who makes the rules, the user or the data. Decision trees are data driven, where the data set (sample data) is divided with the objective of separating the data set into pure or homogenous classes. At each node, an independent variable that most cleanly divides the data set is chosen. Recursively, splits are made with the objective of creating homogenous groups. To mitigate over-fitting, trees are pruned, where branches are cut back to a higher node. After the tree is grown using training data, unknowns are thrown down the tree, and the class for that point is determined.

Moran and Bui (2002) used decision trees as a manner of machine learning or data mining. They input spatial data for multiple environmental variables to see if the computer could derive a set of rules to mimic or re-create the soil map of a previously mapped area. They were able to remap the area with 70% agreement with the original mapping of the soil scientists.

Moran and Bui (2002) also employed boosting and area-weighting, which increased the accuracy and qualitative look of their prediction. Area-weighting samples in proportion to the spatial extent of the class, which can only work when the extent of the class is previously known. Boosting is a technique that reduces bias. Initially, a single tree is grown with all cases receiving equal weight. A new tree is grown where misclassified cases are given more weight relative to correctly classified cases. This process is repeated a user-specified number of times. Each pixel is assigned to a class based on the modal result or “majority of votes” from the trees (Moran and Bui, 2002).

If the computer can capture the same soil patterns on the map, i.e. the conceptual model developed by the soil scientist, then the tree (the set of rules) could potentially predict the soils in unmapped areas. Scull et al. (2005) was able to take the next step with decision trees, extracting randomly sampled points from an existing soil survey, to train the trees and then extrapolate into areas that were not previously mapped.

Saunders (2005) was able to model soil map units as part of a third order soil survey in Wyoming in a previously unmapped area using classification tree analysis. The sample points were observations made in the field. To capture the range of environmental variables within a map unit, a buffer was set up around each point to sample the data layers (independent variables), reducing prediction error.

Developed by Breiman and Cutler (2009), random forests (RF) is an ensemble of classification and regression trees (CART). Random forests is said to be as accurate as or better than adaptive boosting, yet computationally faster (Breiman, 2001; Gislason et al., 2006). Instead of growing just one tree, many (hundreds to thousands) unpruned, independent trees are grown. This ensemble of trees is referred to as a grove. Each tree is trained from an independent and random bootstrap sample, where a random subset of the sample is used to grow (train) the tree and the remaining points are left ("out of bag sample") to test or validate the tree. Also, at each split, a random subset of predictor variables is chosen (e.g., if there were 100 predictor variables, a subset of ten could be selected at random). From this random subset the strongest variable is selected to split the data. Because of the random bootstrap sample and the random subset of predictive variables at each node, random forests is said to be doubly random. Unlike boosting, each tree is grown independent of each other to the maximum depth (no pruning). Like boosting, the modal result of the entire grove determines the class membership. By making many weak, independent trees, random forests discern patterns in the data that otherwise may be overlooked when few strong trees are grown.

STUDY AREA DESCRIPTION

The study area is in the Basin and Range physiographic province, northwest of Milford in Beaver County, Utah (Figure 1).

The study area encompasses The Big Wash watershed and adjoining areas south of the Beaver-Millard County line, east of the crest of the San Francisco Mountains and west of the Beaver River bed (Beaver Bottoms), covering ~47,000 ha (~117,000 ac) (Figure 2). Each of the following sections addresses the five soil-forming factors of climate, organisms, parent material (geology), relief, and time within the study area.

Climate (c)

Situated between the Sevier and Escalante Deserts, the study area has an arid continental climate, with warm summers and cold winters. Precipitation estimates, from PRISM (Parameter-elevation Regressions on Independent Slopes Model) developed by Oregon State University, range from 20 cm (8 in.) at the Beaver Bottoms to 41 cm (16 in.) atop the San Francisco Mountains (see Figure 2) (USDA-NRCS, 2000). Milford has the nearest climate station, located at the airport with an elevation of 1533 m (5030 ft). National Climate Data Center (NCDC) 1961-1990 normal for Milford are 25.0 cm (9.84 in.) of annual precipitation, mean annual temperature of 9.3°C (48.8° F), mean summer temperature of 21.4° C (70.5° F) and -1.7° C (29.0° F) mean winter temperature (WRCC, 2005). The wettest months are March and April, when storms from the Pacific Ocean bring widespread rain and snow events. The driest time of the year is June into the beginning of July. Monsoonal moisture enters into the area in late July into September.

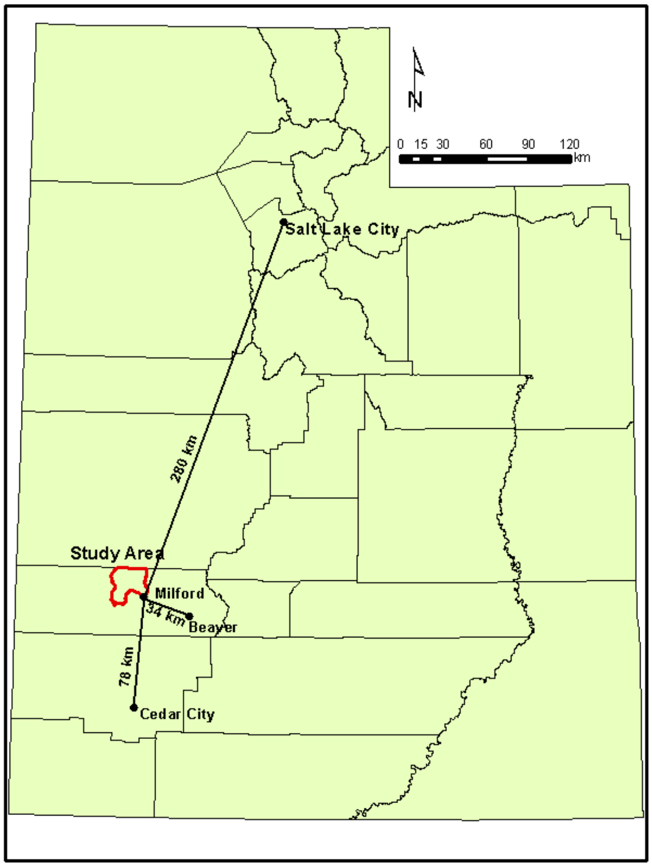


Figure 1. Study area location in the state of Utah.

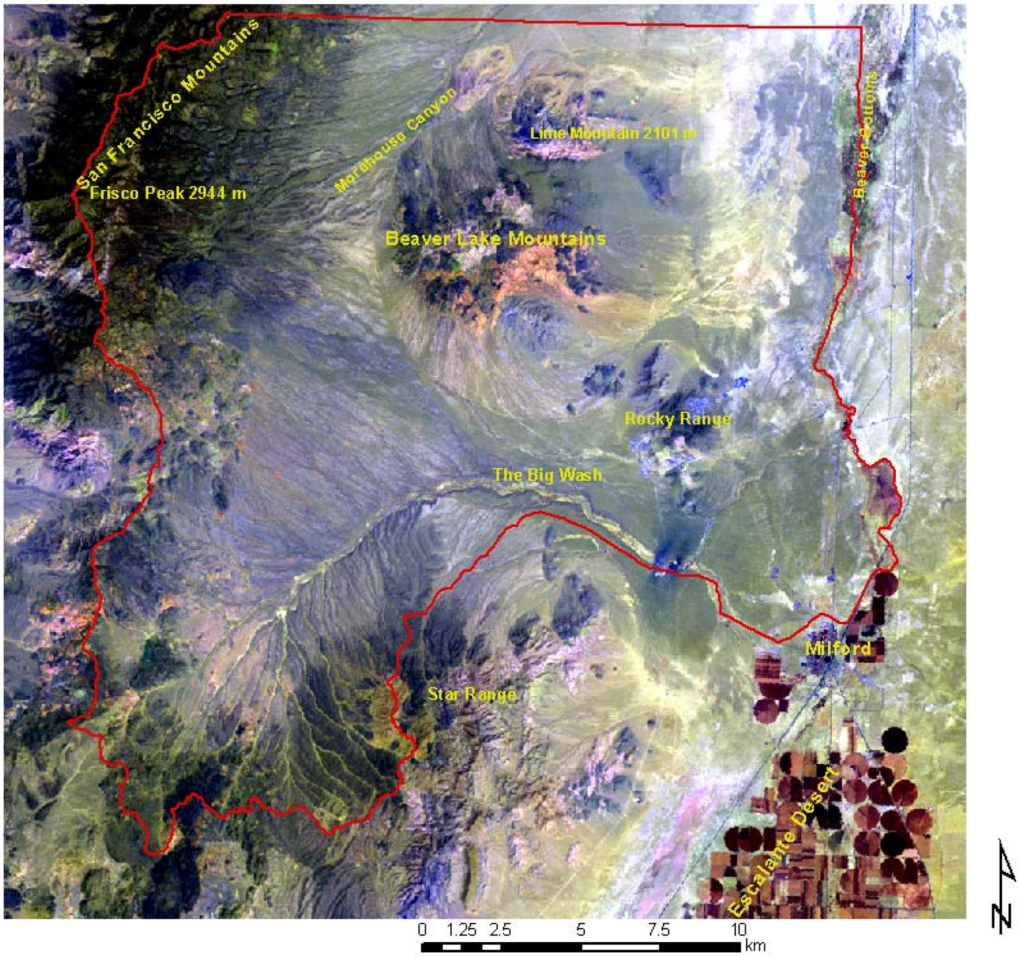


Figure 2. The Big Wash study area shown in a Landsat 7 scene false color (bands 5, 7, 1).

Precipitation during the summer is often associated with intense, convective thunderstorms which are often isolated.

The soil moisture regime across most of the area is aridic bordering on xeric (xeric aridic); meaning the soil is dry 50 to 75 percent of the time when the soil temperature is above 5ºC (Soil Survey Staff, 2003). In Utah, areas that are xeric aridic generally receive 8-12 inches of precipitation (Kent Sutcliffe, USDA-NRCS Utah, personal communication, 2005). The soil moisture regime of the Beaver Bottoms is typic aridic (dry >75 percent of the time when the soil temperature is above 5ºC). Much of the San Francisco Mountains have a xeric soil moisture regime (dry for 45 or more consecutive days in the four months following the summer solstice and moist for 45 or more consecutive days in the four months following the winter solstice). The soil temperature regime is mesic (mean annual soil temperature of 8-15ºC with ≥6ºC difference between mean summer and mean winter soil temperatures) across the whole area, except for the top of the San Francisco Mountains where it is frigid (mean annual soil temperature <8ºC with ≥6ºC difference between mean summer and mean winter soil temperatures).

Organisms (o) – Vegetation

Vegetation in the Great Basin can be an important indicator of soil and climate characteristics. The vegetation in the study area has been grouped into four broad categories of commonly geographically associated species. Scientific names are from Winward (2004) or the Range Plants of Utah web page (USU Extension, 2009). The first group is a salt-desert community, generally found in the valley bottoms and playas where soils are often saline, finer textured, and more alkaline (pH >8.5) in the rooting zone. The plants are shadscale (*Atriplex confertifolia*), black greasewood (*Sarcobatus vermiculatus*), four-wing saltbush (*Atriplex canescens*), budsage (*Artemisia spinescens*), winterfat (*Krascheninnikovia lanata)*, and squirrel tail (*Elymus elymoides*).

The second group is a sagebrush scrubland, which is the most prevalent vegetation type in the study area. The plants include black sage (*Artemisia nova*), Wyoming big sage (*Artemisia tridentata* ssp. *wyomingensis*), basin big sage (*Artemisia tridentata* ssp. *tridentata*), spiny hopsage (*Grayia spinosa*), pygmy sage (*Artemisia pygmaea*), winterfat (*Krascheninnikovia lanata)*, squirrel tail (*Elymus elymoides*), needle-and-thread (*Hesperostipa comata*), indian rice grass (*Achnatherum hymenoides*), galleta (*Pleuraphis jamesii*), scarlet globemallow (*Sphaeralcea coccinea*), cliffrose (*Purshia stansburiana*), rubber rabbitbrush (*Chrysothamnus nauseosus*), Douglas rabbitbrush (*Chrysothamnus viscidiflorus*), broom snakeweed (*Gutierrezia sarothrae*), ephedra (*Ephedra viridis*), and various species of Penstemon, Phlox, and Eriogonum.

The higher elevation terrain that surrounds the area is covered by open woodland of Utah juniper (*Juniperus osteosperma*) and singleleaf pinyon (*Pinus monophylla)*. The understory vegetation is black sage (*Artemisia nova*), Wyoming big sage (*Artemisia tridentata* ssp. *wyomingensis*), antelope bitterbrush (*Purshia tridentata*), bluebunch wheatgrass (*Agropyron spicatum*), lupine (*Lupinus* sp.), Indian rice grass (*Achnatherum hymenoides*), and needle-and-thread (*Hesperostipa comata*).

The higher elevations of the San Francisco Mountains are predominantly covered by woodland composed of singleleaf pinyon (*Pinus monophylla),* Rocky Mountain juniper (*Juniperus scopulorum*), mountain big sage (*Artemisia tridentata* var. *pauciflora),* ponderosa pine (*Pinus ponderosa*), curlleaf mountain mahogany (*Cercocarpus ledifolius)*, limber pine (*Pinus flexilis*), lupine (*Lupinus* sp.), aspen (*Populus tremuloides*), and white fir (*Albies concolor*).

Parent Material (p) – Geology

The soils in the study area have formed in parent materials derived from three distinct lithologies: metamorphic rocks from the Proterozoic, sedimentary rocks from the Paleozoic and early Mesozoic, and igneous rocks from the Tertiary and early Quaternary (Table 3).

Proterozoic to Early Cambrian

The oldest rocks exposed in the area make up the summit crests of the San Francisco Mountains to the west and the very northern end of the Beaver Lake Mountains (East, 1966; Woodward, 1973; Hintze et al., 1984). These rocks represent six concordant units, from Proterozoic to early Cambrian, formed from initial deposits of the Cordilleran miogeosyncline: Cambrian Prospect Mountain Quartzite, Pre-Cambrian Mutual Quartzite, Pre-Cambrian Inkon Slate, Proterozoic Caddy Canyon Quartzite, undivided Proterozoic Papoose Creek Argillite and Proterozoic Blackrock Canyon Limestone, and the upper member of Proterozoic Pocatello Quartzite (Woodward, 1973). While the map by Hintze et al. (1984) indicates that Frisco Peak is composed of Mutual Quartzite, a purple conglomerate quartzite, my observations indicate it is more likely the light pink to tan quartzite, Prospect Mountain Quartzite.

Table 3. Geologic chronology (U.S. Geological Survey Geologic Names Committee, 2007).

Eon	***Era***	***Period***	***Age [Ma]***
Phanerozoic	Cenozoic	Quaternary	Present to 1.8
		Tertiary	1.8 to 65.5
	Mesozoic	Cretaceous	65.5 to 145.5
		Jurassic	145.5 to 199.6
		Triassic	199.6 to 251.0
	Paleozoic	Permian	251.0 to 299.0
		Carboniferous	299.0 to 359.2
		Devonian	359.2 to 416.0
		Silurian	416 to 443.7
		Ordovician	443.7 to 488.3
		Cambrian	488.3 to 542.0
Proterozoic			542.0 to 2500

Paleozoic

Throughout most of the Paleozoic, the study area was covered by shallow seas and lagoons that ultimately deposited several dolomite and limestone formations. The Cambrian Orr Limestone, late Cambrian to early Ordovician Notch Peak Limestone Cherty Marble, Ordovician Pogonip Limestone, Ordovician Kanosh Shale, and Ordovician Watson Ranch Quartzite are exposed on the lower eastern flanks of the San Francisco Mountains and in the northern part of the Beaver Lake Mountains (Welsh, 1973a, 1973b; Hintze et al., 1984; Lemmon and Morris, 1984; Best et al., 1989). Also occurring in the Beaver Lake Mountains are the Silurian Laketown Dolomite, Devonian Sevy Dolomite, Devonian Siminson Dolomite, Devonian Crystal Peak Dolomite, and Mississippian Monte Cristo Limestone (Welsh, 1973a, 1973b; Lemmon and Morris, 1984). The southern end of the Rocky Range has Permian Toroweap Limestone and undifferentiated Permian Kaibab-Plympton Limestone (Baer, 1973; Welsh, 1973a, 1973b; Best et al., 1989).

Mesozoic

During the Triassic, the formative environment transitioned from oceanic deposition to continental processes. This transition was recorded in the Triassic Moenkopi Mudstone interlayered with Limestone, the remnants of a broad coastal plain (Hintze, 1993). Continental rocks, such as shale, siltstone, sandstone, and conglomerate, were deposited in the Chinle flood plain in the Late Triassic, as the region rose above sea level. On the western slope of the Star Range, Late Triassic to Early Jurassic Navajo Sandstone is believed to be the remnant of a coastal-inland dune field (Hintze, 1993).

Some portions of the Navajo formation were silicified into dense quartzite and may be confused with Proterozoic Prospect Mountain Quartzite or Permian Talisman Quartzite (Baer, 1973; Best et al., 1989; author's observations).

The landscape started to take on some familiar forms late in the Cretaceous as the area began to rise during the Sevier Orogeny, part of the Cordilleran Orogeny (Fiero, 1986). The North American plate overrode the Farallon Plate, compressing the region, metamorphosing Proterozoic Pocatello Quartzite through Cambrian Prospect Mountain Quartzite (Woodward, 1973; Fiero, 1986). These older rocks were then pushed over younger Late Cambrian to Ordovician sedimentary rocks at the Frisco Thrust (East, 1966; Woodward, 1973). Brecciated material and slip faces can be observed at the contact of the Frisco Thrust (East, 1966, on the west slope; author’s observation on the east slope). It is believed that the older Proterozoic to Cambrian units are allochtonous, having been thrust eastward some 65 to 100 km (40-60 mi.) (East, 1966; Welsh, 1973; Woodward, 1973; Fiero, 1986).

Cenozoic

Tertiary

Uplift during the Late Cretaceous was followed by a long period of erosion from which no major geologic record remains (Fiero, 1986). Evidence of this erosion exists as coarse debris deposits east of the Great Basin region (Stokes, 1988).

Volcanic activity moved eastward through the Great Basin during the Oligocene (Erickson, 1973; Fiero, 1986). The southern part of the study area is the northern extent of the Tonoquints Volcanic Field (Stokes, 1988). Fairly extensive deposits of andesite, quartz latite, and dacitic and rhyolitic ignimbrites are associated with the Tonoquints Volcanic Field. Mineral enrichment in the north is associated with the Wah Wah-Tushar Mineral Belt (Stokes, 1988). Extensive mineral enrichment of granodiorite, quartz monzonite, and Paleozoic carbonates prone to hydrothermal enrichment, occurred in this region (Baer, 1973; Erickson, 1973; Best et al., 1989).

The southern flank of the San Francisco Mountains (Cactus Stock) and an exposed pluton in the southeast corner of the Beaver Lake Mountains and northern Rocky Range are composed of granodiorite and quartz monzonite 28.7 to 31.2 Ma (Welsh, 1973b; Best et al., 1989). Both of these locations have rich deposits of copper ore (Whelan, 1973a). All sedimentary rocks have been thermally metamorphosed in the Rocky Range, as have many in the Beaver Lake Mountains (Whelan, 1973b). Copper deposits in the Beaver Lake Mountains and Rocky Range are still mined today.

Shauntie Hills Andesite, 31-34 Ma, occurs along the southeast corner of the area and on the lower slopes of the Star Range. Large areas of Horn Silver Porphyritic Andesite, 31.6-35 Ma, occur on the lower flanks of the San Francisco Mountains, Beaver Lake Mountains, and Rocky Range (Best et al., 1989).

Several hot pyroclastic flows blanketed large areas south of the Big Wash in ignimbrites and ash fallout, filling valley bottoms (Erickson, 1973; Fisher and Schmincke, 1984; Fiero, 1986; Best et al., 1989). There are three major ignimbrites mapped in the area: Needles Formation, 29.7-32.3 Ma (strongly to moderately welded); Isom Formation, 22.5 Ma (Intensely welded); and the Quichapa Formation 22.3 Ma (moderately to loosely welded) (Erickson, 1973; Best et al., 1989).

The Squaw Peak formation, a coarsely porphyritic latite (23 Ma), occurs in the southwestern perimeter of the area. There are also smaller deposits of volcanic rock litter and some basalt about 13 Ma in age.

Before all volcanic activity ended, the Basin and Range began to subside and stretch (Stokes, 1988). Many of the ridges and basins started forming in this area around 10-15 Ma, during the Miocene (Hintze, 1993). Hundreds of normal faults, running north to south, formed a series of parallel ridges and basins across the Great Basin (Crosby, 1973; Erickson, 1973; Fiero, 1986; Stokes, 1988). It is estimated that the Basin and Range stretched some 100 to 160 km (60-100 mi.) (Stokes, 1988). Block faults in many of the basins may be listric, flattening at the bottom. Extension occurred when hanging blocks moved down relative to the foot wall, while the foot wall was moving horizontally from the hanging wall (Fiero, 1986; Hintze, 1993).

The asymmetric geometry of the San Francisco Mountains evolved from this process. The western slope rises dramatically over the Wah Wah Valley, 1414 m to 2944 m (4639 ft to 9660 ft). The eastern slope drops quickly to around 2010 m (6600 ft) and then gently slopes to 1510 m (4950 ft) over the course of about 16 km (10 mi) (East, 1966). Many of the volcanic bodies formed in the Oligocene were faulted and fractured from extension and local subsidence (Best et al., 1989). Newly formed basins have continually filled in with sediment, accumulating to depths greater than 1000 m in the valley bottom near Milford (Best et al., 1989).

Relief (r) and Time (t) – Quaternary History and Geomorphology

The area can be broken into three representative landforms and soil-forming environments: San Francisco Mountain Range and Beaver Lake Mountains, The Big Wash Basin (fan piedmont), and the valley bottom below the Lake Bonneville shorelines.

The slopes of the San Francisco Mountains are deeply mantled by colluvium and several large talus aprons are visible from several km away. The range also has several prominent cliff bands. Rock fall, rock avalanches and frost wedging seem almost certain to occur on these slopes. The colluvium is very angular and the entire range has an average slope just greater than 40%. Quartzite gravel from the San Francisco Mountains has been carried several km from the mountain range along drainages. There is no evidence of glaciation anywhere in the study area.

The Beaver Lake Mountains, Rocky Range, and the Shauntie Hills have much less relief, rising 1800 to 2300 m (6000-7500 ft) in elevation. They also exhibit lower gradients and shallower colluvial deposits than the San Francisco Mountains. Overland flow and diffusive transport of sediment seem more prevalent than mass movement. The Rocky Range has several active alluvial fans on both its eastern and western slopes. There are several alluvial fans and slopes coming off the neighboring Beaver Lake Mountains, and many are relict fans, having been deeply incised (see Figure 3). Many of the alluvial features in the survey appear to be relict features.

The Big Wash Basin and adjoining watersheds are a patchwork of alluvial fans, alluvial slopes, relict fans and alluvial surfaces, pediment surfaces, and numerous gullies and washes. Many of the alluvial features are highly incised, isolating higher surfaces.

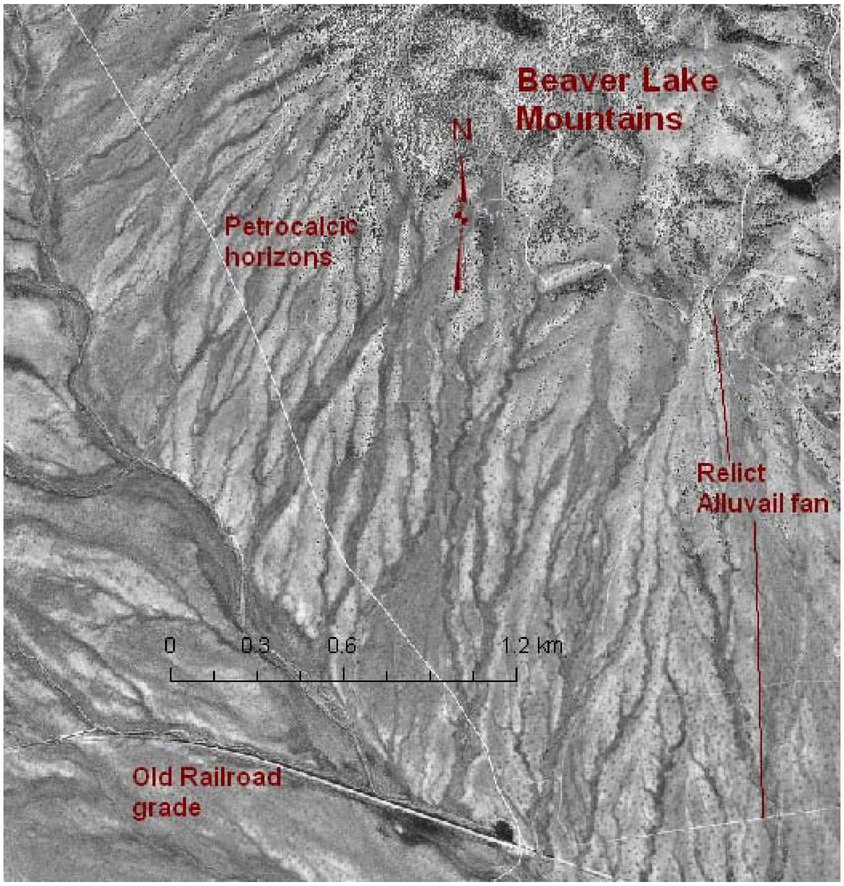


Figure 3. Aerial photograph of the SW slope of the Beaver Lake Mountains. Relict alluvial fans are being incised and higher surfaces are isolated.

These higher surfaces have well developed soils, further suggesting that they are relict features (Figure 3).

The large washes in the study area are underfit streams: steeply walled with wide, flat bottoms and relatively small active channels. When water does run in these drainages, the flows quickly dissipate, seeping into the coarse sediment. The average clast size increases upstream.

The slopes below the Star Range are mapped as Quaternary Alluvium. Upslope, the Lamdorf Tuff member of the Needles Formation and some undifferentiated volcanic rock are mapped (Baer, 1973; Best et al., 1989). Some hillslopes appear to have bedding plane morphology. Deeply incised gullies and ridges run parallel up the slope. Most southwestern slopes are steep, often exceeding 30%, whereas northeastern slopes are more gradual, 5-15%. The steeper slopes have an abundance of surface gravel and cobbles and the soils are skeletal (>50% rock fragments). The shoulder positions have appreciably less surface rock fragments, the soils are still skeletal (35-50% rock fragments) and with rock fragments that are thickly covered with silica and carbonate pendants.

The Big Wash has exposed the toe of one of these ridges. The bedding planes are parallel and have a dip that appears to be reflected in the hillslope geometry, which is possible evidence of faulting since the material has been deposited. The rock fragments appear sorted (pea gravel) with an occasional large cobble. The matrix is a pink-grey fine sediment, very fine sand or finer.

During the Pleistocene, Lake Bonneville filled up the basins of western Utah and smaller portions of eastern Nevada and southern Idaho. Lake Bonneville continued to rise until its shoreline reached a maximum elevation of 1561.9m (5124.3ft), a depth of 73m (239ft) from the local valley bottom at about 15Ka. At this level, the lake etched a distinct shoreline known as the Bonneville high stand. The valley bottom was merely a small inlet on the very southern end of Lake Bonneville.

The Bonneville high stand is not the same elevation from north to south in the study area. There is approximately a 4 m (13ft) difference between 1558.2 m in the south, to 1561.9 m in the north. The entire basin of Lake Bonneville was upwarped due to isostatic rebound when the lake drained and evaporated (Gilbert, 1890; Crittenden, 1963). The distribution of the deformation across the state of Utah is fairly elliptical, the major axis running north to south. The valley bottom in the study area is along the south end of the major axis of deformation. This, combined with the valley bottom being relatively narrow, 6.5 to 21 km wide, resulted in negligible deformation east to west within the study area. The result is nearly linear deformation north to south in this valley.

Because of the presence of Lake Bonneville, soil formation below the Bonneville shoreline was reset to time 0 approximately 15ka ago. Therefore deposits below the Bonneville shorelines can be assumed to be lacustrine materials and recent alluvium from the Beaver River and other drainages with a geomorphic surface age of ~15ka or younger. The relief is generally low and currently diffusive transport (slope alluvium) is the dominant process in action.

Above the Bonneville shoreline, many surfaces have been isolated by a network of gullies and washes. These relict surfaces above the Bonneville shoreline likely predate Lake Bonneville, as they are truncated by shoreline features. The drainages are periodically reworked with high energy flows from summer convective storms. Closer to the mountain front there is evidence of debris flows within the channels.

Soil (s)

Soils have been classified according the 9th edition of Soil Taxonomy (Soil Survey Staff, 2003). Aridisols are the most extensive soil order in the study area, with Typic and Xeric Haplocalcids, Typic and Xeric Calciargids, Typic Natrargids, Calcic Petrocalcids and Durinodic Xeric Calciargids and Haplocalcids covering most of the alluvial fan/piedmont and Lake Bonneville terraces. Entisols also occur, mainly as Torriorthents. Drainage bottoms have weakly developed Haplocalcids or Torriorthents. The mountains and ridges are dominated by Aridisols (Lithic Xeric Haplargids, Xeric Calciargids, Xeric Haplocalcids, and Xeric Lithic Haplocalcids), with minor Entisols (Xeric Lithic Torriorthents). At higher elevations of the San Francisco Mountains, Haploxeralfs are common (Figure 4).

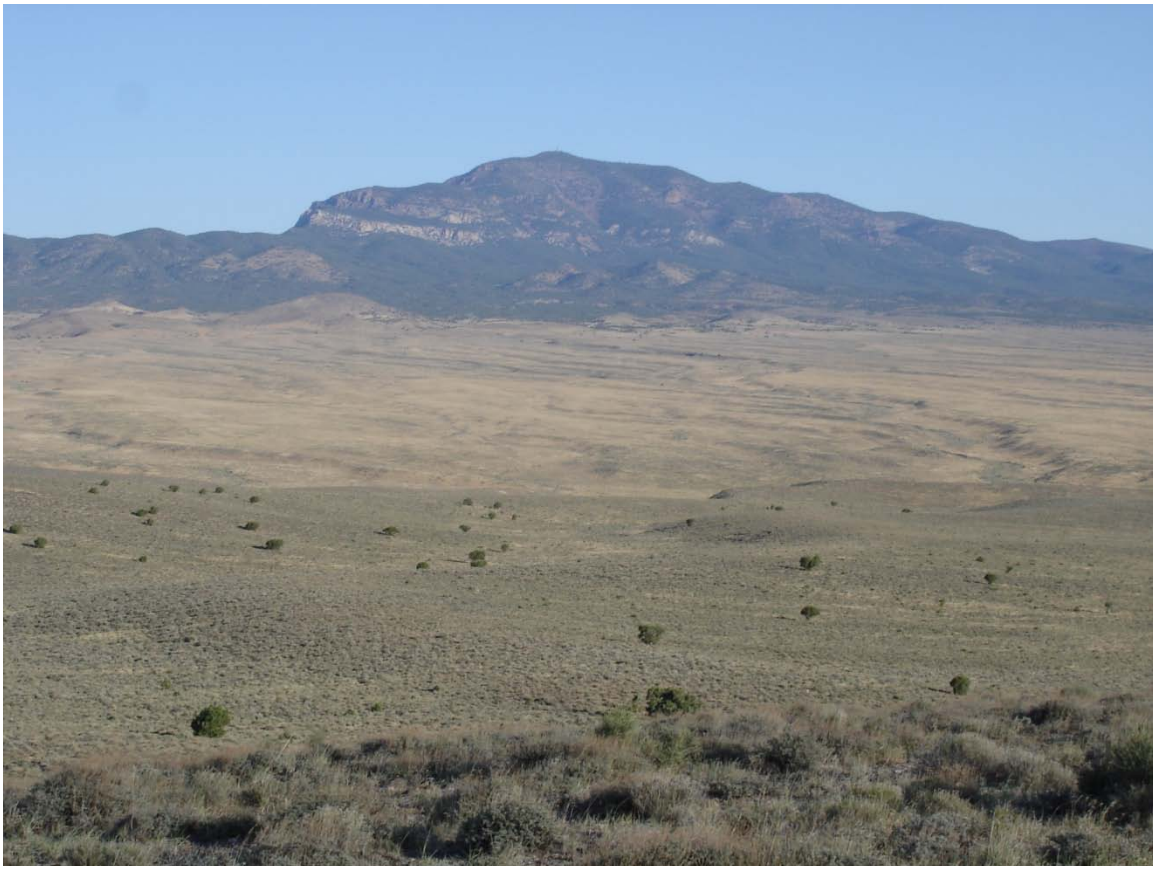


Figure 4. Photograph looking northwest towards the San Francisco Mountains. The Big Wash is in the middle of the picture, with fan remnants coming off the Shauntie Hills in the foreground.

METHODOLOGY

Two types of data were required for this research: digital geospatial data that represent the environmental covariates (soil forming factors) in the scorpan empirical model, and field observations of soil and landscape properties. Field observations were used to train the random forest models. Each field observation was attributed with values from the environmental covariates.

Digital Data

The scorpan environmental covariates in the study area were represented by 22 digital data layers (Table 4). These covariates were principally derived from two types of raster data: Landsat 7 ETM+ and DEM. Much of the processing to prepare the Landsat image was done with ERDAS Imagine 9.1™. The DEMs were processed in ArcGIS 9.2™.

All digital data were projected into Universal Transverse Mercator (UTM) and North American Datum 1983 (UTM 12S North, datum NAD83). All data layers were subset to the rectangular extent of the study area with about a 2-km buffer (Table 5).

Landsat-Derived Data

The entire study area is covered by one Landsat 7 ETM+ scene, path 038 and row 033, acquired July 31, 2000, which was obtained from the Intermountain Region Digital Image Archive Center (IRDIAC, 2006; Figure 5; Figure 6). The Landsat scene was standardized using the cosine theta (COST) method without tau (Chavez, 1996; RSGIS, 2003: script no. 3; Nield et al., 2007). The values for the dark object subtraction were sampled from Fish Lake, Utah, (deep lake) and shadows cast by cumulus clouds. These

Table 4. Environmental covariates represented by digital data.

Covariate	**Source Data**	**Intermediate Data**	**Final Data**
Vegetation	Landsat 7	-	NDVI: Bands (4-3)/(4+3)
Climate	10-m DEM	Elevation	Soil Moisture Regime
		Aspect ($-\pi$ to π)	Xeric vs. Aridic
	Landsat 7	Bands 3 & 4	
Relief	10-m DEM	-	Slope
		-	CTI
		CTI	Filtered CTI (5x5)
		-	Aspect ($-\pi$ to π)
		-	Elevation
	30-m DEM	Filtered DEM (11x11)	Slope Curvature
Parent Material and Soil	Landsat 7	-	Bands 1-5,7
	Landsat 7	-	Normalized Difference Ratios: Bands (4-5)/(4+5) Bands (3-7)/(3+7) Bands (5-2)/(5+2) Bands (5-1)/(5+1) Bands (4-7)/(4+7) Bands (3-1)/(3+1)
Age	10-m DEM	-	Lake Bonneville Shoreline

values were also compared with the histogram for each spectral band to establish the minimum reflectance in the scene.

Individual bands of Landsat 7 (1-5, 7) and several normalized difference band ratios were used to represent the scorpan covariates of vegetation, soil, and parent material in the study area:

$$\frac{\rho_{Band\ A} - \rho_{Band\ B}}{\rho_{Band\ A} + \rho_{Band\ B}} = Normalized\ difference\ band\ ratio$$

where $\rho_{\text{Band A}}$ is the reflectance in Band A and $\rho_{\text{Band B}}$ is the reflectance in Band B. The normalized difference ratio of bands 4 and 3 represented vegetation, known as the Normalized Difference Vegetation Index (NDVI). The normalized difference ratio of bands 5 and 2 distinguished most igneous geologic formations (andesite) from sedimentary formations (limestone). In addition, normalized band ratios 4 and 5, 3 and 7, 5 and 1, 4 and 7, and 3 and 1 exhibited unique patterns wherein distinct landforms and vegetation communities were visually identified and thought to be useful in the model (Cole, 2004; Bodily, 2005; Scull et al., 2005; Nield et al., 2007; Saunders and Boettinger, 2007) (Figure 7, 8, and 9).

Table 5. The bounding coordinates of each independent variable source and the study area.

	Northeast corner	*Southwest corner*
10 m DEM	298484.8 E, 4272046.8 N	328659.8 E, 4243332.3 N
30 m DEM	298471 E, 4272051 N	328651 E, 4243341 N
Landsat 7 ETM+	298456 E, 4272022 N	328696 E, 4243312 N
Study area	300486.4 E, 4271733.5 N	326646.4 E, 4245333.5 N

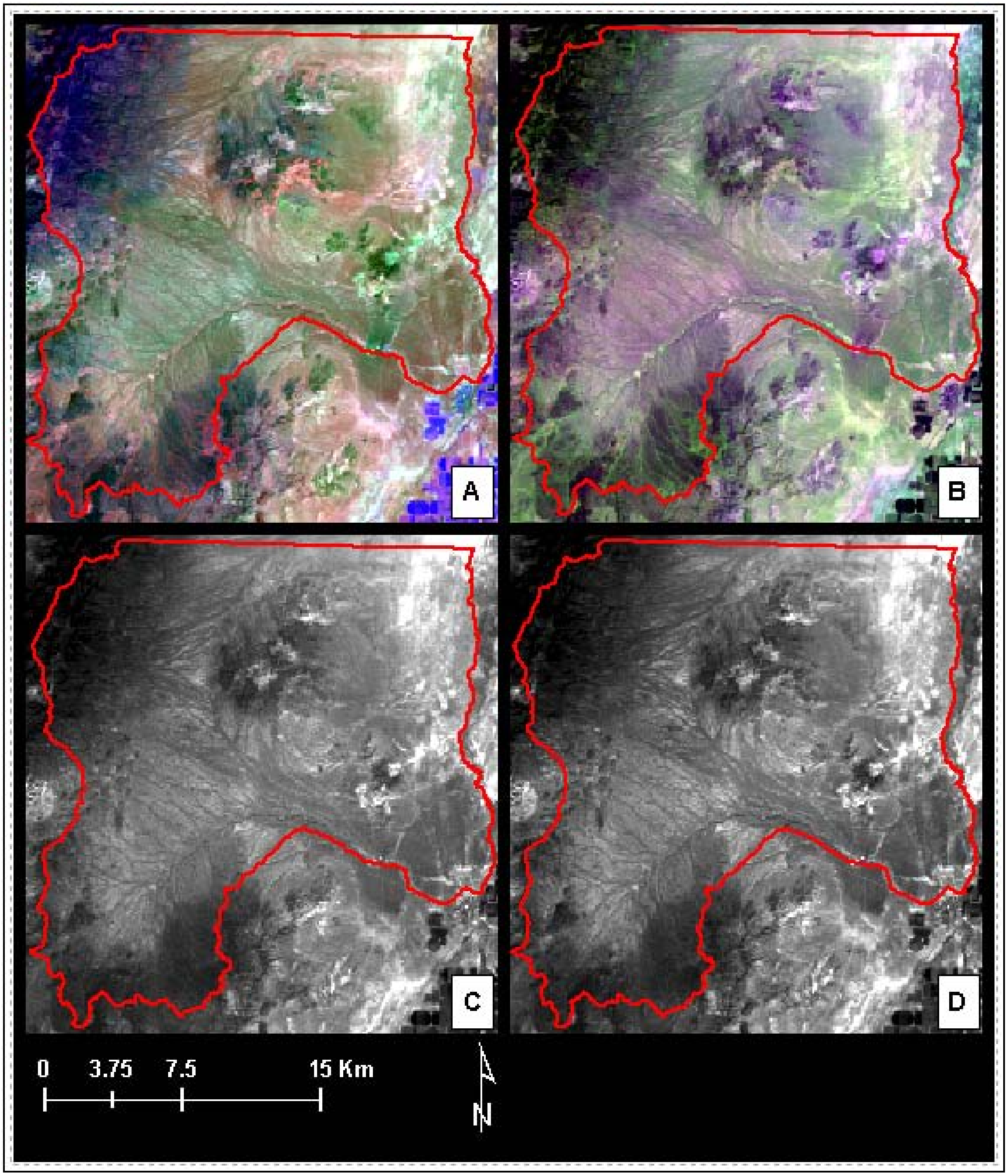


Figure 5. Landsat 7 ETM+ imagery. A: False color composite of bands 5 (red), 2 (green), 4 (blue); B: False color composite of bands 3 (red), 7 (green), 1 (blue); C: band 1; D: band 2.

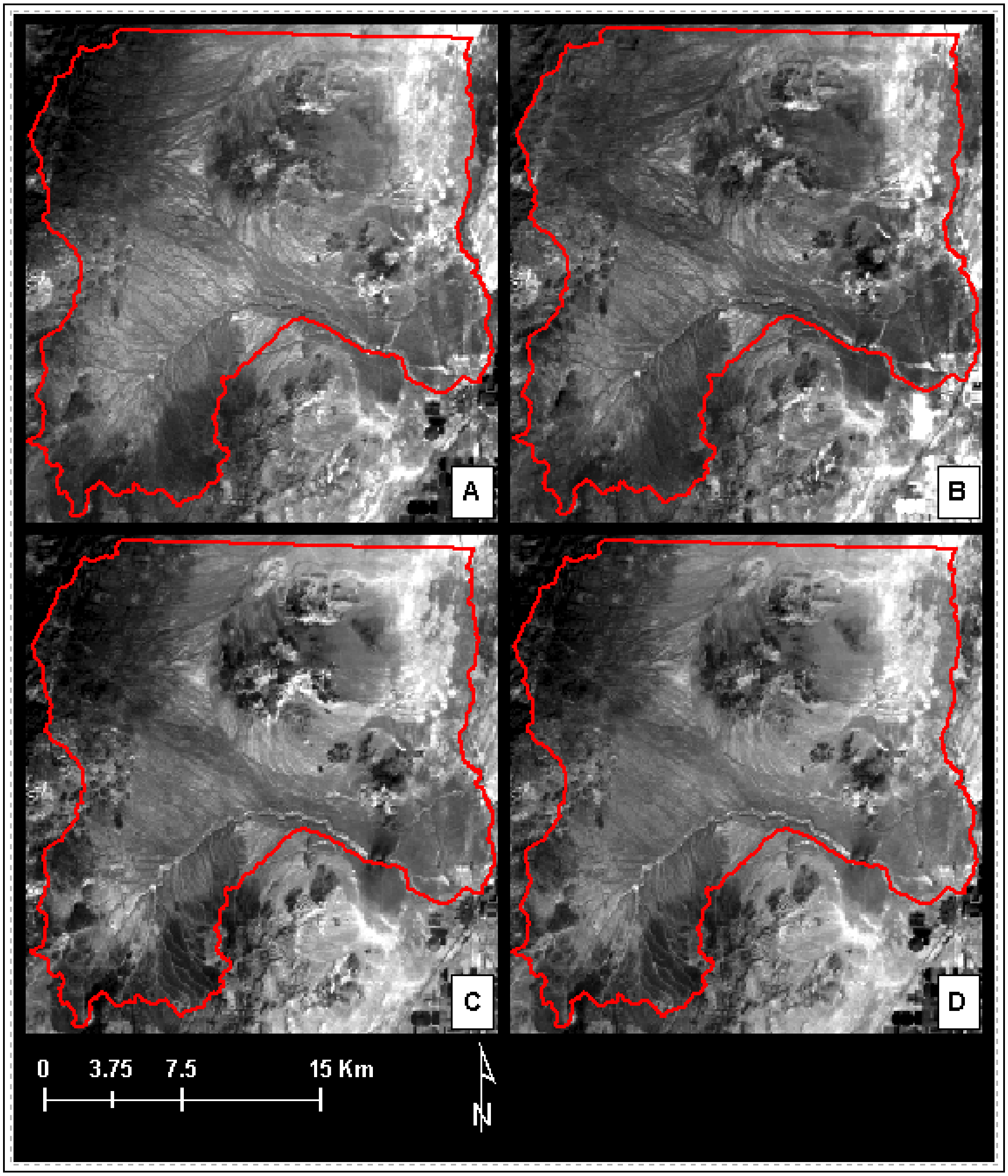


Figure 6. Landsat 7 ETM+ imagery. A: band 3; B: band 4; C: band 5; D: band 7.

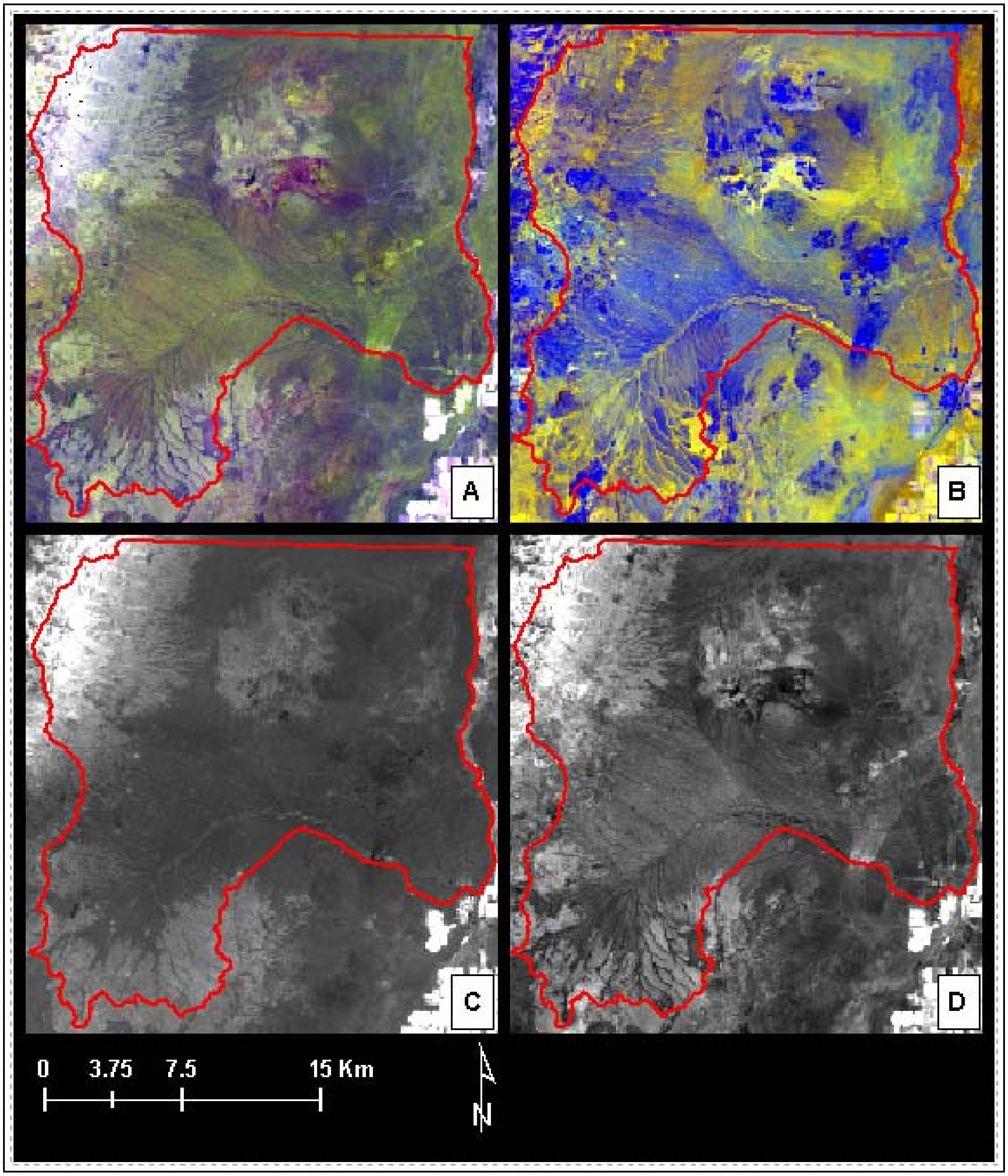


Figure 7. Normalized difference ratios of Landsat 7 ETM+ data. A: False color composite of ratios (4-7)/(4+7) (red), (4-5)/(4+5) (green), (4-3)/(4+3) (blue); B: false color composite of ratios (5-2)/(5+2) (red), (5-1)/(5+1) (green), (4-7)/(4+7) (blue); C: ratio (4-3)/(4+3); D: ratio (4-5)/(4+5).

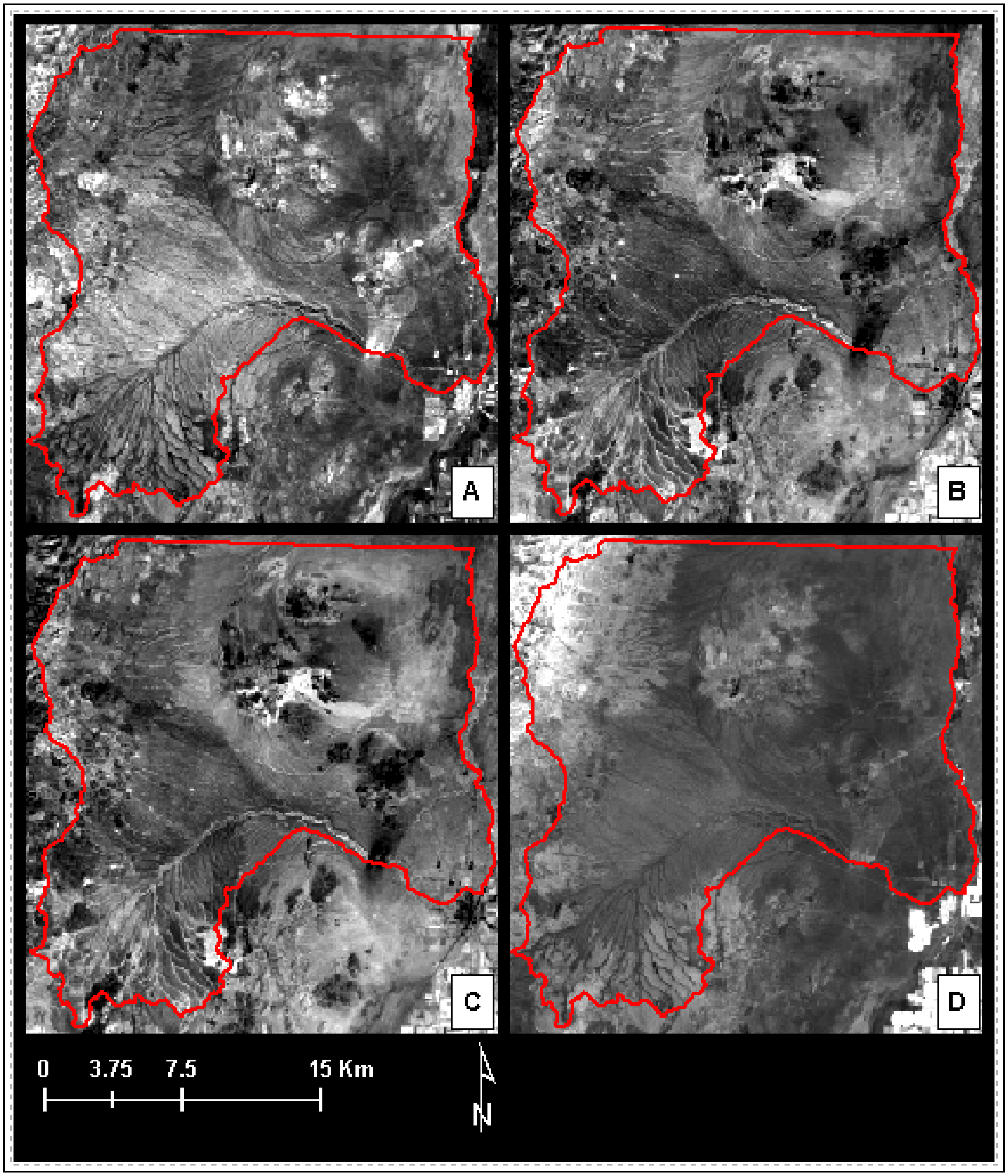


Figure 8. Normalized difference ratios of Landsat 7 ETM+. A: ratio (3-7)/(3+7); B: ratio (5-2)/5+2); C: ratio (5-1)/(5+1); D: ratio (4-7)/(4+7).

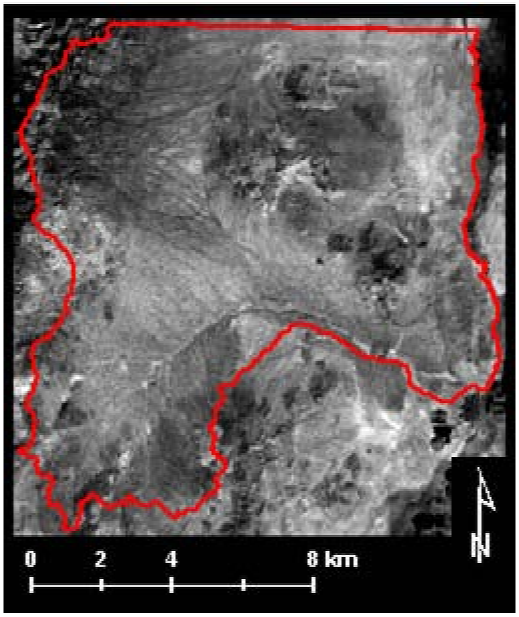


Figure 9. Normalized difference ratio of Landsat 7 ETM+, (3-1)/(3+1).

Digital Elevation Model-Derived Data

Two raster DEM from the national elevation dataset were obtained from the Utah Automated Geographic Reference Center (AGRC, 2008; Figure 10A); one at 9.19-m grid cell resolution (referred to as the 10-m DEM) and another at 30-m resolution. The terrain analysis software, TauDEM (a toolbar addition for ArcGIS), was used to fill sinks in the 10-m and 30-m DEM data sets (Tarboton, 2005). The 10-m DEM was the highest resolution dataset obtainable at the time and offered the most detailed representation of the landscape.

An 11x11 low pass filter was applied to the 30-m DEM to add spatial context to each pixel. For example, consider two pixels that each has a slope of 10 percent: one is on structural bench perched on a steep mountain side and the other is on a small rise on gently sloping fan piedmont. By taking the average across the 330 m by 330 m area the general slope of the landform that each of these pixels are found on can be determined from this filtered 30-m DEM layer.

The flow direction raster was calculated from the 10-m DEM using TauDEM (Tarboton, 1997; Figure 10B), which uses the d-infinite algorithm in the slope algorithm. TauDEM was also used to calculate slope for both the 10-m DEM (Figure 10C) and the filtered 30-m DEM (Figure 10D). An ArcToolbox Spatial Analyst tool was used to calculate curvature of the filtered 30-m DEM (Figure 11A). Compound topographic index (CTI) was derived from the 10-m DEM and the flow direction raster using an ArcInfo avenue script (.aml) (Evans, 2004) (Figure 11C). A 5x5 low pass filter was run over the original CTI to produce an additional filtered CTI layer (Figure 11D).

Aspect and elevation were derived from the 10-m DEM . Aspect was calculated in degrees (0-360º) then transformed to a range of -π to π, where north is –π, south is π, and east and west are equal to 0 (Figure 11B):

$$Transformed\ \ Aspect\ \ N / S = \begin{cases} 0 \quad if \ \ Aspect^\circ = -1 \\ \left(270^\circ - Aspect^\circ\right) / \, 90^\circ \times \pi \quad if \ 180^\circ < Aspect^\circ \leq 360^\circ \\ \left(Aspect^\circ - 90^\circ\right) / \, 90^\circ \times \pi \quad if \ \ 0^\circ \leq Aspect^\circ \leq 180^\circ \end{cases}$$

This transformed aspect is essentially a measure of northness vs. southness. While there are microclimatic contrasts in microclimates between east- and west-facing slopes (aspect = 0), north- and south-facing slopes exhibit more pronounced differences in soil formation (Figure 12).

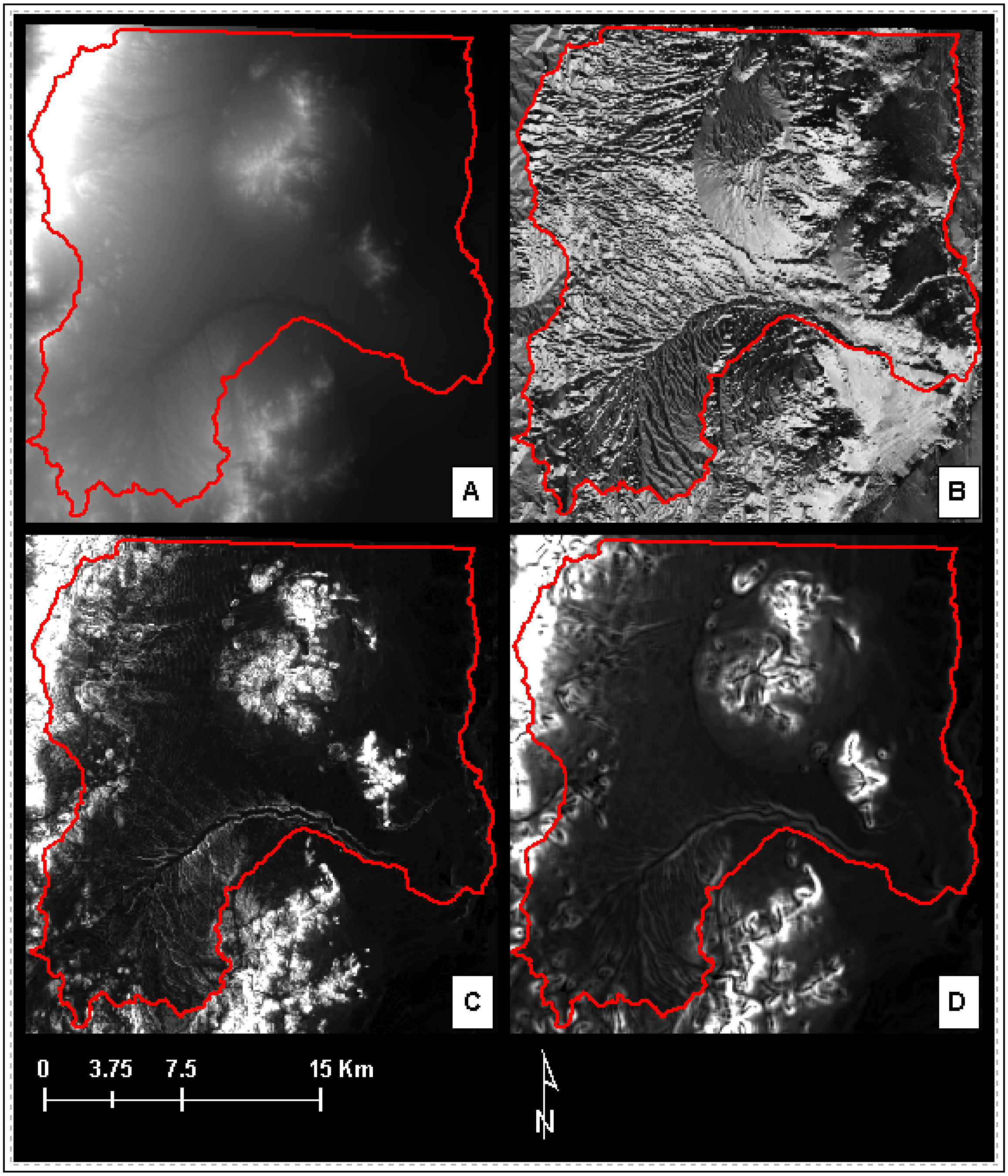


Figure 10. DEM-derived data. A: 10-m DEM; B: Flow direction from 10-m DEM; C: Slope from 10-m DEM; D: Slope from the filtered 30-m DEM.

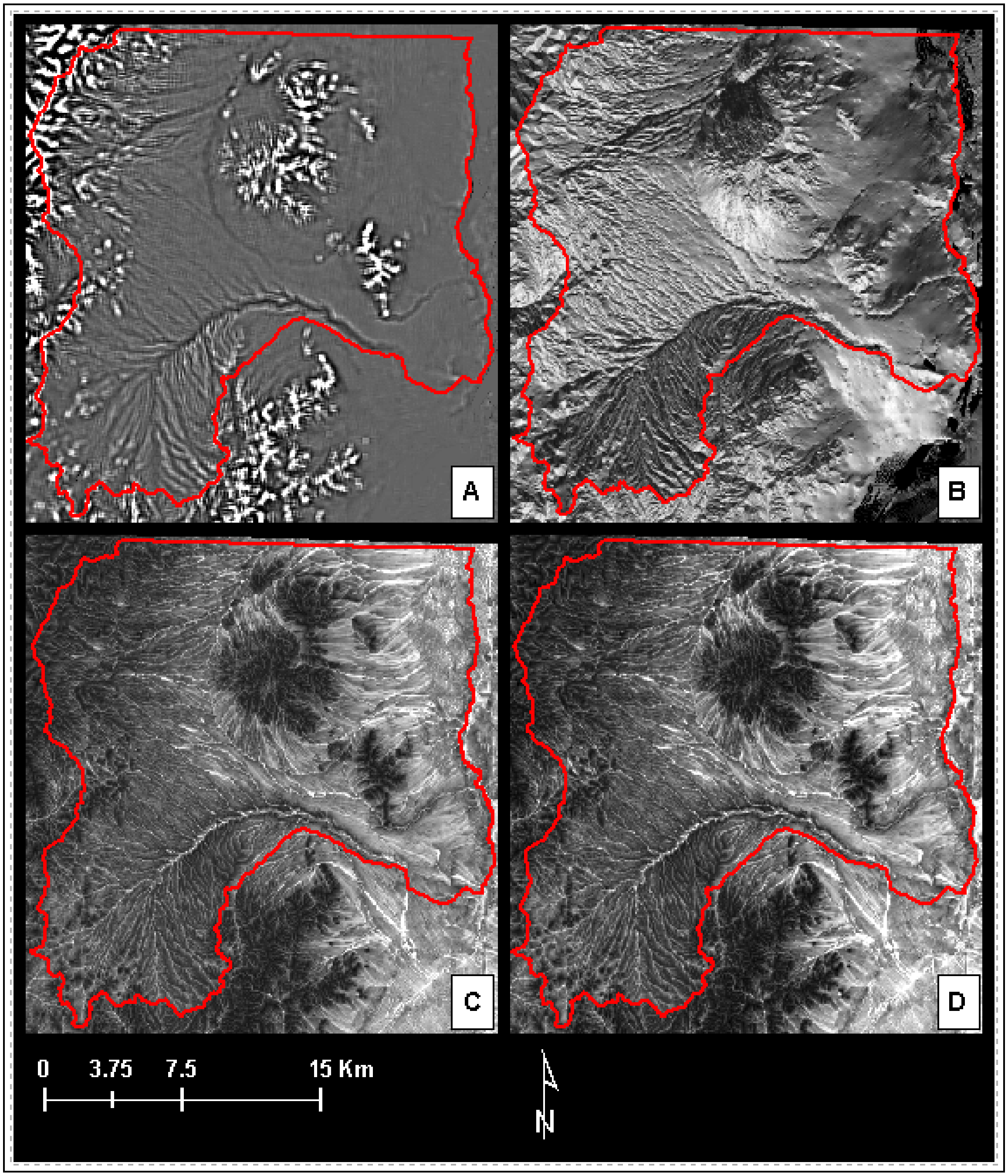


Figure 11. DEM-derived data. A: Curvature from 30-m DEM; B: Transformed aspect from 10-m DEM; C: CTI form 10-m DEM; D: CTI with 5x5 low pass filter from 10-m DEM.

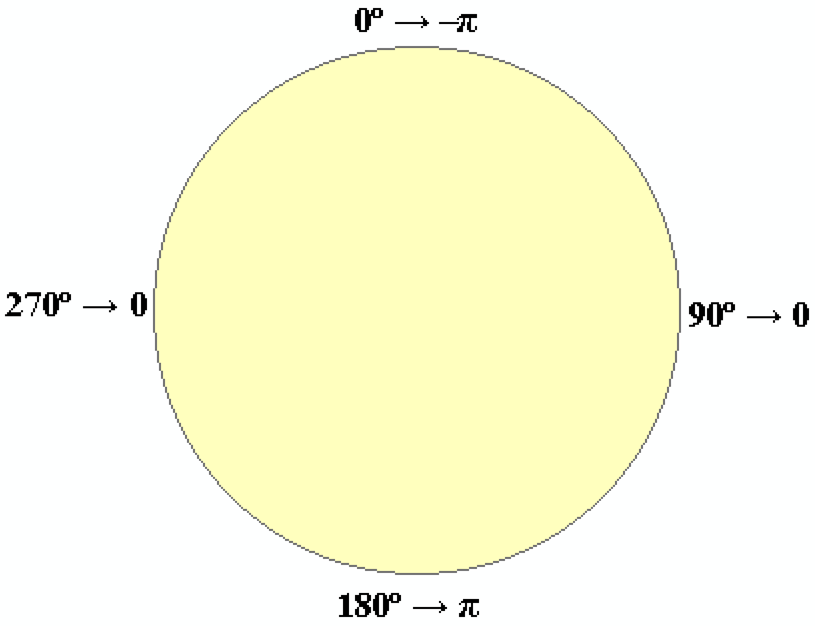


Figure 12. Illustration of the transformation of aspect in degrees to continuous variable of north-south ranging from -π to π.

Digital Data Exploration and Transformation

Unsupervised and supervised classifications of the digital data using Imagine helped identify patterns used to develop conceptual models and guide field data collection, and to develop customized data layers used in the random forest classification. Unsupervised classification requires no *a priori* knowledge of the study area because it is completely driven by the digital data. Class means and clusters are found with the Iterative Self-Organizing Data Analysis Technique (ISODATA). Each pixel is initially assigned to a cluster based on the spectral distance in feature space to the nearest cluster center (mean). Once each pixel has been assigned, a census of each cluster is made. Based on the average pixel value from the census in each cluster, the cluster center is shifted to the new cluster mean to reflect the membership. Once again, all pixels are assigned to the nearest cluster center. This process is recursive, being reiterated until a user specified convergence percentage is reached or a specified number of iterations have been run. The convergence percentage refers to the percentage of pixels that do not change membership, e.g., when 95% convergence is reached, 95 % of the pixels did not change membership after the cluster mean was recalculated (Leica, 2005). Spectral signatures that are identified can be refined using supervised classification.Supervised classification requires *a priori* knowledge. Cluster means for the concept are calculated from the pixel(s) in a training site, which can be a point or a polygon.

Cluster means for classes may also be identified in spectral feature space (2D histogram) (Leica, 2005). When developing classes with the seeding tool, only a few pixels of a given class are sampled. From these pixels, the Imagine software computes the cluster mean of the class from which a parallelepiped is created in n-dimensional feature space. All pixels are then assigned to a class based on Euclidean distance in feature space. In contrast, the user can draw an area of interest (AOI) to select pixels in feature space. The main difference with editing in feature space is that the user is not merely sampling a few pixels of a class but rather the user is literally assigning pixels to a class – essentially this is direct supervision of pixel assignment. One limitation of the feature space analysis is that a multi-dimensional feature space is represented in only 2-dimensions at a time.

Customized Data Layers

Two customized data layers, the Lake Bonneville shoreline and the Xeric-Aridic soil moisture regime (SMR) raster layers, were created to help stratify the study area into distinct pedo-geomorphic regions. The 10-m DEM was incorporated into both Lake Bonneville and Xeric-Aridic SMR models. Landsat 7 data was also used in the Xeric-Aridic SMR model.

There were vector representations available of Lake Bonneville (AGRC, 2008), but they were inaccurate, off by several kilometers from the true shoreline (Figure 13B). While many prominent shoreline features (spits, deltas, shoreline scarps) can be clearly seen in aerial photography there were larger surfaces where shoreline features were not evident, making it difficult to heads-up digitize the shoreline.

The Lake Bonneville layer is a simple binary (true or false) raster layer, where surfaces below ancient Lake Bonneville are “true,” and surfaces that remained above the highest lake level, the Bonneville high stand, are “false.” As explained in the Quaternary History and Geomorphology section, the elevation of this shoreline feature ranged from1558.2 m in the south to 1561.9 m in the north. This northward trend was estimated with simple linear regression. Several prominent shoreline features of the Bonneville high stand were identified in the field and with the aerial photography. These points were attributed with the UTM northing and the elevation value from the 10-m DEM. Using Interactive Data Language (IDL) the elevation trend of the shoreline was estimated to be 1.99×10^{-4} m rise in elevation per meter in distance northward. A 10-m raster representing the hypothetical surface elevation of Lake Bonneville’s shoreline was created where the elevation was calculated as a function of the northing of each cell center (Figure 13A). All raster cells in the 10-m DEM found to be lower than the Lake Bonneville shoreline trend were assigned “true” as they were below ancient Lake Bonneville. The final output

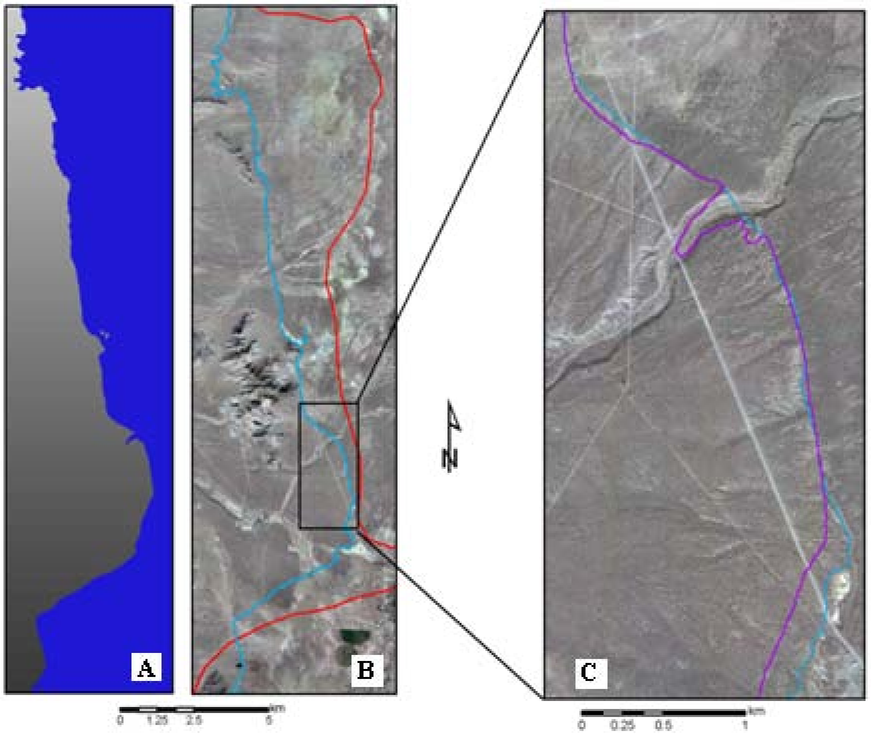


Figure 13. The Lake Bonneville shoreline prediction. A: Hypothetical surface elevation raster of the Lake Bonneville shoreline (gray shading) and the predicted extent of Lake Bonneville (blue). B: Previously available vector layer of the shoreline (red) (AGRC, 2008). Final shoreline output used as a predictive variable in random forests (blue). D: Predicted shoreline feature with linear regression (purple); final edited shoreline (blue).

was vectorized for further editing where minor adjustments (never more than 200 m) were made to match prominent shoreline features (Figure 13C).

The break between xeric and aridic SMR is characterized by single leaf pinyon trees becoming the dominant tree over Utah juniper trees. Spectrally, these two plant communities can be distinguished. Areas of interest (AOI) were delineated over known juniper stands and dominantly pinyon stands in the original Landsat image (non-standardized). These pixels were then identified in a feature space plot (a two dimensional histogram) of Landsat bands 3 and 4 produced in Imagine. Dominantly pinyon and dominantly juniper stand pixels were found to be in two distinct but contiguous clusters in feature space (Figure 14).

Each cluster was then delineated in feature space (Figure 14) to perform a supervised classification with three general classes: 1) vegetation typical of the xeric SMR, including singleleaf pinyon, fir and others (see vegetation section in Area Description), 2) woodlands dominated by Utah juniper, which are characteristic of the xeric aridic SMR, and 3) vegetation typical of the xeric aridic and typic aridic SMR, which are non-woodland or a shrub steppe (e.g. sagebrushes). This output is referred to as the pinyon-juniper (PJ) classification as it was based on the presence of single-leaf pinyon or Utah juniper (Figure 15A).

Not all areas in the xeric SMR are covered by woodland, and areas of irrigated cropland and tamarisk at low elevations in the aridic SMR are spectrally similar to true woodlands at higher elevations. To account for these areas the PJ classification was combined with transformed aspect and elevation in a model (Figure 16). At all aspects, juniper was not observed in the field to occur below 1700 m; also, all farmland and

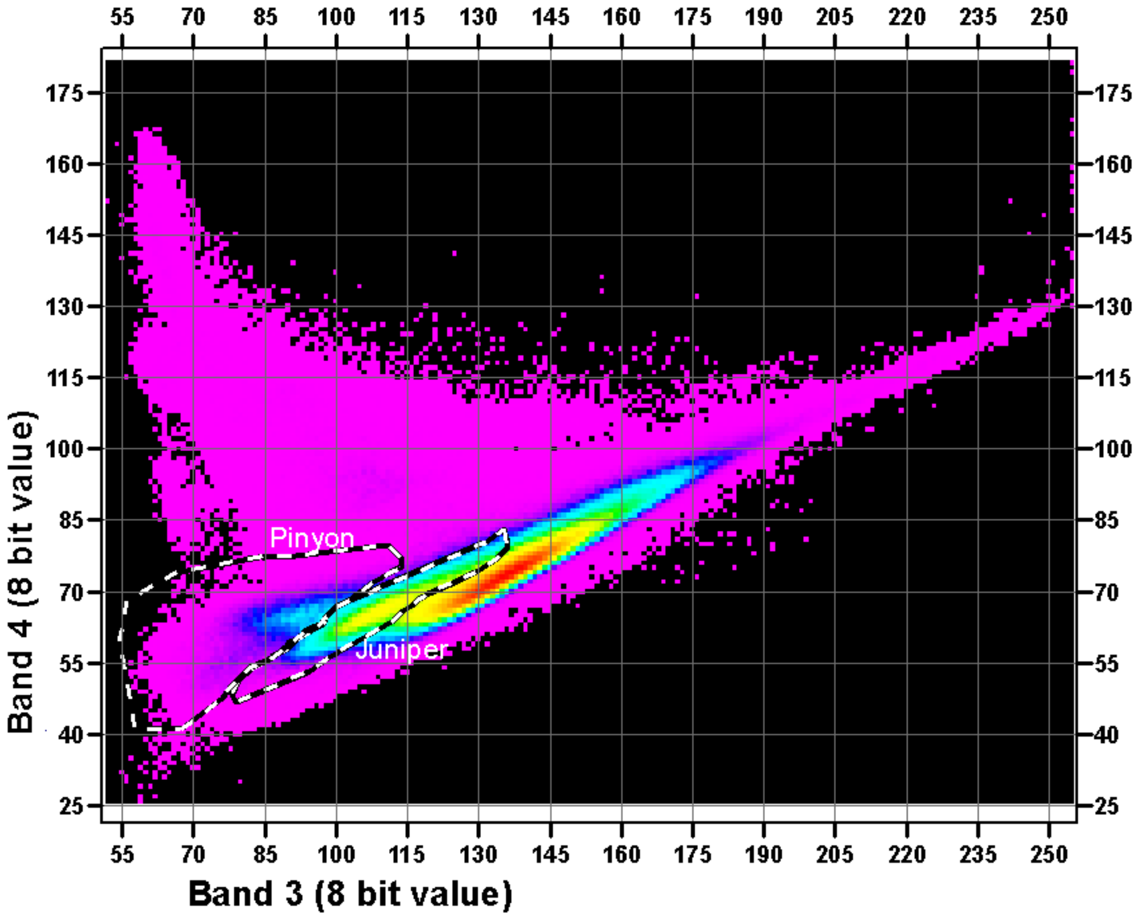


Figure 14. Feature space plot of Landsat 7 ETM+ (non-standardized) bands 3 and 4. The two areas of interest (AOI) delineate two distinct clusters of pixels, one where the dominant vegetation is pinyon and the other is juniper.

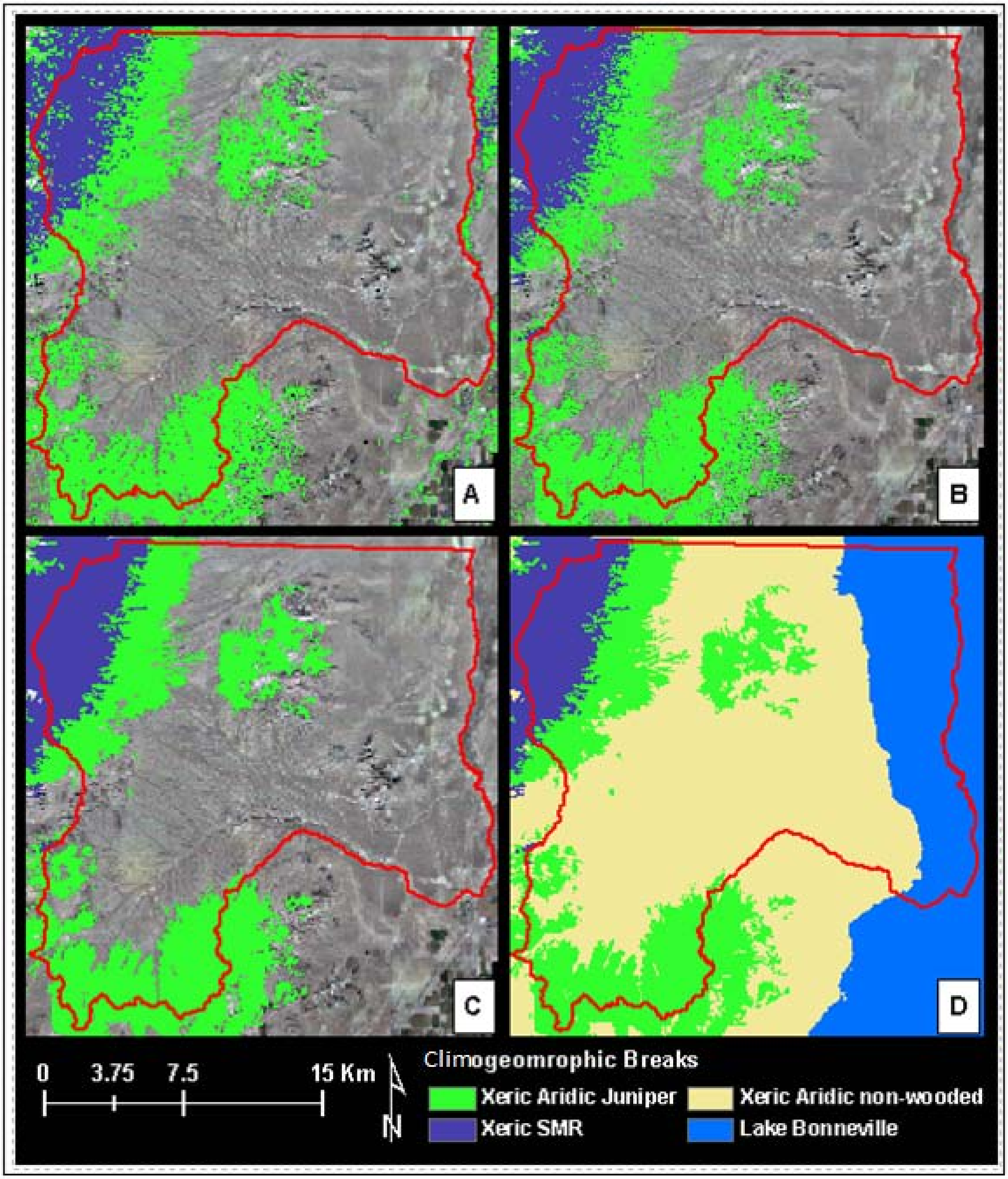


Figure 15. Climogeomorphic breaks A: The PJ classification. B: PJ classification with elevation constraints applied. C: Final SMR classification. D: The four climogeomorphic breaks incorporated as a predictive variable in random forests.

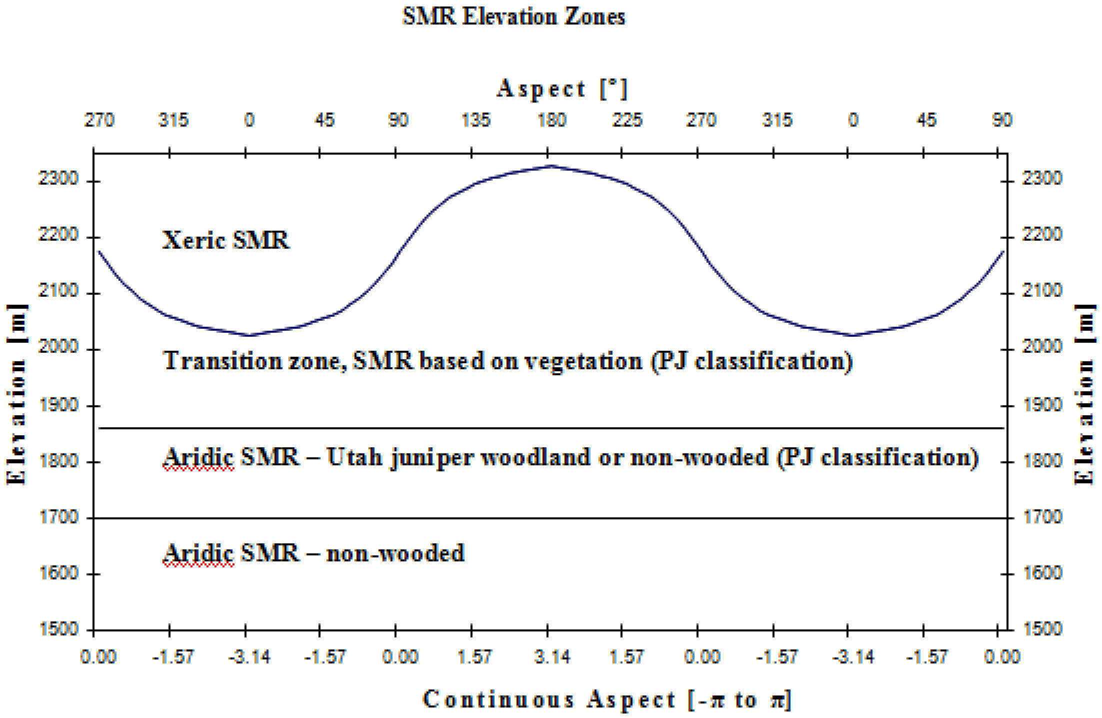


Figure 16. The elevation range relative to aspect for each SMR.

tamarisk occurred on the valley floor below 1600 m. Therefore, a conservative threshold of 1700 m was set, where all points below this elevation were classified as non-woodland and aridic SMR. As mentioned above, woodland that is dominantly pinyon is indicative of a xeric SMR though some pinyon does grow in areas with an aridic SMR.

Based on field observations, areas between 1700 m and 1860 m that were classified as pinyon or juniper in the original PJ classification were classified as woodlands with an aridic SMR. The PJ classification did not account for many areas in the xeric SMR which were not wooded, such as talus slopes, rock outcrops, and other non-wooded areas in the xeric SMR. The transition from aridic SMR to xeric SMR was observed at elevations between 1860 m and about 2325 m relative to aspect. All areas above 2325 m in elevation regardless of aspect and areas above 2025 m on north aspects are thought to be in the xeric SMR. A conservative estimate of 1860 m was set as the lowest extent of the xeric SMR. Both the transformed aspect (-π to π) and elevation rasters were incorporated into a model to refine the xeric-aridic break between the elevations of 2025 to 2325 m. I empirically fit an inverse tangent function to estimate maximum threshold elevation (m): $2175 + 119 \times \tan^{-1}(Aspect)$

where 2175 is the midpoint elevation between 2025 m and 2325 m, transformed aspect ranges from –π (North) to π (South), and 119 is a coefficient that converts $\tan^{-1}(\pm \pi)$ to the elevation range of ±150m. With the calculation of the maximum threshold elevation relative to aspect and the minimum elevation of 1860 m, an envelope of transition was created. All pixels within this envelope are determined to be xeric or aridic SMR based on the original PJ classification, where a pinyon classification is xeric SMR and juniper as aridic SMR. The conditions of this classification are defined in the simplified argument below, where the elevation zone of each SMR class is shown.

$$SMR\ Classification = \begin{cases} Xeric\ SMR \quad if\ \ Elevation \geq Maximum\ PJ\ Threshold\ \ Elevation \\ \qquad\qquad OR\ Elevation \geq 1860m\ AND\ \ pinyon\ in\ PJ\ \ classification \\ Wooded\ \ Aridic\ SMR \\ \qquad\qquad\quad if\ \ Pinyon\ OR\ \ Juniper\ in\ PJ\ \ classification \\ \qquad\qquad\quad AND \geq 1700m \\ Non-wooded\ \ Aridic\ SMR \quad Otherwise \end{cases}$$

A clump and eliminate procedure was run in ERDAS Imagine on the final SMR classification where all clumps less than 500 pixels (5 ha) were eliminated (Figure 15C). The clump procedure groups individual pixels with neighboring pixels that have the same identity into groupings called clumps, similar to creating polygons but the individual pixels still remain but are also labeled with a clump ID. The eliminate procedure identifies clumps smaller than a user-specified size (area or number of pixels) and then eliminates them. This is done by an iterative process, where individual pixels within these identified clumps are reassigned to another class with a majority filter.

The four final climogeomorphic units derived from the customized data layers are shown in Figure 15D. These are the Xeric SMR, Xeric Aridic SMR with juniper, Xeric Aridic SMR non-wooded, and the area below the Lake Bonneville shoreline which has xeric aridic and typic aridic SMRs.

Field Work

Field observations of soils, vegetation, and landscapes were gathered over two field seasons, the summers of 2005 and 2006. Soils were described from profiles exposed in small holes (<1-m diameter x 1-m deep) excavated by hand and, in a few cases, larger exposures excavated with a backhoe. Field observations included full pedon descriptions (description of soil morphology) and abbreviated soil descriptions. Full pedon descriptions included the depth, color, texture, rock fragment content, roots/pores, structure, boundary, presence of secondary carbonates/silica, clay films, pH/reaction and other unique features for each soil horizon. Also, slope (%), and aspect (compass direction in degrees) of the site, presence or absence of biological soil crust, percent of surface covered by rock fragments and the rock fragment lithology, and type of vegetation present in order of dominance. An abbreviated soil description ranged from having almost all the elements of a full pedon description to stating only the soil classification. In most cases, slope, vegetation, depths to major horizons, rock content, presence of secondary carbonates/silica, and some textures are recorded for each abbreviated description. At each observation, the soil was classified to the family level, according to the ninth edition of the *Keys to Soil Taxonomy* (Soil Survey Staff, 2003) and the UTM coordinates were recorded with a global positioning system (GPS Garmin 76).

The locations of the field observations were not generated randomly. I determined which landforms to investigate based on tacit knowledge. The sampling paradigm of summer 2005 focused on investigating soil-landscape relationships and developing conceptual models to predict those relationships. Observations along linear 10-point transects and at individual points were gathered. Once conceptual soil-landscape models were developed, sampling in summer 2006 was oriented toward discovering the geographic extent of the major soil types and refining the conceptual models. Color aerial photography and an image showing the first three principle components of Landsat data were used to help develop conceptual models and guide field sampling. Approximately 650 sampling points (~250 during 2005 and ~400 in 2006) were logged using a Garmin 76S GPS unit. Waypoints were downloaded using DNR Garmin software.

Predicting Soil Classes Using Random Forests

Soil Classes

All observations made in the field were compiled into an Excel spreadsheet where fields for the particle size family classification, diagnostic horizons and features, depths to top of calcic and bottom of argillic horizons, dominant vegetation, slope, and taxonomic classification were populated. Based on this information each observation was assigned to one of 23 soil classes. A 24th class was added to include mine dumps and other severely disturbed areas. The following is a summary of each class:

Class 1: Dixie soils were the most commonly observed soil found on stable fan remnants throughout the survey area. They formed from mixed alluvium and were vegetated with Wyoming big sage.

Class 2: Garbo soils are similar to Dixie soil except they have durinodic properties (partially cemented Bkkq horizons) and were more limited in extent. Dixie and Garbo were often found together on the landscape.

Class 3: Crestline soils were found on younger fan remnants, fan skirts and lake terraces. They formed in mostly mixed alluvium and were vegetated with Wyoming big sage. Most often they had a cambic horizon and were non-effervescent to the surface.

Class 4: Heist soils were found on inset fans and fan skirts. They formed from mixed alluvium and were vegetated with winterfat, Basin big sage, and Douglas rabbitbrush. They were often less developed than Crestline soils and often calcareous to the surface.

Class 5: Sugarloaf soils were found on stream terraces. They formed from mixed alluvium and were vegetated with rabbitbrushes and ephedra. They had weak calcic horizons and were calcareous to the surface. They included some soils that were coarse-loamy but had less than 10% clay in all horizons.

Class 6: Taylorsflat soils were found on lake shore remnants and terraces. They were lacustrine or reworked lacustrine deposits and were vegetated with Basin big sage or Wyoming big sage. They were minor in extent.

Class 7: Biblesprings soils were found on fan skirts and lake terraces. They were vegetated with Wyoming big sage and were very minor in extent.

Class 8: Hiko Peak soils were found on fan remnants and alluvial fans. They formed from mixed alluvium. Hiko Peak-like soils were found under three different vegetation types (see Classes 21 and 22). Class 8 includes Hiko Peak with Wyoming big sage and was closely associated with Crestline soils on lower fan remnants.

Class 9: Moderately deep Petrocalcids were found on fan remnants with Crestline and Hiko Peak. They were vegetated with Wyoming big sage and were rarely observed.

Class 10: Thermo Springs soils were found on valley floors below Lake Bonneville shoreline. They were vegetated with shadscale, winterfat, greasewood and budsage.

Class 11: Typic Calciargids were found on lake terraces and valley floors below Lake Bonneville shoreline. They were vegetated with shadscale, winterfat, Douglas rabbitbrush and budsage. These soils may have been sodic and or saline. They were associated with Thermo Springs.

Class 12: Uvada soils were found on valley floors below Lake Bonneville shoreline. They were vegetated with greasewood, similar to Thermo Springs, Class 10.

Class 13: Loamy-skeletal Xeric Calciargids were found on fan remnants and were vegetated with Utah juniper and Wyoming big sage. They were minor in extent.

Class 14: Pyrat soils formed on fan remnants. Some soils appeared to be residuum from weakly consolidated fanglomerate. They were vegetated with black sage and shadscale.

Class 15: Fluvents were found on inset fans and drainages. They were dominated by sandy-skeletal textures and were subject to occasional flash flooding. They were vegetated with rubber rabbitbrush, Wyoming big sage, and spiny hopsage.

Class 16: Olac soils were found on hills and foothills composed of andesite. They were vegetated by Utah juniper and black sage.

Class 17: Pibler soils were found on fan remnants and were vegetated with black sage.

Class 18: Saxby were found on hills and foothills composed igneous and some sedimentary rock. They were vegetated by black sage.

Class 19: Deep Haploxeralfs were found on the foothills and structural benches in the San Francisco Mountains. The dominant vegetation was single-leaf pinyon.

Class 20: Haploxeralfs and Haploxerepts were found on the steep mountain faces of the San Francisco Mountains. This was the broadest soil class in the legend and includes limber pine, white fur, curleaf mountain mahogany and other vegetation.

Class 21: Hiko Peak soils were found on fan remnants and alluvial fans, formed from mixed alluvium, and vegetated with black sage (see also Classes 8 and 22.

Class 22: Hiko Peak soils were found on fan remnants and alluvial fans, formed from mixed alluvium, and vegetated with Utah juniper (possibly invasive).

Class 23: Carbonatic soils were found on fan remnants, vegetated with pygmy sage, and very minor in extent.

Class 24: This class included mine dumps and other severely disturbed sites.

Four classes were composed of more than one soil type (associations and complexes), classes 15, 19, 20, and 23. These soils in these classes commonly occurred together on the same landform. They also supported similar vegetation communities.

Many of the individual soil components of these class combinations did not have sufficient sample numbers to be predicted individually. The soil components in these broader classes, and were not classified to the family level of Soil Taxonomy (Soil Survey Staff, 2003). The best examples are classes 19 and 20 which cover the remote areas of the San Francisco Mountains. The taxonomic classification for the soil classes is listed in Table 6.

Sampling of Digital Data

All field observations used to train the random forests were at least 90 m apart so that no pixel was double-sampled, which resulted in a final set of 561 field points (Figure 17). Classes 6, 15, 18, and 24 had low sample sizes but could be easily identified in the aerial photography. Polygons were digitized over these areas which were identified as these four classes and points were randomly generated within these polygons. Points that were at least 90 m apart were selected to supplement the sample of these four classes. An additional 111 points were added through this case-based reasoning approach (Shi et al., 2004) (Figure 17). The total number of sample points was 672. The numbers of observations by class are reported in Table 6.

Each observation point was attributed with the values of each environmental covariate using a sampling tool from the ArcGIS Spatial Analyst toolbox, essentially piercing through the stack of covariates. Nearest neighbor assignment was used to attribute each point. The resulting table was exported as a .txt file and read into R, a language and environment for statistical computing (R, 2007), to be formatted for importation into the Random Forests software (all R scripts are found in Appendix A).

Table 6. The general taxonomic class and soil series name (if available) of each predicted class. The number of cases used to train the models and the minimum confidence threshold for each class is shown.

Class Code	General Taxonomic Class	Soil Series & Notes	Grove 1A/B	Grove 2A/B	Confidence Threshold
1	fine-loamy, mixed, mesic Xeric Calciargids	Dixie	77	77	0.25
2	fine-loamy, mixed, mesic Durinodic Xeric Calciargids	Garbo	38	38	0.25
3	coarse-loamy, mixed, mesic Xeric Haplocalcids	Crestline	64	61[a]	0.25
4	coarse-loamy, mixed, mesic Xeric Haplocambids	Heist	19	19	0.25
5	sandy, mixed, mesic Xeric Haplocalcids	Sugarloaf	18	18	0.15
6	fine-loamy, mixed, mesic Xeric Haplocalcids	Taylorsflat	3	16[b]	0.25
7	coarse-loamy, mixed, mesic Durinodic Xeric Haplocalcids	Biblesprings	14	13[a]	0.25
8	loamy-skeletal, mixed, mesic Xeric Haplocalcids	Hiko Peak: Big Sage	68	65[a]	0.15
9	mixed, mesic Calcic Petrocalcids, moderately deep	none	10	10	0.15
10	fine-loamy, mixed, mesic Typic Natrargids	Thermosprings	5	5	0.15
11	fine-loamy, mixed, mesic Typic Calciargids	none	10	10	0.25
12	fine, mixed, mesic Typic Natrargids	Uvada	2	2	0.15
13	loamy-skeletal, mixed, mesic Xeric Calciargids	none	13	13	0.15
14	loamy-skeletal, mixed, mesic Durinodic Xeric Haplocalcids	Pyrat	18	18	0.15
15	Fluvents	none: washes	21	41[b]	0.15
16	fine-loamy, mixed, mesic Lithic Xeric Haplargids	Olac	26	26	0.15
17	loamy-skeletal, mixed, mesic shallow Calcic Petrocalcids	Pibler	20	20	0.1
18	loamy-skeletal, mixed, mesic Lithic Xeric Torriothents & Haplocalcids	none	29	62[b]	0.15
19	deep Haploxeralfs	none	12	12	0.15
20	Lithic Haploxeralfs & Haploxerepts	none	11	11	0.15
21	loamy-skeletal, mixed, mesic Xeric Haplocalcids	Hiko Peak: Black Sage	54	53[a]	0.15
22	loamy-skeletal, mixed, mesic Xeric Haplocalcids	Hiko Peak: Juniper	29	29	0.15
23	loamy-skeletal, carbonatic, Xeric Haplocalcids mesic	none: pygmy sage	NA	8[a]	0.25
24	Mine dumps	none	NA	45[b]	0.5

[a] Eight cases were reassigned to create soil class 23 in Groves 2A and 2B.

[b] Classes supplemented by CBR.

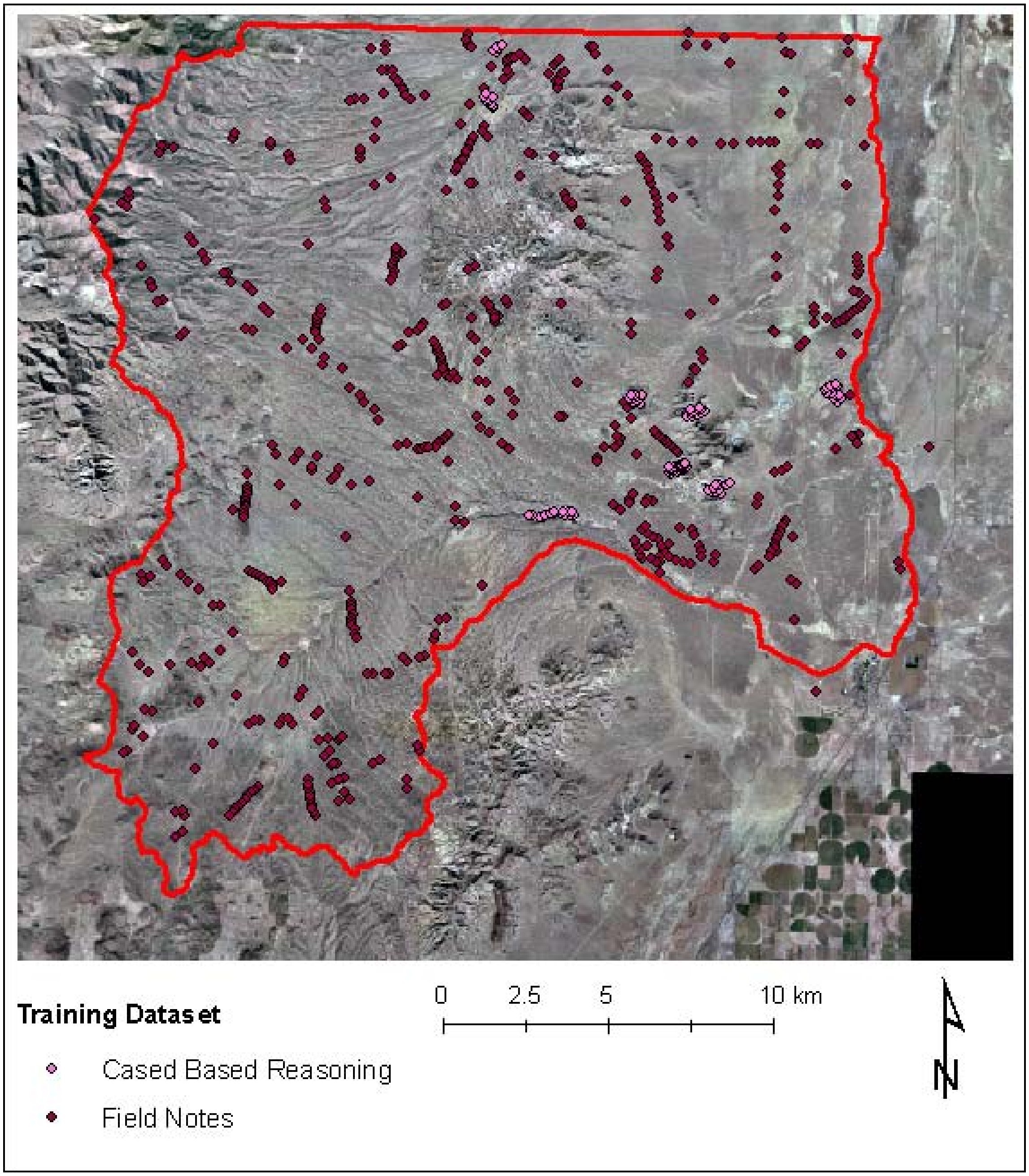


Figure 17. Training data set for the random forests models. The 561 points (maroon) were observations made in the field. The 111 observations made by case based reasoning are shown in pink.

The unknown sample, a set of unknown points (515,731 points), was generated by creating a 30-m raster layer of the study area, which was then converted to a grid of points evenly spaced 30-m apart. The method of sampling the covariates at unknown points was the same as that described above for sampling at known points.

Random Forests Model

Random Forests software by Salford Systems (2004) was used to grow the grove of trees to make the soil class predictions. In addition to predicting soil classes, RF was applied to predict the presence of diagnostic soil features, such as the presence of an argillic horizon (see Appendix B). Various model outputs were validated with “out of the bag” (OOB) testing. Each tree was trained with an independent bootstrap sample, which is a random selection of sample points with replacement. Within an individual bootstrap sample some points may be drawn one or more times while others may not be drawn. On average, one-third of cases are not selected for an individual bootstrap sample (Breiman, 2001). The points not drawn into a bootstrap sample (left out) are the “out of the bag” (OOB) samples. As these OOB points are not used to train that tree, they are used to test the tree. The OOB samples are thrown down the tree, and the tree predicts their class.

After all trees were grown, each OOB sample point was assigned a final classification, which is the majority class from each time that point was left out of the bag. The results of this are then summarized in an error matrix. The overall OOB error is the proportion of OOB misclassifications of all the sample cases. The class OOB error is the proportion of OOB misclassifications for a particular class.

For all iterations, the bootstrap sample size was equal to the total sample size and 500 trees were grown for each grove. All variables were selected as potential predictors. Several iterations were run where the number of predictive variables that were randomly selected at each node was changed (1 to 21). Several iterations were run to assess the effect of weighting. The effect of changing of these parameters was gauged by the overall and class OOB error rates. The modal result of the entire grove determined the class membership for all sample points in the study area (515,731 points), and output maps were generated.

Four groves were selected for comparison of outputs which will be referred to as Groves 1A, 1B, 2A and 2B (Table 7). Initially, Groves 1A and 1B were grown with the 561 field-gathered points and only classes 1 through 22 were predicted. To improve upon the outputs of Groves 1A and 1B, groves 2A and 2B were 1) trained with the 561 points gathered in the field and the 111 supplemental points generated using case-based reasoning, for a total of 672 points; 2) two additional classes were added, classes 23 and 24, and classes 1 through 24 were predicted; and 3) the normalized ratio of Landsat bands 3 and 1 was excluded in Groves 2A and 2B as it was the least important predictive variable for Groves 1A and 1B. Groves 1A and 2A were weighted inversely proportional to sample size, whereas no weighting was applied to groves 1B and 2B. Three predictive variables were selected at each node for groves 1A, 2A, and 2B, whereas four variables were selected at each node for grove 1B (Table 7).

Table 7. The four groves selected for comparison of outputs.

Grove	*Weighted*	*Sample*	*Predicted Classes*
1A	Yes	561 field points	1-22
1B	No	561 field points	1-22
2A	Yes	All 672	1-24
2B	No	All 672	1-24

Variable Importance

Variable importance was evaluated for the final RF groves. There are two methods in RF to determine the importance of the predictive variables, the standard measure and the Gini measure. The standard measure of variable importance replaces the true values of the variable with randomly generated (likely incorrect) values for each tree in the grove, and assesses the impact on classification (Salford Systems, 2004). If there is no impact on the error of the tree the significance of the variable decreases. Conversely, if the tree's ability to predict the OOB observations is diminished, the variable is considered important. The Gini importance ranks variables according to how cleanly the variable separated classes when selected at a node (Salford Systems, 2004).

Components and Map Units

The likelihood that an individual pixel belongs to a soil class is based on the individual predictions (votes) of each independently grown tree within the grove. Salford Systems refers to this ratio of votes for a given class as a "probability" (Salford Systems, 2004). In reference to the OOB performance of the training data set, Breiman (2001) says, "The margin measures the extent to which the average number of votes…for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification (p. 7).” Peters et al. (2007) interpret the proportion of trees as a probability. Williams and Abernethy (2008) refer to it as the prediction confidence and suggest that it could be used in fuzzy logic algorithms. In this thesis I will refer to the proportion of trees (votes) that predict a given class as the likelihood or confidence in the prediction. As mentioned above the final classification is the mode of all the trees. The proportion of trees that voted or predicted the modal class is known, as well as the proportion of votes for all other classes. This gives the likelihood of membership for each class in the legend. These ratios were used to estimate uncertainty, determine the extent of individual components, and create new soil map units (associations).

Minimum confidence thresholds were established in determining the second and third most likely classifications of each pixel. Three threshold values were established based on potential limitations to use and management (Table 6), where the most limiting soils (e.g., Petrocalcids) had the lowest minimum threshold (0.1), while soils with some limitations, such as >35% rock fragments, had a minimum threshold of 0.15 and all other soil classes had a minimum threshold of 0.25. These threshold values were based on the minimum composition of a soil component to be considered a major component in a map unit by the NRCS (Table 1 in USDA-NRCS, 2009). However, it should be noted that these confidence values should not be interpreted as composition of a cell or even the composition of an aggregation of cells. A threshold value of 0.50 was established for miscellaneous class 24 as a pixel is either a mine dump or not.

In the Knowledge Engineer in Imagine, a simple argument (rule) was created where all pixels that had a relatively high likelihood (≥ the minimum confidence threshold [Table 6]) of belonging to class α were identified (Figure 18). The purpose in identifying these alternate classifications of each pixel is to determine a measure of proximity between pixels. For example, if pixel *i* is predicted to be class α with a likelihood of 0.60 and pixel *y* is predicted to be class β with a likelihood of 0.55, the two pixels are said to be in different classes. However, if pixel *i* also has a 0.31 likelihood of belonging to class β and pixel *y* has a 0.37 likelihood of belonging to class α, pixels *i* and *y* are arguably similar in some respect.

Similar pixels can be indentified and aggregated (clumped) to represent map unit associations and complexes by extending the above argument. All pixels that have a relatively high likelihood (≥ minimum confidence threshold) of belonging to both classes α AND β can be grouped together in a map unit.

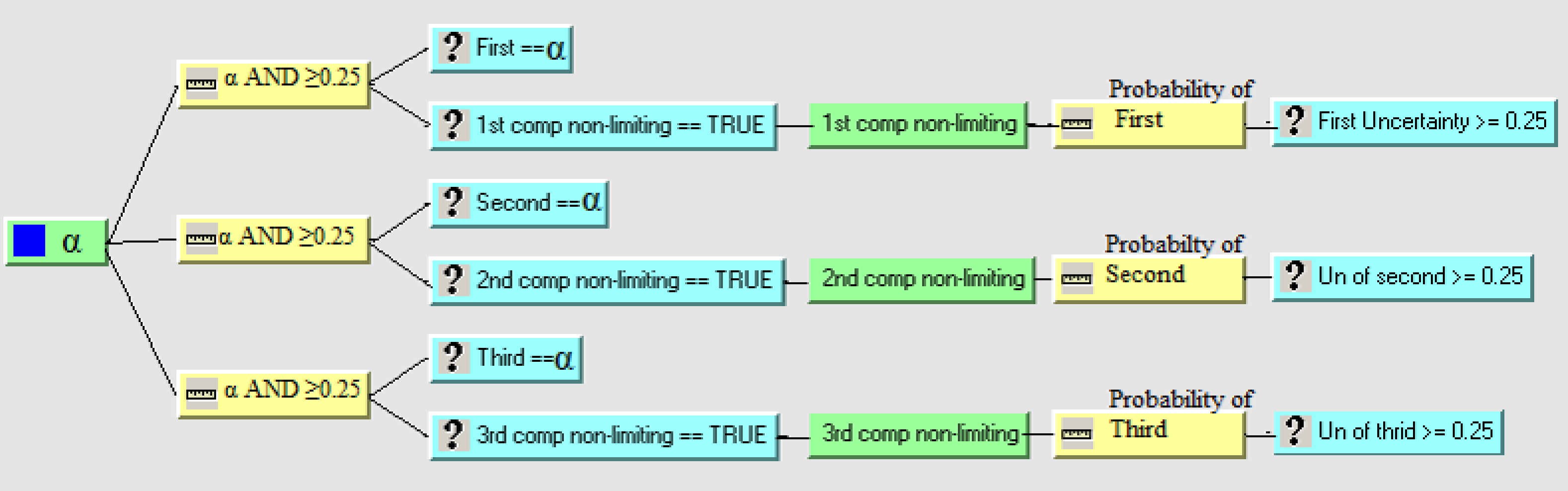


Figure 18. Conceptual example of an Imagine .ckb model. Green boxes represent hypotheses, yellow boxes represent the rules, and blue boxes represent conditions. For hypothesis α to be true one of the three conditions must be met.

RESULTS AND DISCUSSION

Groves 1A and 1B

Weighting increased the overall OOB error rate while reducing the OOB error rate for individual classes with smaller sample sizes (Table 8). For example, Grove 1A, which was weighted inversely proportional to sample size, had an overall OOB error of 74.7%, whereas the error for Grove 1B, which was not weighted, was 61.1%. The OOB error for class 1, which had the largest sample size (n=77), was 57.1% in Grove 1B, but increased to 97.4% in the weighted Grove 1A. In contrast, the OOB error for class 15 (n=21) decreased from 90.5% in Grove 1B to 61.9% in weighted Grove 1A (Table 8).

The differences in error by class are apparent when the outputs of Groves 1A and 1B (Figures 19A and 19B, respectively) are compared (Figure 19C). Only 36.6% of the pixels were predicted to be in the same class. This is illustrated by class 1, which are Dixie soils frequently observed on fan remnants. Class 1 was rarely predicted with the weighted Grove 1A; fan remnants, where Dixie was observed in the field, were often classified as Biblesprings (class 7) and Garbo (class 2). Biblesprings, which was rarely observed in the field, was overrepresented by weighting. Garbo and Dixie are fairly similar soils and did often co-occur on older fan remnants. Another example is Hiko Peak with black sage (class 21) and Pyrat (class 14), which also has black sage. Based on field observations, Hiko Peak was the most common soil found on steeper fan remnants, whereas Pyrat was only found on a couple of highly dissected fan remnants. Again, the weighted Grove 1A overrepresented the distribution of the less common class (Pyrat). Groves 1A and 1B had general agreement on soils in the northeast corner of the study

Table 8. Summary of the OOB error matrix for Groves 1A and 1B showing the number of cases from each class (*Number of Cases*), the number of cases that were misclassified (*Number Misclassified*) when left out of the bag, and the percent OOB error (*Percent Error*) for the class.

Grove 1A

Class	**Number of Cases**	**Number Mis-classified**	**Percent Error**
1	77	75	97.4
2	38	11	29.0
3	64	62	96.9
4	19	12	63.2
5	18	6	33.3
6	3	3	100.0
7	14	10	71.4
8	68	56	82.4
9	10	8	80.0
10	5	1	20.0
11	10	3	30.0
12	2	1	50.0
13	13	13	100.0
14	18	12	66.7
15	21	13	61.9
16	26	18	69.2
17	20	12	60.0
18	29	20	69.0
19	12	7	58.3
20	11	4	36.4
21	54	49	90.7
22	29	23	79.3
Overall Error			**74.7**

Grove 1B

Class	**Number of Cases**	**Number Mis-classified**	**Percent Error**
1	77	44	57.1
2	38	21	55.3
3	64	42	65.6
4	19	12	63.2
5	18	1	5.6
6	3	3	100.0
7	14	14	100.0
8	68	26	38.2
9	10	10	100.0
10	5	4	80.0
11	10	5	50.0
12	2	2	100.0
13	13	13	100.0
14	18	12	66.7
15	21	19	90.5
16	26	17	65.4
17	20	16	80.0
18	29	16	55.2
19	12	7	58.3
20	11	5	45.5
21	54	41	75.9
22	29	13	44.8
Overall Error			**61.1**

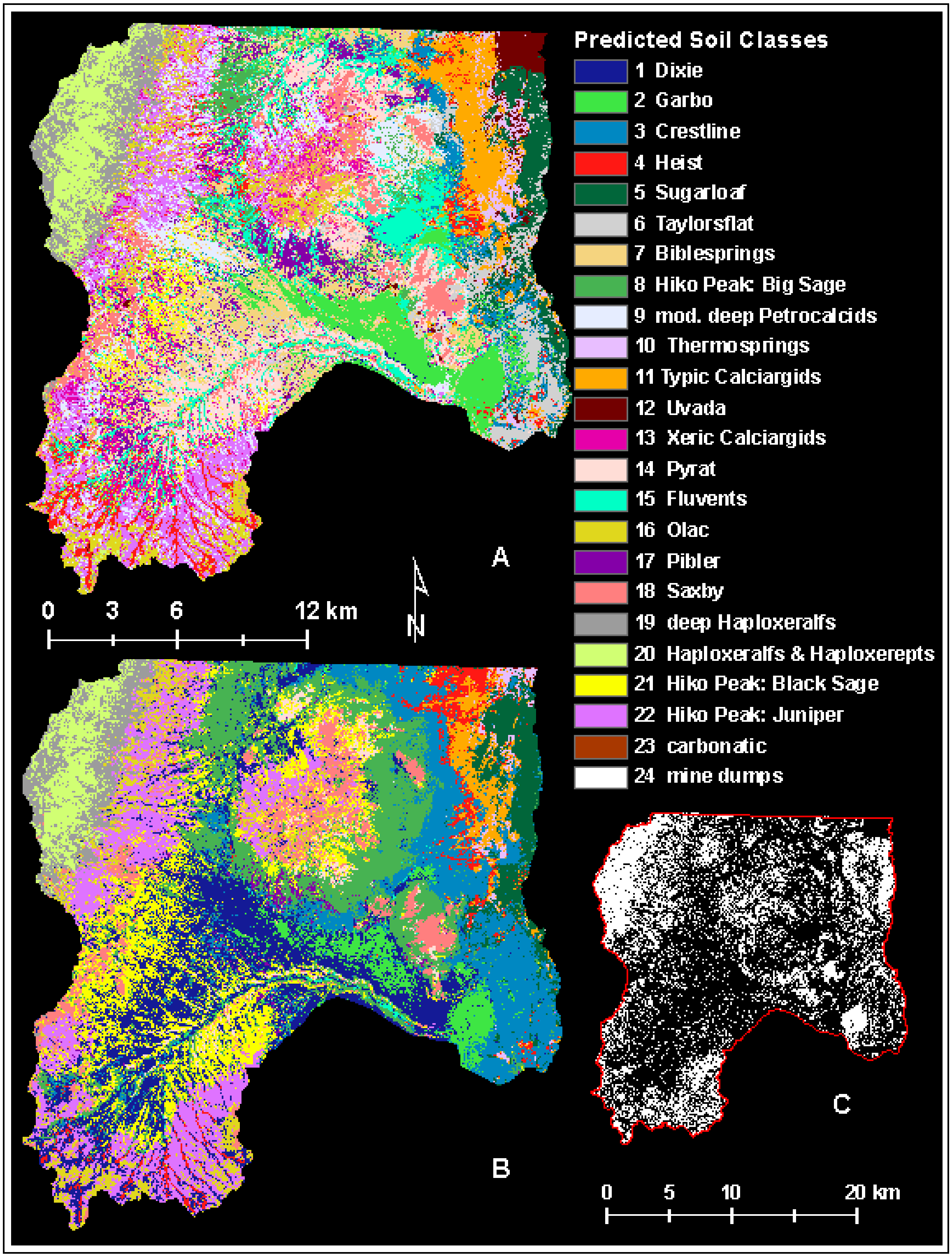


Figure 19. The results of Groves 1A (A) and 1B (B). C: The white pixels indicate where Groves 1A and 1B were in agreement.

area with classes 4, 5, 10, and 11, which were observed on the valley floor. The distribution of classes predicted in the San Francisco Mountains (far west side of the survey) showed strong agreement between Groves 1A and 1B. While the unweighted Grove 1B had a much lower overall error and best represented the soils as observed in the field, several classes that I was familiar with in the field were poorly predicted, such as Fluvents (class 15) as only a small number of observations were made of these classes in the field.

Groves 2A and 2B

Grove 2A, which was weighted inversely proportional to sample size, had an overall OOB error of 64.9%, whereas the error for Grove 2B, which was not weighted, was 55.2% (Table 9). The grove with weighting (2A) overrepresented minority classes and neglected the same large classes as Grove 1A; for example, the OOB error for class 1 in Grove 2A was 100% (Table 9). When Grove 2A (Figure 20A) was compared to Grove 2B (Figure 20B) only 37.6% of pixels were predicted to be the same class (Figure 20C), which is very similar to the comparison of Groves 1A and 1B (Figure 19C).

The addition of 111 points by case-based reasoning (CBR) and the addition of two classes decreased overall error. The OOB errors for Groves 2A and 2B were notably lower than for Groves 1A and 1B, respectively (Tables 8 and 9). The differences between Groves 1B and 2B were largely a result of the better predictions of classes 6, 15, 18, and the addition of two new classes 23 and 24. For example, the OOB error for class 6 decreased from 100% to 43.8%, class 15 from 90.5% to 56.1%, and class 18 from 55.2%

Table 9. Summary of the OOB error matrix for Groves 2A and 2B, showing the number of cases from each class (*Number of Cases*), the number of cases that were misclassified (*Number Misclassified*) when left out of the bag, and the percent OOB error (*Percent Error*) for the class.

Grove 2A

Class	Number of Cases	Number Misclassified	Percent Error
1	77	77	100.0
2	38	12	31.6
3	61	59	96.7
4	19	9	47.4
5	18	4	22.2
6	16	8	50.0
7	13	9	69.2
8	65	47	72.3
9	10	9	90.0
10	5	3	60.0
11	10	3	30.0
12	2	1	50.0
13	13	12	92.3
14	18	12	66.7
15	41	22	53.7
16	26	17	65.4
17	20	12	60.0
18	62	31	50.0
19	12	6	50.0
20	11	3	27.3
21	53	52	98.1
22	29	22	75.9
23	8	0	0.0
24	45	6	13.3
Overall Error			**64.9**

Grove 2B

Class	Number of Cases	Number Misclassified	Percent Error
1	77	42	54.6
2	38	24	63.2
3	61	41	67.2
4	19	12	63.2
5	18	5	27.8
6	16	7	43.8
7	13	13	100.0
8	65	28	43.1
9	10	10	100.0
10	5	4	80.0
11	10	5	50.0
12	2	2	100.0
13	13	13	100.0
14	18	13	72.2
15	41	23	56.1
16	26	17	65.4
17	20	16	80.0
18	62	15	24.2
19	12	6	50.0
20	11	5	45.5
21	53	45	84.9
22	29	15	51.7
23	8	6	75.0
24	45	4	8.9
Overall Error			**55.2**

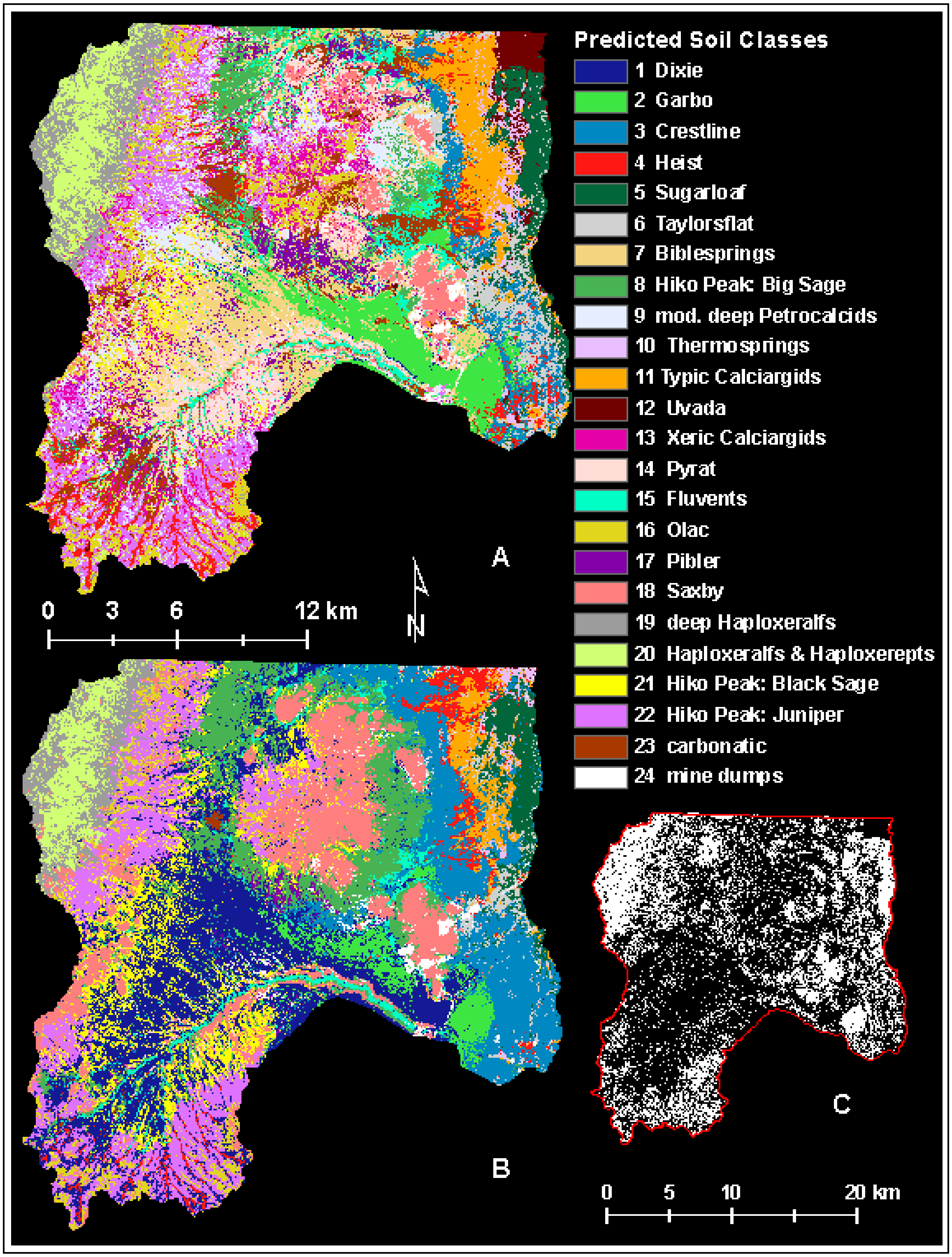


Figure 20. The results of Groves 2A (A) and 2B (B). C: The white pixels indicate where Groves 2A and 2B were in agreement.

to 24.2% (Tables 8 and 9). When Groves 1B and 2B were compared, 77.2% of pixels were predicted to be the same (Figure 21C). While the sample size of each class was not perfectly proportional to the spatial extent of each class in the study area, it was indicative of their relative extent. The dominant landform in the study was fan remnants, and Dixie (class 1; 77 observations), Hiko Peak (classes 8 and 21; 65 and 53 observations, respectively), and Crestline (class 3; 61 observations) were the commonly occurring soils on fan remnants. Therefore, when Grove 2A predicted 9.1% of all pixels to be Biblesprings (class 7) with only 13 observations, and only 0.12% of pixels were predicted to be Dixie (class 1) with 77 observations, it raised a red flag.

Individual classes can be weighted differentially to improve the prediction of these specific classes. While different weights could have been applied to those poorly predicted classes in Groves 1A and 1B, I felt that manipulating the weights on a class by class basis introduced bias into the model. If one desired lower prediction errors for specific classes, perhaps to facilitate the prediction of rare plant habitat, individual classes could be weighted.

The purpose of a third order soil map is to characterize the general distribution of soils on landforms, not to overpredict less common soils at the cost of the dominant soils. In the outputs from Groves 1B and 2B, landforms can be clearly observed by the pattern of the predicted soil classes. For example, both the Beaver Lake Mountains and Rocky Range are flanked by steep fan remnants. In Groves 1B and 2B, these fans are clearly visible as green aprons (class 8, Hiko Peak: Wyoming big sage). Groves 1A and 2A outputs look messy in comparison, making it difficult to identify such soil-landscape relationships. Based on my

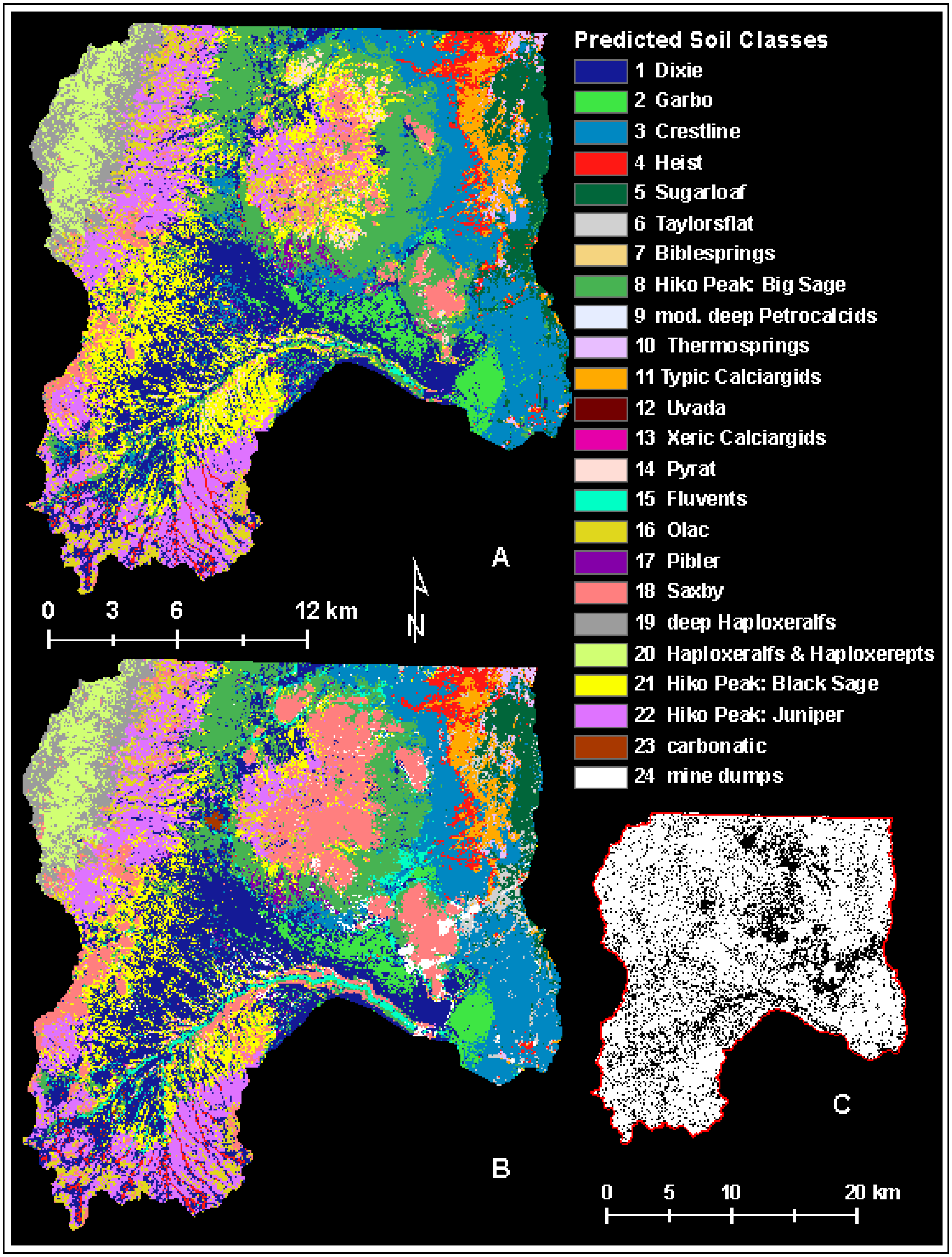


Figure 21. A comparison of Groves 1B (A) and 2B (B). C: The white pixels indicate where Groves 1B and 2B were in agreement.

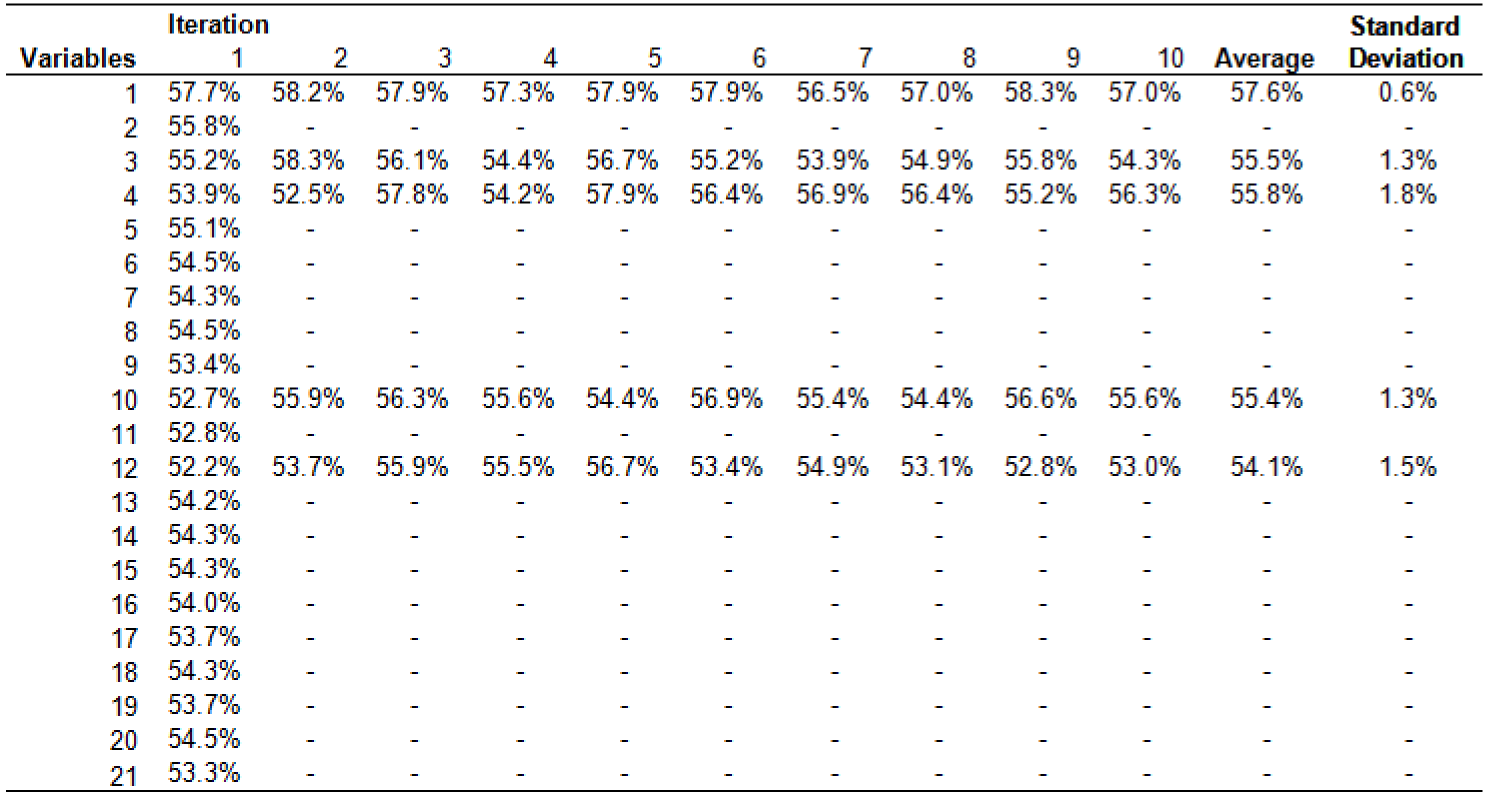
field experience in the area, Groves 1A and 2A did not represent the soil distribution across the study area as well as Grove 2B or even Grove 1B.

As RF is doubly random, two groves grown with the same parameters may be unique if the random number seed is changed. While groves with the same parameters may be different, growing 500 trees does make the model quite stable. To quantify this variability, 21 groves were grown using the same parameters as Grove 2B, except changing the number of variables at each node. Nine additional iterations were done for groves having one, three, four, 10 and 12 variable subsets at each node (45 groves) to determine average OOB errors and standard deviations. A unique random number seed was used for each of these 45 additional groves. The overall OOB error of these 66 groves ranged from 52.2% to 58.3% (Table 10). The standard deviations for ten iterations ranged from 0.6% to 1.8%. There was a 4.4 percentage point range, 53.9% to 58.3%, in the OOB error from the 10 groves that had the exact same parameters as Grove 2B (3 variables). This demonstrates that how much of the OOB error between these groves is random. Weighting had a much greater impact on the OOB error rate than changing the number of independent variables that were subset at each node.

Grove 2B

Based on the OOB error and the observations made previously about the soil-landscape relationships, Grove 2B was selected as the best model output, on which the remainder of the results and discussion are based. The spectral and topographic characteristics of each class (observations and predictions) were further explored in

Table 10. A comparison of OOB error rates for 66 groves grown using the same parameters as Grove 2B, except for varying the number of variables at each node (21 groves) and the number of iterations of selected models (45 groves).



	Iteration											Standard
Variables	1	2	3	4	5	6	7	8	9	10	Average	Deviation
1	57.7%	58.2%	57.9%	57.3%	57.9%	57.9%	56.5%	57.0%	58.3%	57.0%	57.6%	0.6%
2	55.8%	-	-	-	-	-	-	-	-	-	-	-
3	55.2%	58.3%	56.1%	54.4%	56.7%	55.2%	53.9%	54.9%	55.8%	54.3%	55.5%	1.3%
4	53.9%	52.5%	57.8%	54.2%	57.9%	56.4%	56.9%	56.4%	55.2%	56.3%	55.8%	1.8%
5	55.1%	-	-	-	-	-	-	-	-	-	-	-
6	54.5%	-	-	-	-	-	-	-	-	-	-	-
7	54.3%	-	-	-	-	-	-	-	-	-	-	-
8	54.5%	-	-	-	-	-	-	-	-	-	-	-
9	53.4%	-	-	-	-	-	-	-	-	-	-	-
10	52.7%	55.9%	56.3%	55.6%	54.4%	56.9%	55.4%	54.4%	56.6%	55.6%	55.4%	1.3%
11	52.8%	-	-	-	-	-	-	-	-	-	-	-
12	52.2%	53.7%	55.9%	55.5%	56.7%	53.4%	54.9%	53.1%	52.8%	53.0%	54.1%	1.5%
13	54.2%	-	-	-	-	-	-	-	-	-	-	-
14	54.3%	-	-	-	-	-	-	-	-	-	-	-
15	54.3%	-	-	-	-	-	-	-	-	-	-	-
16	54.0%	-	-	-	-	-	-	-	-	-	-	-
17	53.7%	-	-	-	-	-	-	-	-	-	-	-
18	54.3%	-	-	-	-	-	-	-	-	-	-	-
19	53.7%	-	-	-	-	-	-	-	-	-	-	-
20	54.5%	-	-	-	-	-	-	-	-	-	-	-
21	53.3%	-	-	-	-	-	-	-	-	-	-	-

Appendix C. Aerial photography (1-m resolution) was used to enhance the 30-m resolution of the Grove 2B output, the results of which can be found in Appendix D.

Variable Importance

The 21 environmental covariates used in Grove 2B are shown in order of importance in Table 11. Many of the variables were gauged similarly by the standard measure and the Gini method of calculating variable importance.

Six of the seven DEM-derived data variables were among the top ten most important variables, indicating the importance of topography in predicting soil classes. However, transformed aspect ("Aspect") was the least important variable when estimated using the standard measure and second least important variable when estimated using the Gini method. Therefore, aspect appeared to have little impact on soil formation in the relatively low relief landscapes (alluvial fans, low hills, and valley bottom) that comprised the majority of the study area and had the greatest numbers of field and case-based observations.

Two normalized band ratios derived from Landsat data were among the top ten most important variables for both standard and Gini measures: (4-3)/(4+3) [NDVI] and (5-2)/(5+2) [distinguished igneous from sedimentary rock sources]. Before running the model I felt that these two variables were the most significant band ratios in terms of predicting soils across the study area.

The customized discrete variables (Xeric SMR and Bonneville [Lake Bonneville shoreline]) both scored relatively low. These two variables helped separate mountains (Xeric SMR) and valley floors (Lake Bonneville shoreline) from the mid-elevation alluvial fans and low hills that constituted the majority of the study area (Appendix C). Mountains and valley floors had much fewer observations, which may have contributed to the lower scores of these variables, especially with the Gini measure.

To test the importance of both spectral and topographic variables on soil class prediction, one new grove was grown where only the DEM-derived covariates were used and another new grove was grown where only the Landsat-derived covariates were used as predictive variables. Neither grove used the two customized variables, while all other parameters were the same as grove 2B. The OOB error for the DEM-only grove was 58.9% and 69.1 % for the Landsat-only grove, confirming the greater importance of DEM-derived environmental covariates for predicting soils classes in the study area. These results would seem to be consistent with the literature, where topographic data is the most commonly used covariate (McBratney et al., 2003). However, the topographic and spectral variables both contributed to the final model performance, and variables derived from both sources scored well with the standard and Gini measures of variable importance.

Table 11. Variable importance from Random Forests Grove 2B. Variables are in order of most important to least important.

	Standard			Gini		
Order of Importance	Variable	Score	Relative Importance	Variable	Score	Relative Importance
1	Elevation	9.66	100	(4-3)/(4+3)	9.38	100
2	Slope 11x11	8.7	90.03	Slope	8.56	91.29
3	Slope	8.57	88.68	Slope 11x11	7.84	83.56
4	band 1	7.2	74.55	Elevation	7.68	81.85
5	Curvature	6.94	71.80	band 1	7.01	74.71
6	band 2	5.9	61.09	(5-2)/(5+2)	6.57	70.01
7	CTI	5.67	58.62	CTI	6.24	66.53
8	(4-3)/(4+3)	5.61	58.07	Curvature	6.05	64.50
9	(5-2)/(5+2)	4.93	50.97	CTI 5X5	5.76	61.40
10	CTI 5X5	4.5	46.55	band 2	5.68	60.58
11	(5-1)/(5+1)	4.48	46.35	(5-1)/(5+1)	4.57	48.75
12	band 3	4.35	45.02	band 6	3.23	34.43
13	(4-7)/(4+7)	3.99	41.24	band 3	3.15	33.58
14	(4-5)/(4+5)	3.45	35.71	(4-7)/(4+7)	3.12	33.28
15	band 6	3.26	33.74	Bonneville	3.01	32.09
16	(3-7)/(3+7)	2.95	30.54	(3-7)/(3+7)	2.63	28.06
17	band 5	2.82	29.19	(4-5)/(4+5)	2.53	26.96
18	band 4	2.47	25.51	band 5	2.37	25.25
19	Bonneville	2.34	24.18	Xeric SMR	1.72	18.31
20	Aspect	1.48	15.33	band 4	1.64	17.52
21	Xeric SMR	0.73	7.51	Aspect	1.27	13.58

OOB

Now that a best model has been selected attention can be turned to the performance of individual classes. The error matrix (Table 12) is a summary of the OOB performance of each observation, which provides insight on how observations were misclassified. For example, there were 65 observations of class 8 (Hiko Peak with Wyoming big sage vegetation) and 37 of them (56.9%) were classified correctly as OOB samples. By looking across the row, we can see that observations of class 8 were most commonly misclassified as class 1 (Dixie, 10 or 15%) and class 3 (Crestline, 8 or 12%). All three of these soils are found on fan remnants with Wyoming big sage vegetation communities. Looking down column 8 we can see that classes 1 and 3 were commonly misclassified as class 8. Classes 15 (Fluvents) and 21 (Hiko Peak with black sagebrush) were also commonly misclassified as class 8. Class 21 and class 8 are both Hiko Peak soils but with different sagebrush communities, black sagebrush versus Wyoming big sage. I observed in the field that some Hiko Peak soil sites were a mosaic of both shrub types.

At the bottom of each column of Table 12 is the *Percent False Positive*, which is the percent of observations that were incorrectly classified as that class. The best overall class was class 24, which had the highest *Percent Correct* (91.1%) and lowest *Percent False Positive* (2.4%) (excluding classes 9 and 12 which had zero predictions). Class 24 is the mine dumps class and is spectrally unique based on Landsat 7 reflectance (Appendix C). Class 8 had 65% false positive, which follows from the above discussion of other classes (e.g., 1 and 3) being classified as class 8.

Table 12. The error matrix summarizes the OOB results by class. The *Predicted As Class* columns shows how the training cases were predicted when left out of the bag. The bolded numbers on the diagonal is the number of correctly classified cases. *Times Class Predicted* row shows the total number cases that were predicted as for a class. *Percent False Positive* is the proportion of cases that were incorrectly classified as that class.

Actual Class	Total Cases	Percent Correct	Predicted As Class																							
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	77	45.5	**35**	6	2	2	0	0	0	11	0	0	0	0	1	1	1	2	1	3	0	0	5	6	1	0
2	38	36.8	16	**14**	1	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
3	61	32.8	16	2	**20**	0	0	0	0	15	0	0	0	0	0	0	2	0	0	1	0	0	3	2	0	0
4	19	36.8	4	0	1	**7**	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0
5	18	72.2	0	0	1	0	**13**	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	16	56.3	0	0	2	1	4	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	13	0.0	5	3	0	0	0	0	**0**	2	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0
8	65	56.9	10	0	8	1	2	0	0	**37**	0	0	0	0	0	0	2	0	0	4	0	0	1	0	0	0
9	10	0.0	1	0	0	0	0	0	0	2	**0**	0	0	0	0	0	0	0	0	2	0	0	2	3	0	0
10	5	20.0	0	0	1	1	0	1	0	0	0	**1**	1	0	0	0	0	0	0	0	0	0	0	0	0	0
11	10	50.0	0	0	0	3	1	0	0	0	0	1	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0
12	2	0.0	0	0	0	0	1	1	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
13	13	0.0	2	0	0	0	0	0	0	1	0	0	0	0	**0**	3	0	0	0	1	0	0	1	5	0	0
14	18	27.8	2	0	0	0	0	0	0	3	0	0	0	0	0	**5**	1	0	0	3	0	0	2	1	1	0
15	41	43.9	8	0	2	0	0	0	0	10	0	0	0	0	0	0	**18**	0	0	2	0	0	1	0	0	0
16	26	34.6	4	0	0	0	0	0	0	2	0	0	0	0	0	0	0	**9**	0	4	0	0	5	2	0	0
17	20	20.0	4	0	0	1	0	0	0	6	0	0	0	0	0	0	0	0	**4**	1	0	0	1	2	0	1
18	62	75.8	2	0	0	0	0	0	0	8	0	0	0	0	0	0	0	2	0	**47**	0	0	0	3	0	0
19	12	50.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**6**	4	0	1	0	0
20	11	54.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	**6**	0	0	0	0
21	53	15.1	14	0	5	0	0	0	0	12	0	0	0	0	0	0	3	1	1	6	0	0	**8**	3	0	0
22	29	48.3	4	0	0	1	0	0	0	0	0	0	0	0	1	0	0	3	0	2	0	0	4	**14**	0	0
23	8	25.0	3	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**2**	0
24	45	91.1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	**41**
Times class predicted			131	25	44	17	21	18	1	114	0	2	9	0	2	10	28	18	8	79	11	10	34	44	4	42
Percent False Positive			73.3	44.0	54.5	58.8	38.1	50.0	100	67.5	0.0	50.0	44.4	0.0	100	50.0	35.7	50.0	50.0	40.5	45.5	40.0	76.5	68.2	50.0	2.4

To test whether additional data may improve the OOB error, new groves with the same parameters as 2B were grown by withholding random subsets of the observation data. Results indicate that more data may not greatly improve the OOB error (Figure 22). By extending the trend line in Figure 22 to 120%, or 806 observations, the OOB error is likely to decrease to only 53.7%. However, more observations representing smaller classes, such as class 10, may improve the prediction of individual classes. See Appendix E for further analysis of the OOB error and how individual observations were classified OOB.

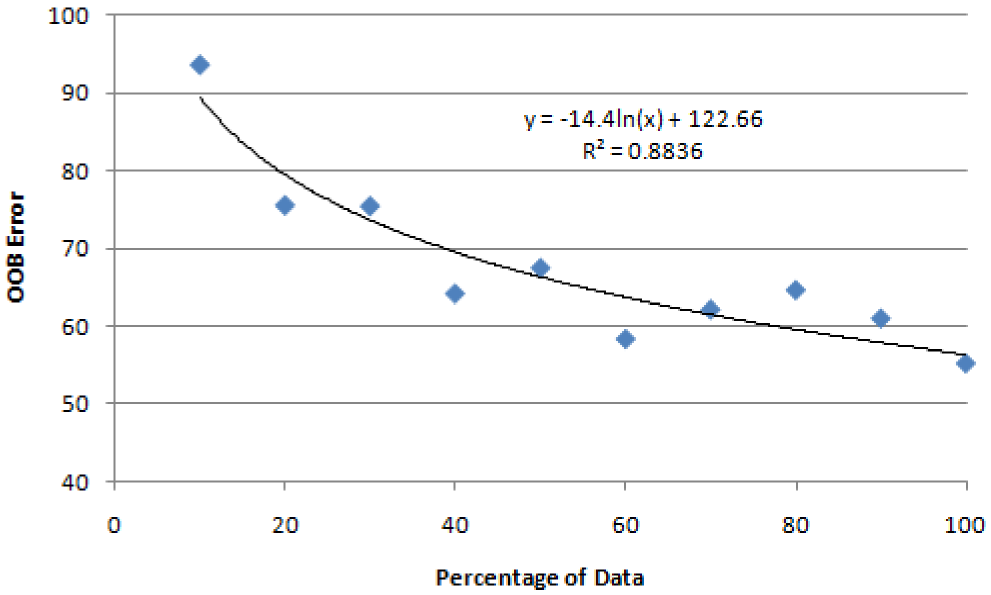


Figure 22. The OOB error of groves grown with random subsets of observation data removed from the whole dataset.

Model Confidence

In addition to the primary model output, the likelihood that a pixel may belong to another class can be gleaned from RF. For example, Table 13 shows the results of an individual pixel located in the San Francisco Mountains. The “PRED” row shows the predicted class for this pixel, based on the result of a majority vote by 500 trees, to be class 19. Row “PROB_0019” shows a prediction confidence of 0.428 that this pixel was class 19, meaning that 214 of the 500 independently grown trees predicted this pixel to be class 19 (recall that Salford Systems [2004] referred to the model confidence as a “probability”). The second most likely class was class 20 with a 0.238 prediction confidence. Class 19 consists of deep Haploxeralfs and class 20 includes Haploxeralfs or Haploxerepts that are shallow or moderately deep to a lithic contact. These are classes with similar soils that are found in the San Francisco Mountains. The third most likely class was class 16 (Olac) with a prediction confidence of 0.126. Olac soils were found on ridges and mountains and are shallow to a lithic contact. So, the three most likely classes share important similarities with regards to the soil morphology and landform.

The primary, secondary, and tertiary predictions of each pixel and the confidence of these outputs are shown in Figures 23 and 24, respectively. Some distinct geomorphic surfaces and landform patterns can be discerned from the confidence images. This is evident in the northeastern edge of the study area, where class 5 (Sugarloaf soils) is the primary prediction and class 6 (Taylorsflat soils) is the secondary prediction (Figures 23A and 23B). Both of these soils are Xeric Haplocalcids that occur below the Lake Bonneville shoreline, but have different family particle-size classes (Sugarloaf is sandy, whereas Taylorsflat is fine-loamy). Clearly the Sugarloaf soils of class 5 have a relatively

Table 13. The Random Forests result from Grove 2B for an individual pixel in the San Francisco Mountains. The final class was predicted to be class 19, which received 42.8% of the votes from the grove. The proportion of votes received for each class in the legend is also shown.

Column Heading	**Value**
MU	NA
PRED	**19**
PROB_0001	0.034
PROB_0002	0
PROB_0003	0
PROB_0004	0.002
PROB_0005	0
PROB_0006	0
PROB_0007	0
PROB_0008	0.002
PROB_0009	0.002
PROB_0010	0
PROB_0011	0
PROB_0012	0
PROB_0013	0.018
PROB_0014	0
PROB_0015	0
PROB_0016	0.126
PROB_0017	0.002
PROB_0018	0.056
PROB_0019	**0.428**
PROB_0020	0.238
PROB_0021	0.048
PROB_0022	0.044
PROB_0023	0
PROB_0024	0

high likelihood of prediction compared to the Taylorsflat soils of class 6 (Figures 24A and 24B).

Other geomorphic surfaces are a composite of two or three classes as primary predictions, and those same classes occur together as secondary predictions (Figures 23A and 23B). For example, the south-central part of the study area is composed of class 2 (Garbo soils) dominant over class 1 (Dixie soils) in the primary prediction, and composed of class 1 over class 2 in the secondary prediction (Figures 23A and 23B). This geomorphic surface is a large, low-slope fan remnant; Dixie soils are fine-loamy Xeric Calciargids and Garbo soils are fine-loamy Durinodic Xeric Calciargids, which are similar except that Garbo soils have patchy silica-cementation. The confidence of both primary and secondary predictions is relatively moderate (Figures 24A and 24B).

Some geomorphic surfaces have less distinct patterns (speckled appearance) in the primary and secondary predictions (Figures 23A and 23B). Coincidentally these same surfaces have relatively lower confidence in both primary and secondary predictions (Figures 24A and 24 B). An example is classes 1 (Dixie soils) and 21 (Hiko Peak soils with black sagebrush vegetation) in the southwestern part of the study area, which is a transition area between Dixie (finer textured soil found on lower stable surfaces) and Hiko Peak (coarser textured soil found on steeper, more active surfaces).

In the tertiary output (Figure 23C) some landform patterns still exist but the distribution is less clustered and not as discernible. The confidence of the tertiary predictions is generally quite low (Figure 24C).

Many pixels had a low confidence of belonging to any class (e.g., Figure 24A). For example, 7.7% of the pixels had ≤0.20 confidence of belonging to any class in the legend. An additional 17.3% of the pixels had 0.20-0.25 confidence of belonging to any class in the legend. Model uncertainty may be caused by: 1) none of the predicted soil classes in the legend represented the soils in the pixel; 2) the pixel represents a transition soil between several other classes; 3) several soil classes may exist in an individual 30-m pixel; 4) there were insufficient predictive variables to distinguish spectrally and topographically similar soil classes, and/or 5) there were insufficient observation data to train the model (Lowry et al., 2008).

Individual Components

As mentioned previously, instead of applying weighting to better predict minority classes, the hypothetical extent was determined for each soil class according to a minimum confidence threshold based on potential limitations for land use (Table 6). Most pixels had a first most likely classification greater than the confidence threshold (Table 14). Many pixels had a second most likely classification greater than the minimum confidence threshold indicated in Table 14. A significant proportion of all pixels had a third component identified above the confidence threshold. Because relatively few pixels had a fourth class higher than the minimum confidence threshold, the fourth most likely soil class was not determined. The hypothetical extent (greater than the confidence threshold in Table 14) of each of the 24 predicted classes is shown in Figures 25 through 30. Overall, the results of the hypothetical extent of each component match very well with the soil-landscape relationships. As no independent accuracy assessment was made, an actual estimate of error cannot be provided beyond the likelihood of prediction for each pixel based on the votes of all trees in the grove.

Crestline (class 3) and Hiko Peak with Wyoming big sagebrush vegetation (class

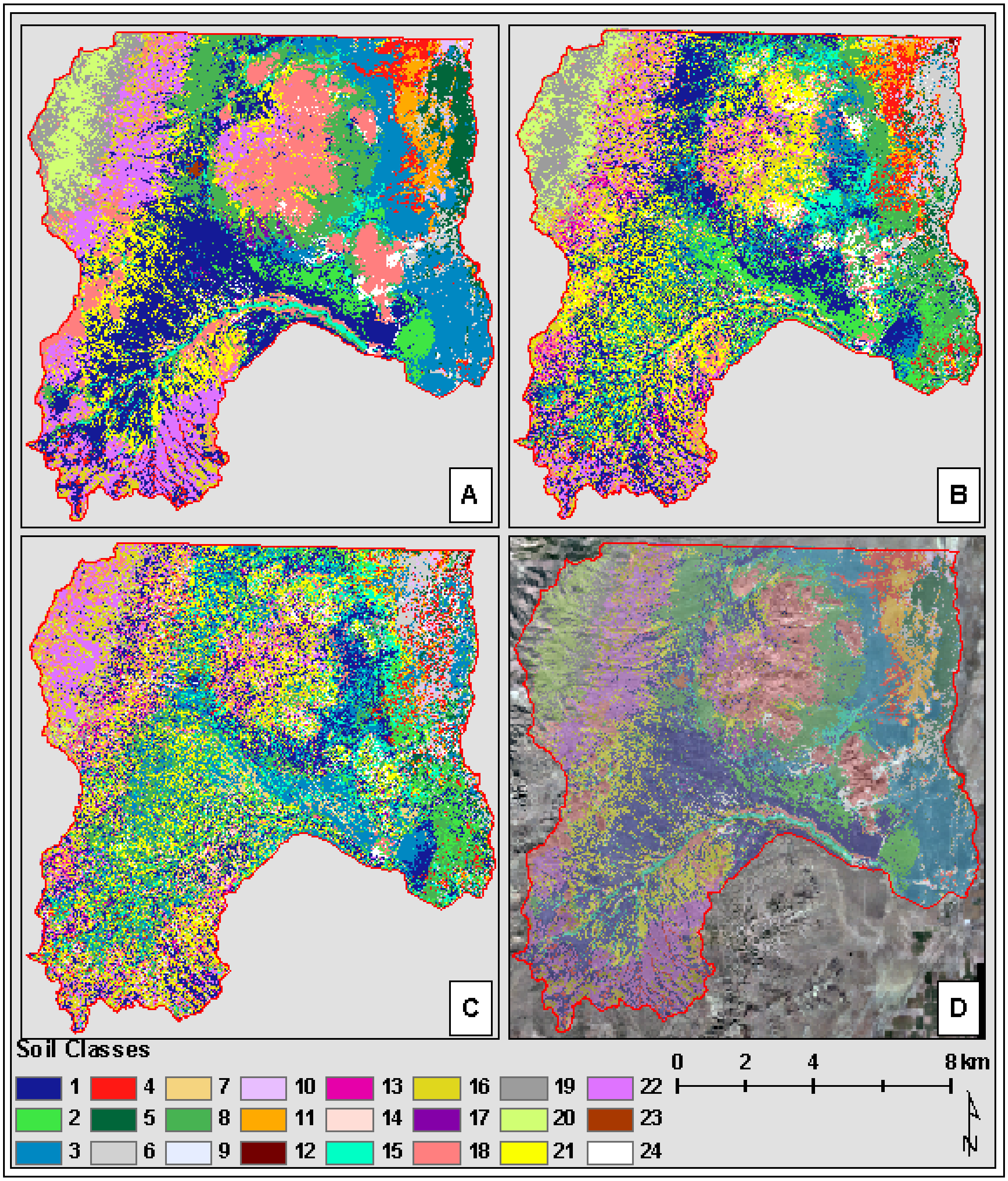


Figure 23. Model prediction outputs of Grove 2B. The primary (A), secondary (B) and tertiary (C) model outputs. D: This is the primary output made transparent to show aerial photography underneath.

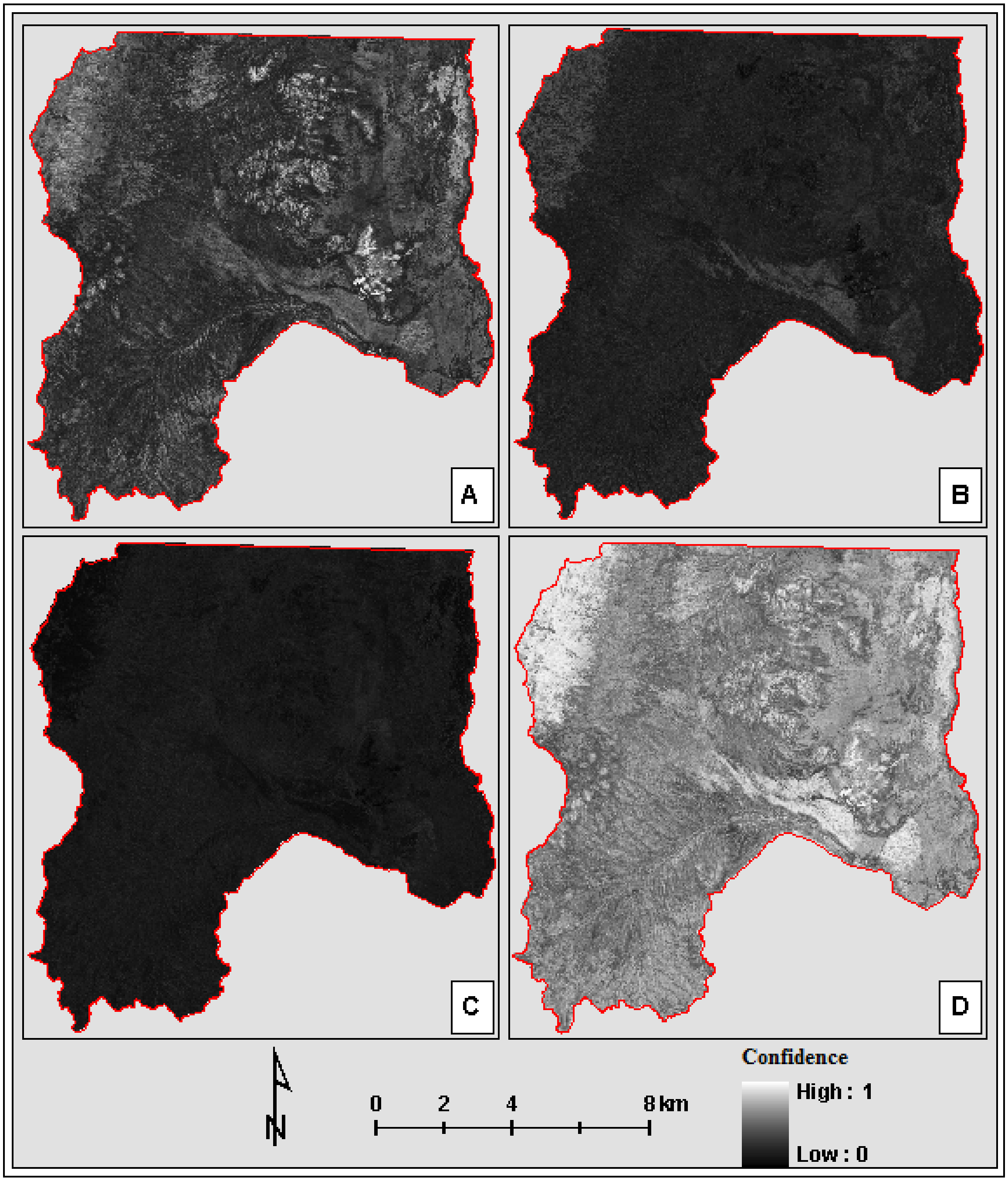


Figure 24. Model confidence outputs from Grove 2B. A: Confidence image showing the likelihood that a pixel belongs to the class predicted by majority of all trees in the grove. B: Confidence image showing the likelihood that a pixel may be the second most predicted class. C: Confidence image showing the likelihood that a pixel may be the third most predicted class. D: Confidence image showing the sum of likelihoods that a pixel may belong to one of the three most predicted classes for that pixel.

8) are very similar morphologically and support Wyoming big sage; the only difference is the quantity of rock fragments, with Crestline being coarse-loamy and Hiko Peak being loamy-skeletal (Figures 25 and 26). Crestline was found on the lower half of alluvial fans while Hiko Peak was found higher on these alluvial fans with steeper slopes. The transition in the middle is where these two classes literally mingle. The soil can change from one class to the other gradually or abruptly within a couple of meters reflecting the complicated formation of alluvial fans and the accumulation of sediment from distinct debris flow events.

Determining the hypothetical extent of Garbo (class 2) identified 42.8% more pixels (6,080 pixels) than were predicted by the primary output (Table 14). While weighting also increased the number of pixels predicted to be Garbo, recall that it poorly predicted commonly occurring classes like Dixie (class 1) while over predicting small classes like Biblesprings (class 7). Garbo (class 2) and Dixie (class 1) are morphologically very similar, and the only significant difference was the durinodic properties (secondary silica accumulations) present in Garbo (Figure 25). The hypothetical extent of Garbo often overlapped (~70% of pixels) with Dixie (Figure 31). Garbo was only found on lower fan remnants and some sites with Garbo had lower vegetation production (shorter stands of Wyoming big sage).

Sugarloaf soils (class 5) (Figure 26A) were consistently the best predicted soil class (mine dumps (class 24) was the best predicted class overall) whether weighting was applied or not (Table 9). Pixels classified as class 5 in the primary output (Figure 23A) had the second highest average confidence, 0.477 (Table 14). The successful prediction of this class can be attributed to the distinct vegetation (low density and low production:

Table 14. A summary of Grove 2B. *Number of Observations* is the number observations per class in the training dataset. *Original Count* is the number of pixels that were predicted by a majority of trees to belong to that class. *Mean of First* is the average confidence of pixels that were predicted by a majority of trees to belong to that class. *OOB error* is the OOB error by class. *Component Count* is the number of pixels that had a confidence greater than the *Confidence threshold* of belonging to that class.

Class	Number of Observations	Original Count	Mean of First	OOB error [%]	Component Count	Confidence threshold
1	77	98179	0.296	54.6	78403	0.25
2	38	14199	0.413	63.2	20279	0.25
3	61	59546	0.359	67.2	52779	0.25
4	19	12245	0.325	63.2	12889	0.25
5	18	15053	0.477	27.8	23671	0.15
6	16	7844	0.311	43.8	8762	0.25
7	13	156	0.186	100.0	19	0.25
8	65	57216	0.326	43.1	119136	0.15
9	10	24	0.254	100.0	781	0.15
10	5	2604	0.289	80.0	6593	0.15
11	10	11837	0.362	50.0	12203	0.25
12	2	71	0.231	100.0	368	0.15
13	13	41	0.284	100.0	3668	0.15
14	18	1688	0.244	72.2	4820	0.15
15	41	11697	0.304	56.1	32832	0.15
16	26	14936	0.299	65.4	40077	0.15
17	20	3696	0.228	80.0	14106	0.1
18	62	65231	0.394	24.2	94759	0.15
19	12	18907	0.422	50.0	38444	0.15
20	11	17082	0.517	45.5	33927	0.15
21	53	39417	0.260	84.9	95265	0.15
22	29	55932	0.344	51.7	81646	0.15
23	8	993	0.247	75.0	406	0.25
24	45	7137	0.365	8.9	1498	0.5

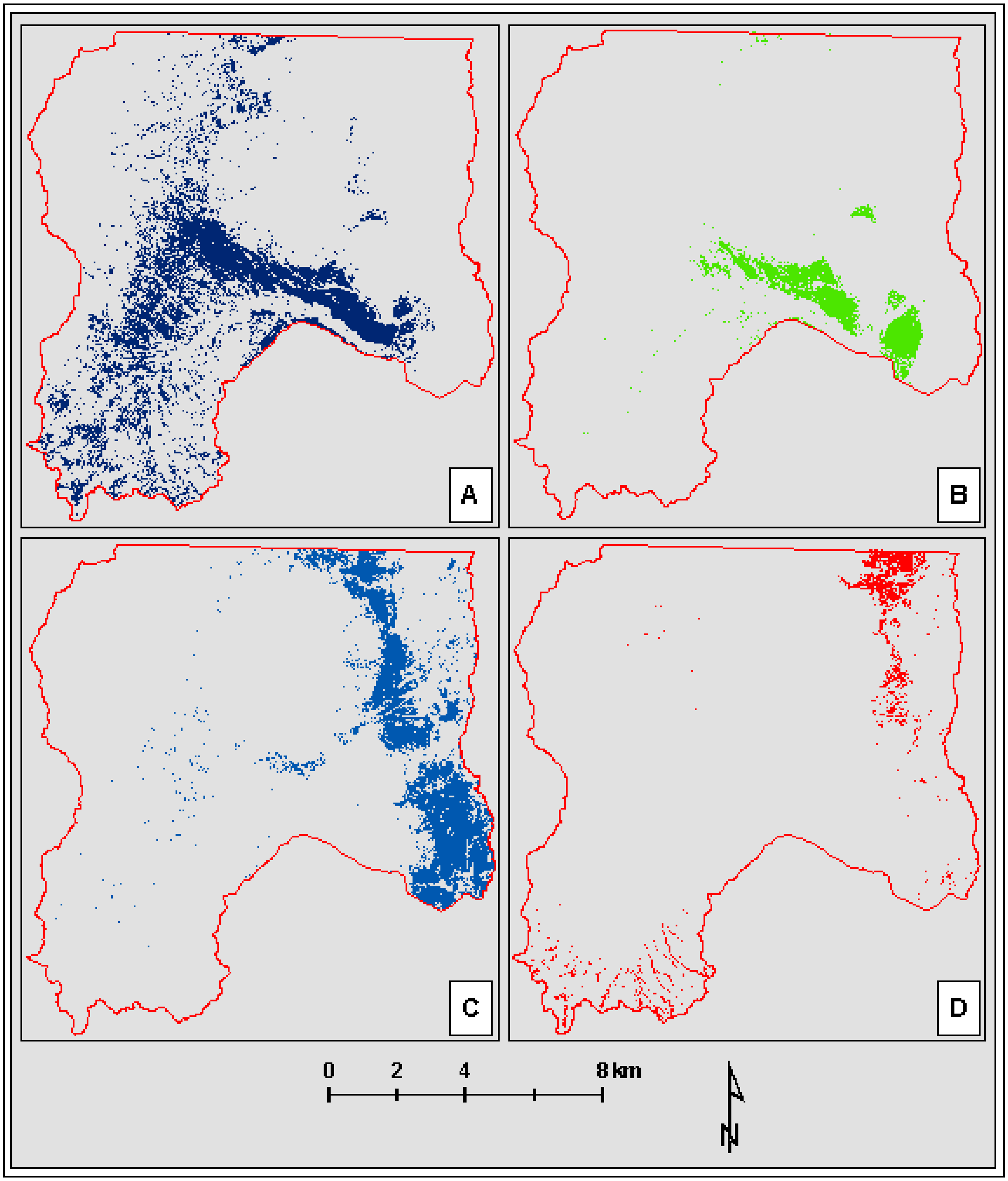


Figure 25. The hypothetical extents of classes 1 (A), 2 (B), 3 (C), and 4 (D).

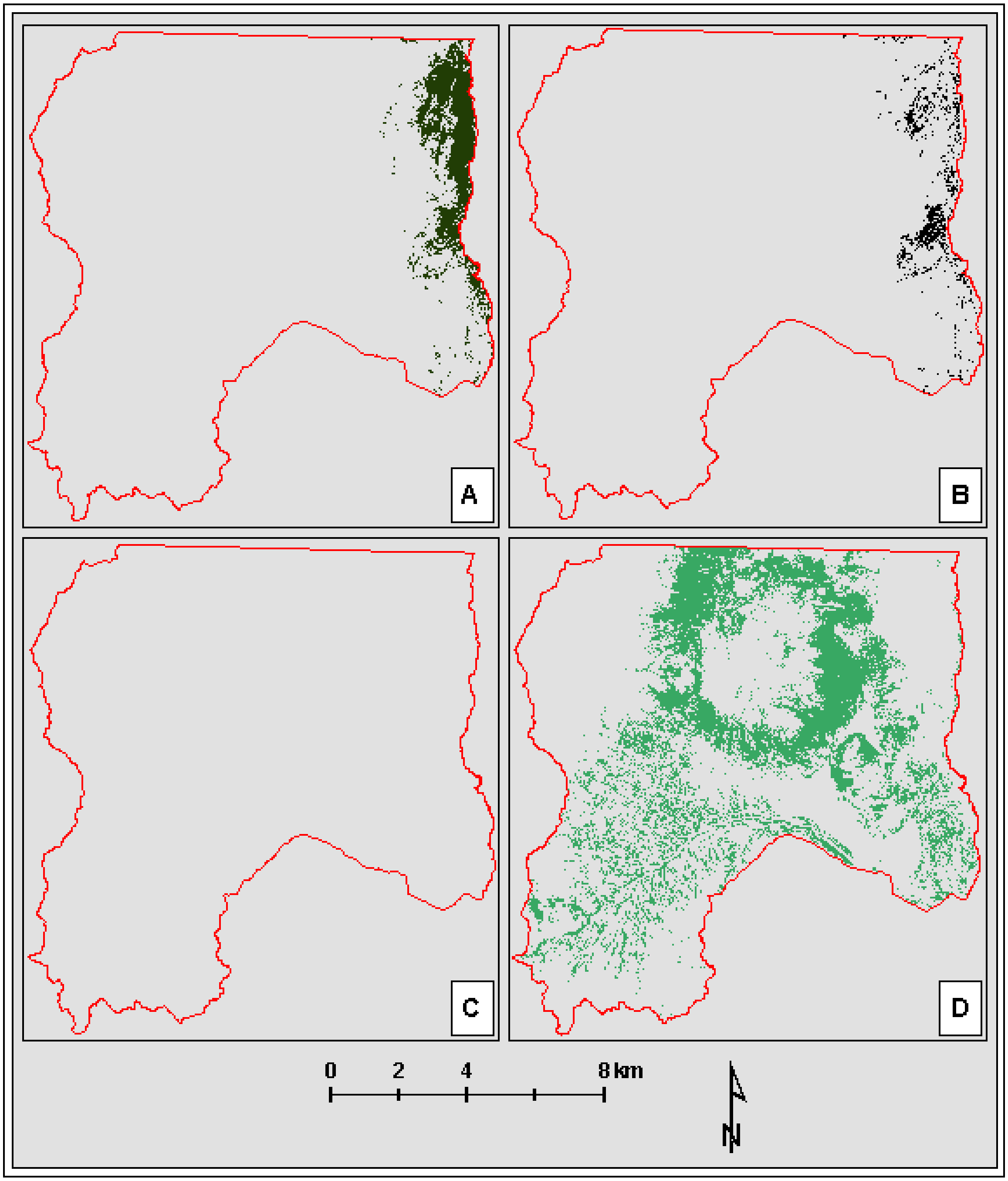


Figure 26. The hypothetical extents of classes 5 (A), 6 (B), 7 (C), and 8 (D).

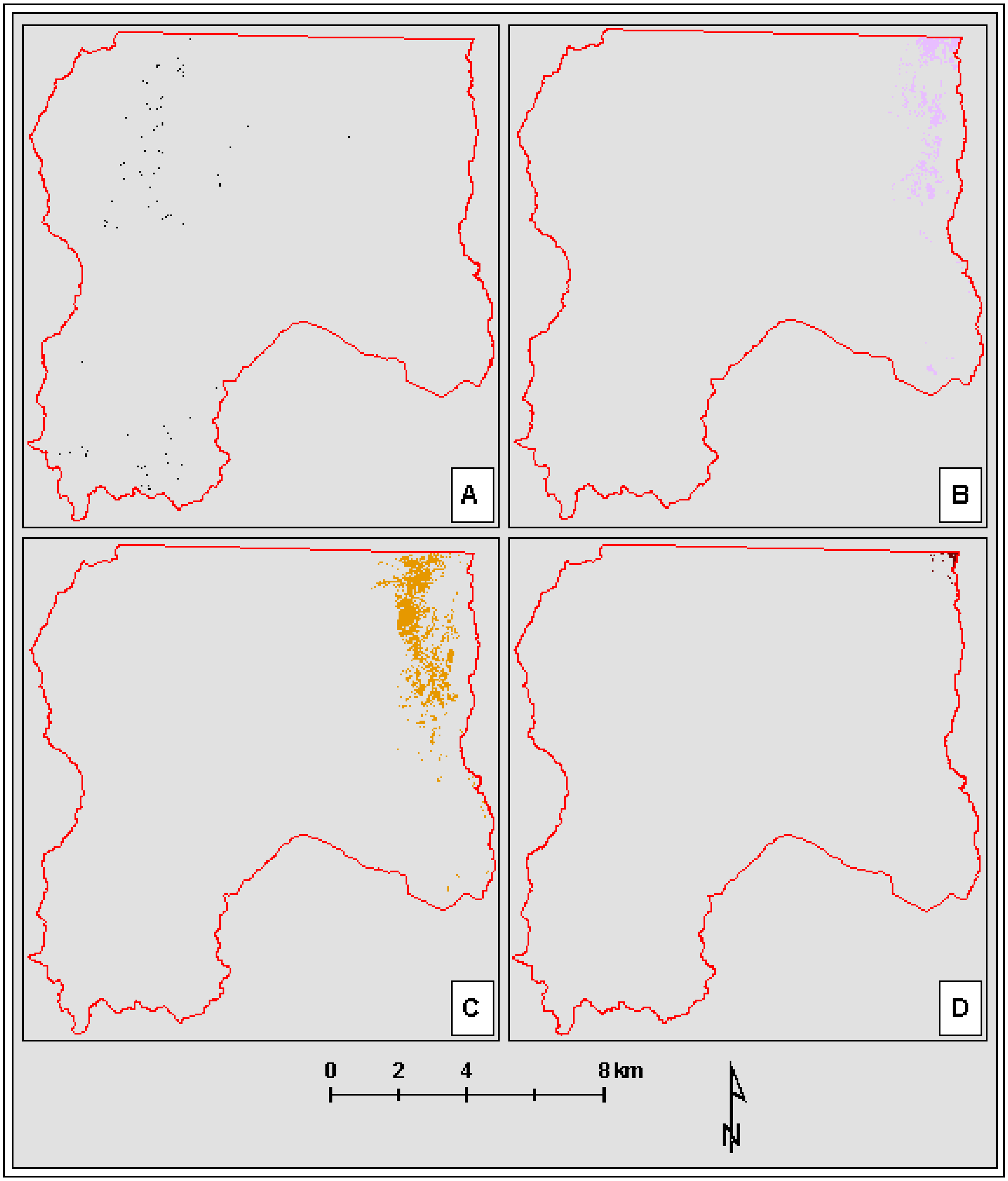


Figure 27. The hypothetical extents of classes 9 (A), 10 (B), 11 (C), and 12 (D).

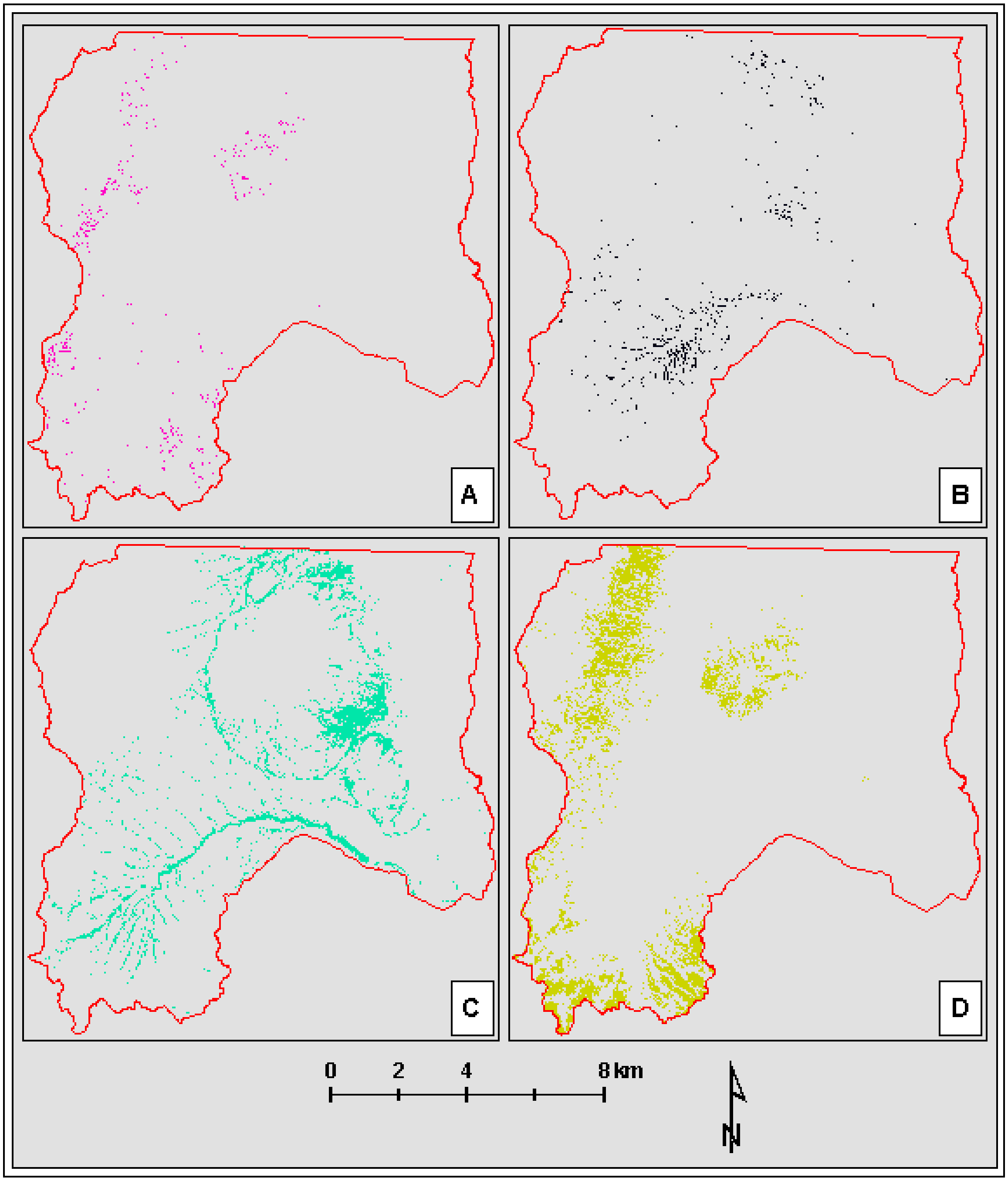


Figure 28. The hypothetical extents of classes 13 (A), 14 (B), 15 (C), and 16 (D).

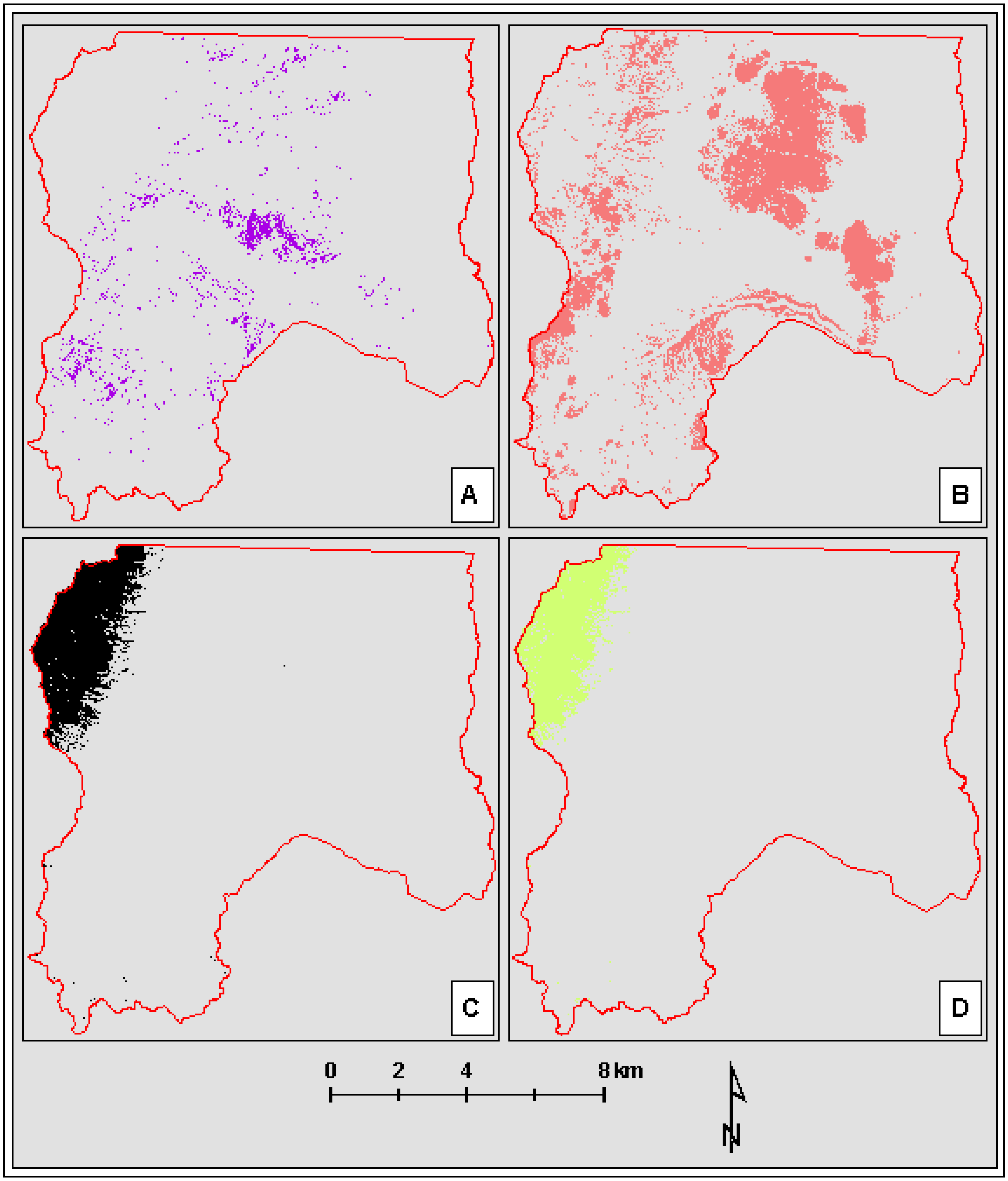


Figure 29. The hypothetical extents of classes 17 (A), 18 (B), 19 (C), and 20 (D).

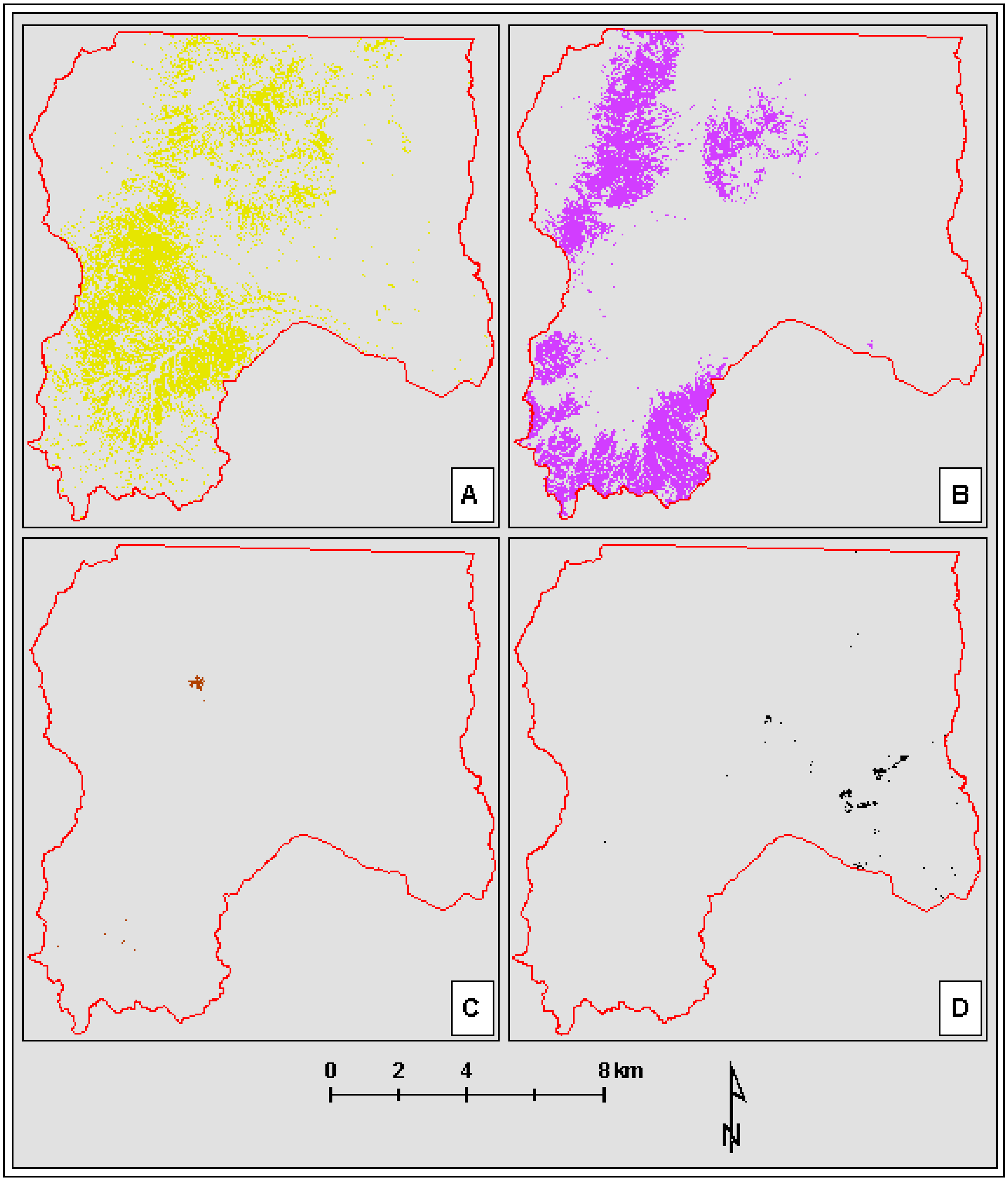


Figure 30. The hypothetical extents of classes 21 (A), 22 (B), 23 (C), and 24 (D).

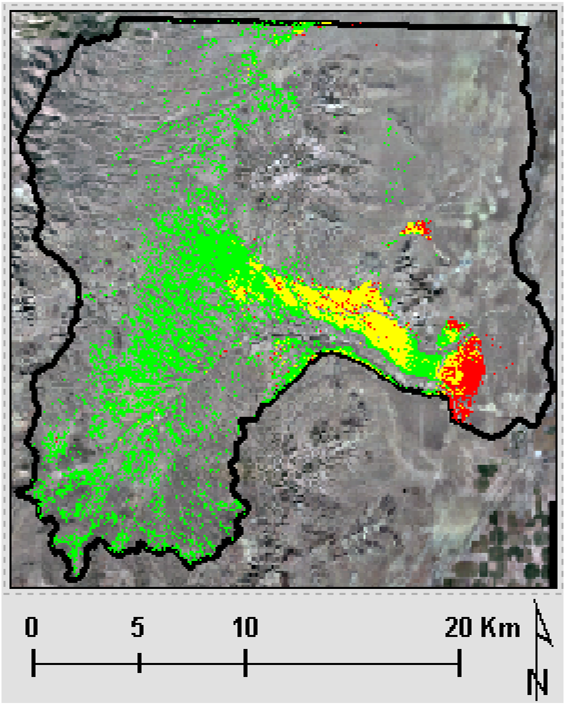


Figure 31. The hypothetical extents of Dixie (green), Garbo (red) and the overlap of Dixie and Garbo (yellow).

rabbitbrushes, ephedra and Indian rice grass) and unique geomorphic surface (large stream terrace) below the Lake Bonneville shoreline (Appendix C), despite only a modest number of observations (n=18, Table 14).

Uvada (class 12) is another of several examples of soil classes where the predicted output was improved by determining the hypothetical extent. Only 71 pixels, all in the northwest corner of the study area, were classified as class 12 (Uvada) in the primary output (Table 14). While this class was only found in the very northwest corner of the study area, 6.4 ha grossly under represented the occurrence of this class based on soil-landscape concepts developed in the field. By determining the hypothetical extent of class 12, 368 pixels (33 ha) in the very northwest corner were identified as Uvada (Figure 27D).

Map Units

While Grove 2B had the lowest overall OOB error of the four groves, it still had significant error. In analyzing the OOB error matrix, and as noted above, there are several soil classes which co-occurred on landforms, were spatially extensive (predicted to cover >60% of the area), and were frequently misclassified as each other (Table 12). Five soil classes (1, 2, 3, 8, and 21) were selected to create soil map units (associations and complexes) because they were frequently misclassified as each other. Hiko Peak classes 8 and 21 were treated as one class in this exercise because both classes had similar vegetation communities: class 8 under Wyoming big sage, class 21 under black sage, and some observation sites had a mosaic of both shrub types. Hiko Peak class 22 was under pinyon and Utah juniper vegetation, and was found at higher elevations on the fan piedmont surface than classes 1, 2, 3, 8, and 21. The number of times the two soil components were misclassified as each other are in parentheses: Hiko Peak [classes 8 or 21] – Dixie [class 1] (40); Crestline [class 3] – Hiko Peak [classes 8 or 21] (31); Dixie [class 1] – Garbo [class 2] (22); Dixie [class 1] – Crestline [class 3] (18) (Table 12).

The results of these new soil combinations are illustrated in Figure 32. The Dixie-Hiko Peak, Dixie-Garbo, and Crestline-Hiko Peak class combinations exhibited clustering on recognizable landforms suggesting they could make good map unit associations and complexes. The Dixie-Crestline combination did not co-occur often, were spatially scattered and, thus, would not make a good map unit complex or association.

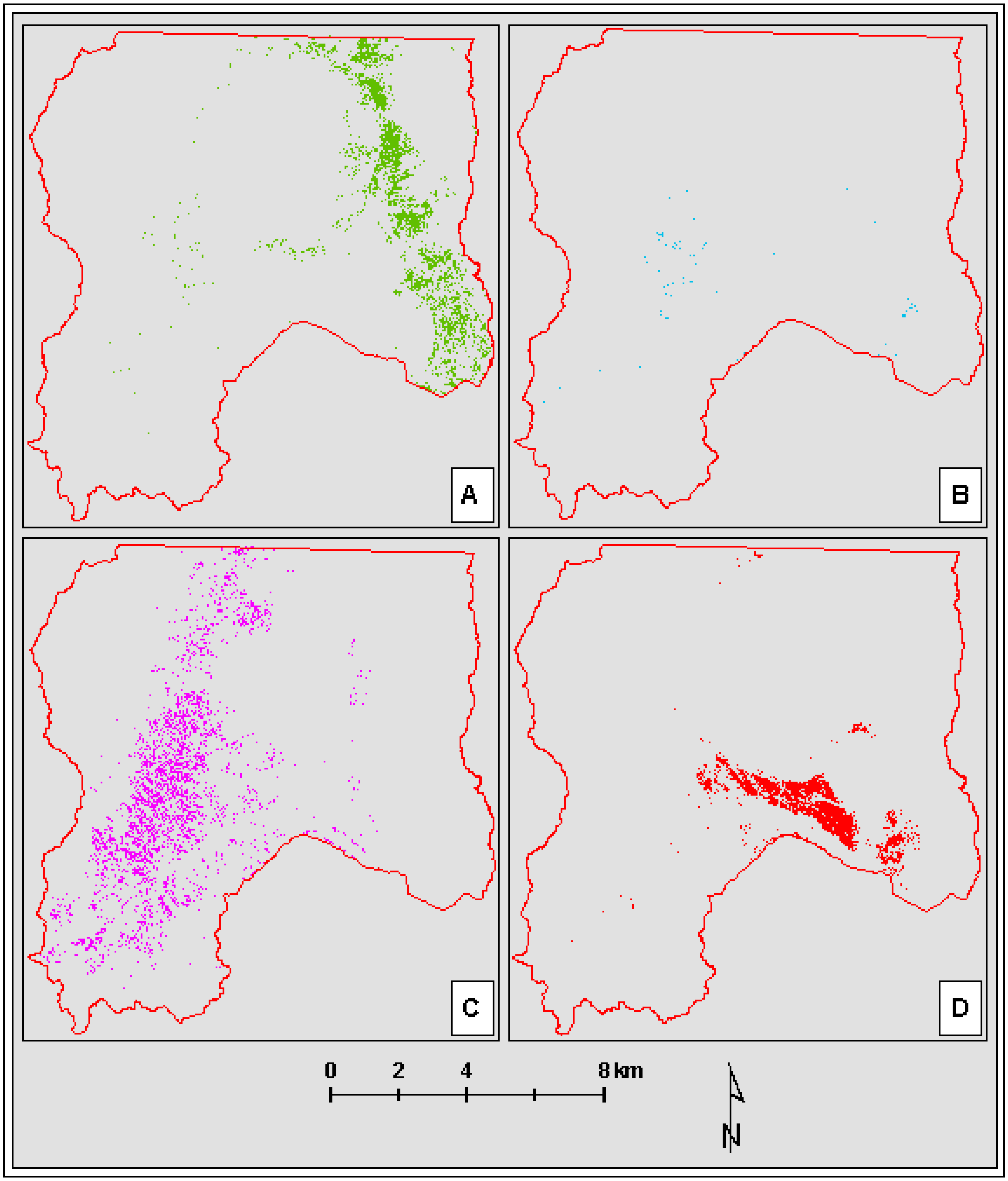


Figure 32. The results of the soil class combinations. A: Crestline (class 3) and Hiko Peak under Wyoming Big Sage or black sage (classes 8 or 21) B: Dixie (class 1) and Crestline (class 3). C: Dixie (class 1) and Hiko Peak under Wyoming Big Sage or black sage (classes 8 or 21). D: Dixie (class 1) and Garbo (class 2).

CONCLUSION

Random forests coupled with GIS was an effective spatial predictor of soil classes in a previously unmapped watershed. The prediction of soil classes was made for individual 30-m pixels. Both Landsat- and DEM-derived variables improved the overall prediction of soil classes. The Gini and the standard measures of variable importance identified Landsat- and DEM-derived variables as strong predictive variables. However, DEM-derived data were the strongest predictive variables based on both variable and importance measures and new groves grown with either only DEM-derived variables or Landsat-derived variables.

Weighting had a much greater overall effect on OOB error than the number of environmental covariates selected at random at each node of a decision tree. Weighting resulted in greater overall OOB error, higher prediction error of the commonly occurring soil classes that covered the majority of landforms in the study area, and lower prediction error for minor soil classes. Furthermore, the spatial predictions from the weighted groves did not reflect expected soil-landscape relationships observed in the field.

Growing 500 independent trees, each with a vote on final classification, provided a mechanism for estimating model confidence for each individual pixel. This also allowed for further analysis of the results beyond the primary (final) classification of each pixel. Identifying the second and third most likely classification of a pixel objectively predicted the hypothetical extent of individual soil components. Using model confidence (likelihood) to expand the hypothetical extent of minority classes seems to be an effective alternative to weighting but without sacrificing the prediction of common classes. Similar pixels were aggregated where there was the likely occurrence (above a confidence threshold) of two or more soils components. Identifying the hypothetical extent of individual soil classes allowed for the user to objectively aggregate similar pixels as possible map unit associations and complexes.

Furthermore, RF outputs included a map of model uncertainty, which can be used to direct further field documentation. Additional field documentation can be collected where model uncertainty is high, and additional data can be used to refine the model. Further research with multiple field observations within close proximity (<90-m) may help explain why multiple likely classes were predicted for one pixel and elucidate the variation of soils with respect to different landforms.

While some of the error is a result of sampling, some error can be caused by limitations of the predictive variables. Soil classes like Dixie (class 1) and Garbo (class 2) so similar morphologically that distinguishing them with a model may not be practical. Error and areas of low confidence may be improved with different predictive variables. Since this study was conducted, 5-m DEMs have become available for the study area. These DEMs were derived from digital orthophotography and have fewer relicts from interpolation than the 10-m DEM from National Elevation Dataset (NED) used in this study. Other DEM-derivations could also be tried, such as stream power or relative distance to hill tops.

Digital soil mapping using random forests has many quantifiable advantages over traditional soil mapping. Traditional soil maps have no error assessment or estimated accuracy. Random forests provided objective estimates of the model’s accuracy via the OOB error and the likelihood of prediction for each pixel. The OOB error is strictly based on out of the bag observations being classified as the right class or not. The likelihood (confidence) can be used to predict hypothetical extent of minor and similar classes. In a third order soil survey, the minimum delineation is a 5 acre (2-ha) polygon and the legend would allow for 3 to 7 possible outcomes as each soil map unit usually consists of 1-3 major soil components and several minor soil components.

REFERENCES

AGRC, 2008. Utah Automated Geographic Reference Center: Utah GIS portal. http://gis.utah.gov/gisresources (last verified April 25, 2010)

Amen, A.E., Foster, J.W., 1987. Soil Landscape Analysis Project (SLAP) methods in soil surveys. Bureau of Land Management technical note 379.

Anderson, K., Croft, H., 2009. Remote sensing of soil surface properties. Prog. in Phys. Geogr. 33, 457-473.

Baer, J.L., 1973. Summary of stratigraphy and structure of the Star Range, Beaver County, Utah. In: Hintze, L.F., Whelan, J.A. (Eds.) Geology of the Milford Area: Utah Geological Association, publication UGA-3, pp. 33-38.

Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Prentice Hall, Harlow, England.

Best, M.G., Lemmon, D.M., Morris, H.T., 1989. Geologic Map of the Milford Quadrangle and East Half of the Frisco Quadrangle, Beaver County, Utah. USGS, Reston.

Birkeland, P.W., 1999. Soils and Geomorphology, 3rd ed. Oxford University Press, New York.

BLM, 2006. U.S. Department of Interior Bureau of Land Management: National: About the BLM. http://www.blm.gov/wo/st/en/info/About_BLM.html (last verified July 28, 2010).

Bodily, J.M., 2005. Developing a digital soil survey update protocol at the Golden Spike National Historic Site. M.S. Thesis, Utah State University, Logan.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5-32.

Breiman, L., Cutler, A., 2009. Random Forests homepage. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (last verified 13 August 2009).

Brungard, C.W., 2009. Alternative sampling and analysis methods for digital soil mapping in southwestern Utah. M.S. Thesis. Utah State University, Logan.

Chaplot, V., Darboux, F., Bourennane, H., Leguedois, S., Silvear, N., Phachomphon, K., 2006. Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. Geomorphol. 77,126-141.

Chavez, P.S., Jr., 1996. Image-based atmospheric corrections – revisited and improved. Photogrammetric Eng. and Remote Sens. 62,1025-1036.

Crittenden, M.D., 1963. New data on the isostatic deformation of the Lake Bonneville: U.S. Geological Survey Professional Paper 454-E, 31.

Cole, N.J. 2004. A pedogenic understanding raster-based classification model for mapping soils in the Powder River Basin, Wyoming. M.S. Thesis, Utah State University, Logan.

Cole, N.J., Boettinger, J.L., 2007. A pedogenic understanding raster classification methodology for mapping soils, Powder River Basin, Wyoming, USA. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam, pp. 377-388.

Crosby, G.W., 1973. Regional structure in Southwestern Utah. In: L.F Hintze and J.A. Whelan (Eds.), Utah Geological Association, publication UGA-3, pp. 27-32.

DeMers, M.N., 2000. Fundamentals of Geographic Information Systems, 2nd ed. John Wiley & Sons, Inc., New York.

Di Paolo, W.D., Hall, L.B., 1983. The use of remote sensing for soils investigations on BLM lands, Technical Note 361, U.S. Department of Interior, Bureau of Land Management.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., Helt, T., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. Geoderma 97,367-391.

East, E.H., 1966. Structure and stratigraphy of San Francisco Mountains, Western Utah. Bulletin of the American Association of Petroleum Geologists 50, no. 5, 901-920.

El Rakaiby, M.L., Ashmawy, M.H., Yehia, M.A., Ayoub, A.S., 1994. *In situ* reflectance measurements and TM data of some sedimentary rocks with emphasis on white sandstone, southwestern Sinai, Egypt. Int. J. Remote Sens. 15,3785-3797.

Erickson, M.P., 1973. Volcanic rocks of the Milford area, Beaver County, Utah. In : L.F Hintze and J.A. Whelan (Ed.) Utah Geological Association, publication UGA-3, pp. 13-21.

Evans, J., 2004. Compound topographic index ArcScript. http://arcscripts.esri.com/details.asp?dbid=11863 (last verified October 6, 2009).

Fiero, B., 1986. The Geology of the Great Basin. University of Nevada Press, Las Vegas.

Fisher, R.V., Schmincke, H.-U., 1984. Pyroclastic Rocks. Springer-Verlag, Berlin, Germany.

Gessler, P.E., Chadwick, O.A., Chamran, F., Althouse, L., Holmes, K. 2000. Modeling soil-landscape and ecosystem properties using terrain attributes. Soil Sci. Soc. Am. J. 64,2046-2056.

Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modeling and spatial prediction of soil attributes. Int. J. Geogr. Inf. Systems 9,421-432.

Gilbert, G.K., 1890. Lake Bonneville: U.S. Geological Survey Monograph 1, (Lake Bonneville Diastrophism). 362-381.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognit. Lett. 27,294-300.

Goetz, A.F.H., 1989. Spectral remote sensing in geology. In: G. Asrar (Ed.) Theory and Applications of Optical Remote Sensing, Wiley, New York, pp. 491-526.

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. Geoderma 146,403-411.

Heimsath, A.M., Dietrich, W.E., Nishiizumi, K., Finkel, R.C., 1997. The soil production function and landscape equilibrium. Nature 388,358-361.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124,383-398.

Hintze, L.F., 1993. Geologic History of Utah. Brigham Young University, Provo, UT.

Hintze, L.F., Lemmon, D.M., Morris, H.T., 1984. Geologic map of the Frisco Peak quadrangle, Millard and Beaver Counties, Utah. USGS, Reston, VA.

Hudson, B.D. 1992. The soil survey as paradigm-based science. Soil Sci. Soc. Am. J. 56,836-841.

IRDIAC, 2006. Intermountain Region Digital Image Archive Center. http://earth.gis.usu.edu/ (last verified October 6, 2009)

Irons, J.R., Weismiller, R.A., Petersen, G.W., 1989. Soil reflectance pp. 66-106 In: G. Asrar (Ed.) Theory and Applications of Optical Remote Sens., Wiley, New York.

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York.

Jenny, H., 1980. The Soil Resource. Springer-Verlag, New York.

Jensen, J.R., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective, 3rd ed. Pearson Prentice Hall, Upper Saddle River, NJ.

Lagacherie, P., Holmes S. 1997. Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geogr. Inf. Sci. 11,183-198.

Leica, 2005. ERDAS Field Guide. Leica Geosystems, Norcross, GA.

Lemmon, D.M., Morris, H.T., 1984. Geologic map of the Beaver Lake Mountains quadrangle, Millard and Beaver Counties, Utah. USGS, Reston, VA.

Lookingbill, T., Urban, D., 2004. An empirical approach towards improved spatial estimates of soil moisture for vegetation analysis. Landsc. Ecol. 19,417- 433.

Lowry, J.H., Ramsey, R.D., Langs-Stoner, L., Kirby, J., Schulz, K., 2008. An ecological framework for evaluating map errors due to class similarity using fuzzy sets. Photogrammetric Eng. and Remote Sens. 74,1509-1519.

McBratney, A.B., Minasny, B., Cattle, S., Veroort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109,41-73.

McBratney, A.B., Mendonça Santos, M.L., Minansy, B., 2003. On digital soil mapping. Geoderma 117,3-52.

Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhancing soil map modeling. Int. J. Geogr. Inf. Sci. 16,533-549.

NASA, 2008. ASTER spectral library. http://speclib.jpl.nasa.gov/ (last verified September 28, 2009).

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

Noller, J., 2010. Applying geochronology in predictive digital mapping of soils. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, Dordrecht, The Netherlands, pp. 43-53.

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with Fuzzy-c-means: Application to classification and soil-landform interrelationships. Soil Sci. Soc. Am. J. 56,505-516.

Peters, J., De Baets, B., Verhoest, N.E.C, Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modeling. Ecol. Model. 207,304-318.

R, 2007. R v. 2.x, a statistical language and package. www.r-project.org (last verified October 6, 2009).

Rees, W.G., 2001. Physical Principles of Remote Sensing, 2nd ed. Cambridge University Press, Cambridge, UK.

RSGIS, 2003. Remote Sensing and Geographic Information Systems Laboratory, Utah State University. http://earth.gis.usu.edu/imagestd/ (last verified 13 August 2009).

Salford Systems, 2004. Random Forests (software help guide). San Diego, CA.

Saunders, A.M., 2005. Incorporating classification tree analysis into the pedogenic understanding raster classification methodology, Green River Basin, Wyoming. M.S. Thesis, Utah State University, Logan.

Saunders, A.M., Boettinger, J.L., 2007. Incorporating classification trees into a pedogenic understanding raster classification methodology, Green River Basin, Wyoming, USA. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping: An introductory perspective. Developments in Soil Science Vol. 31. Elsevier, Amsterdam, p.389-399.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Progress in Phys. Geogr. 27,171-197.

Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecol. Model. 181,1-15.

Shi, X., Zhu, A.X., Burt, J.E., Qi, F., Simonson, D., 2004. A case-based reasoning approach to fuzzy mapping. Soil Sci. Soc. Am. J. 68,885-894.

Soil Survey Division Staff, 1993. Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18, Washington, D.C. Available online at: http://www.soils.usda.gov/technical/manual/ (last verified January 2, 2010).

Soil Survey Staff, 2003. Keys to Soil Taxonomy, 9th ed. U.S. Department of Agriculture, Natural Resources Conservation Service, Washington, D.C. Available online at: http://www.soils.usda.gov/technical/classification/tax_keys/archive.html (last verified April 26, 2010).

Stokes, W.L., 1988. Geology of Utah. Utah Museum of Natural History and Utah Geological Survey, Salt Lake City, UT.

Tarboton, D.G., 1997. A new method for the determination of flow directions and contributing areas n grid digital elevation models. Water Resources Res., 33,309-319.

Tarboton, D.G., 2005. Terrain Analysis Using Digital Elevation Models v. 3.1. http://hydrology.neng.usu.edu/taudem/ (last verified April 27, 2010)

USDA-NRCS, 2000. Parameter-elevation Regression on Independent Slopes Model (PRISM). http://www.wcc.nrcs.usda.gov/climate/prism.html (last verified October 6, 2009).

USDA-NRCS, 2009. National Soil Survey Handbook, title 430-VI, Part 647. Available online at: http://soils.usda.gov/technical/handbook/ (last verified October 31, 2009).

U.S. Geological Survey Geologic Names Committee, 2007. Divisions of geologic time—Major chronostratigraphic and geochronologic units: U.S. Geological Survey Fact Sheet 2007-3015, 2 p. Available online at: http://pubs.usgs.gov/fs/2007/3015/ (last verified January 2, 2010).

USU Extension, 2009. Range Plants of Utah. http://extension.usu.edu/rangeplants/ (Last verified 6 October 2009)

Welsh, J.E., 1973a. Paleozoic and Mesozoic stratigraphy. In: Hintze, L.F., Whelan, J.A. (Eds.) Utah Geological Association, publication UGA-3, pp. 9-12.

Welsh, J.E., 1973b. Geology of the Beaver Lake Mountains, Beaver County, Utah. In: Hintze, L.F., Whelan, J.A. (Eds.), Utah Geological Association, publication UGA-3, pp.49-53.

Western Regional Climate Center (WRCC), 2005. www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?utmilf (last verified October 6, 2009).

Whelan, J.A., 1973a. Mineral resources of the Milford area, Beaver County, Utah. In: Hintze, L.F., Whelan, J.A. (Eds.), Utah Geological Association, publication UGA-3, pp. 1-4.

Whelan, J.A., 1973b. Geology of the Rocky Range, Beaver County, Utah. In: Hintze, L.F., Whelan, J.A. (Eds.), Utah Geological Association, publication UGA-3, p. 55.

Williams, J.K., Abernethy, J., 2008. Using random forests and fuzzy logic for automated storm type identification. American Meteorological Society Sixth Conference on Artificial Intelligence Applications to Environmental Science, New Orleans, LA.

Winward, A.H., 2004. Sagebrush of Colorado: Taxonomy, distribution, ecology and management. Colorado Division of Wildlife, Denver CO.

Woodward, L.A., 1973. Upper Precambrian and lower Cambrian rocks of the Milford area, Utah. In: Hintze, L.F., Whelan, J.A. (Eds.), Utah Geological Association, publication UGA-3, pp. 5-8.

Zhu, A-X., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resources Res. 36,663-677.

APPENDICES

Appendix A: R Code

```
#Read in the tables
#This is your table containing your training data
sample=read.table("train_RF6_7.txt", header=1, sep=",")
#This is the table with the class ID of your training data
dep_id=read.table("soil_CBR_id.txt", header=1, sep=",")
# This is the table containing your unknown points (raster)
unknown=read.table("unknown_final.txt", header=1, sep=",")

#Append sample table with the environmental covaritates
train_RF=cbind(dep_id[,3],sample[,5:25])

#Apply logical column names
colnames(train_RF)=
        c("MU","MIN4_3","MIN4_5","MIN3_7","MIN5_2","MIN5_1","MIN4_7",
        "Land1","Land2","Land3","Land4","Land5","Land6","CTI5x5","slope",
        "curve","slope11x11","CTI","lake","aspect","pj","elev")


predict_RF=cbind(unknown[,5:25])
colnames(predict_RF)=
        c("MIN4_3","MIN4_5","MIN3_7","MIN5_2","MIN5_1","MIN4_7",
        "Land1","Land2","Land3","Land4","Land5","Land6","CTI5x5","slope",
        "curve","slope11x11","CTI","lake","aspect","pj","elev")


#Export in CSV format – best format to use in RF
write.csv(train_RF, file="train_RF6_7.csv", row.names=FALSE)
write.csv(predict_RF, file="predict_RFinal.csv", row.names=FALSE)

#exporting results
# Read in the table produced from RF
results=read.table("results_final.csv",header=1,sep=",")
# Append the UTM coordinates
results_p=cbind(unknown[,3:4],results[,2])
colnames(results_p)=c("X","Y","prediction")
# Save it all to a .csv to import into ArcGIS
write.csv(results_p, file="result_points6_7b.csv", row.names=FALSE)

#Finding the second and third most likely class

blank=results[,3:26]
high=results[,2]
```

```
y=c(1:24)
mat=matrix(y,dim(blank)[1],24,byrow=1)
high_index=mat==high
blank[high_index]=0
second=max.col(blank)

second_index=mat==second
blank[second_index]=0
third=max.col(blank)

second_best=cbind(unknown[,3:4],second)
colnames(second_best)=c("X","Y","second")
write.csv(second_best,file="second6_7b.csv",row.names=FALSE)

third_best=cbind(unknown[,3:4],third)
colnames(third_best)=c("X","Y","third")
write.csv(third_best,file="third6_7b.csv",row.names=FALSE)


#Uncertainty
# Below is exporting tables of uncertainty for the three most likely choices
blank=results[,3:26]
blank=t(blank)
third_index=mat==third
uncertain=blank[t(high_index)]
sec_uncertain=blank[t(second_index)]
third_uncertain=blank[t(third_index)]
sum=uncertain+sec_uncertain
third_sum=sum+third_uncertain

uncertain=cbind(unknown[,3:4],uncertain)
colnames(uncertain)=c("X","Y","Uncertainty")
write.csv(uncertain,file="uncertain6_7b.csv",row.names=FALSE)

sec_uncertain=cbind(unknown[,3:4],sec_uncertain)
colnames(sec_uncertain)=c("X","Y","Uncertainty")
write.csv(sec_uncertain,file="sec_uncertain6_7b.csv",row.names=FALSE)

sum=cbind(unknown[,3:4],sum)
colnames(sum)=c("X","Y","Uncertainty")
write.csv(sum,file="sum6_7b.csv",row.names=FALSE)

third_uncertain=cbind(unknown[,3:4],third_uncertain)
colnames(third_uncertain)=c("X","Y","Uncertainty")
write.csv(third_uncertain,file="third_uncertain6_7b.csv",row.names=FALSE)
```

```
third_sum=cbind(unknown[,3:4],third_sum)
colnames(third_sum)=c("X","Y","Uncertainty")
write.csv(third_sum,file="third_sum6_7b.csv",row.names=FALSE)

#Summarize variables by class
class=results[,2]
variables=unknown[,5:25]
class_mean=aggregate(variables,list(class),mean)
class_sd=aggregate(variables,list(class),sd)
class_sum=rbind(class_mean,class_sd)
write.csv(class_sum,file="class_summary.csv")

#Summarize thematic class PJ

wood=cbind(0,0,0,1:24)[,1:3]
jun=variables[,20]==1
pin=variables[,20]==2
non=variables[,20]==3
c=jun
c[]=1
count=aggregate(c,list(class),sum)

a=aggregate(variables[jun,20],list(class[jun]),sum)
a=type.convert(as.matrix(a))
wood[a[,1],1]=a[,2]

a=aggregate(variables[pin,20],list(class[pin]),sum)
a=type.convert(as.matrix(a))
wood[a[,1],2]=a[,2]/2

a=aggregate(variables[non,20],list(class[non]),sum)
a=type.convert(as.matrix(a))
wood[a[,1],3]=a[,2]/3

percent=wood/count[,2]
write.csv(percent,file="wooded.csv")

#Component statistics
first=readBin("sub/first",integer(),size=1,n=767360,signed=0)
p1=readBin("sub/p1",double(),size=4,n=767360,signed=0,endian="swap")
second=readBin("sub/second",integer(),size=1,n=767360,signed=0)
p2=readBin("sub/p2",double(),size=4,n=767360,signed=0,endian="swap")
third=readBin("sub/third",integer(),size=1,n=767360,signed=0)
p3=readBin("sub/p3",double(),size=4,n=767360,signed=0,endian="swap")
```

```
min=.15
d=cbind(first,p1,1)
index=(d[,1]==12|d[,1]==14|d[,1]==21|d[,1]==13|d[,1]==10|d[,1]==16|d[,1]==15
|d[,1]==8|d[,1]==22|d[,1]==18|d[,1]==19|d[,1]==5|d[,1]==20)&d[,2]>=min
sum1=aggregate(d[index,2],list(d[index,1]),sum)
count1=aggregate(d[index,3],list(d[index,1]),sum)

d=cbind(second,p2,1)
index=(d[,1]==12|d[,1]==14|d[,1]==21|d[,1]==13|d[,1]==10|d[,1]==16|d[,1]==15
|d[,1]==8|d[,1]==22|d[,1]==18|d[,1]==19|d[,1]==5|d[,1]==20)&d[,2]>=min
sum2=aggregate(d[index,2],list(d[index,1]),sum)
count2=aggregate(d[index,3],list(d[index,1]),sum)

d=cbind(third,p3,1)
index=(d[,1]==12|d[,1]==14|d[,1]==21|d[,1]==13|d[,1]==10|d[,1]==16|d[,1]==15
|d[,1]==8|d[,1]==22|d[,1]==18|d[,1]==19|d[,1]==5|d[,1]==20)&d[,2]>=min
sum3=aggregate(d[index,2],list(d[index,1]),sum)
count3=aggregate(d[index,3],list(d[index,1]),sum)

#Preparing attributes
#Read in the tables
sample=read.table("sample7_19.txt", header=1, sep=",")
dep_id=read.table("soil_points_table.txt", header=1, sep=",")
unknown=read.table("unknown30mB.txt", header=1, sep=",")

#Append sample table with the observed environmental covariates
attributes=matrix(cbind(dep_id[,13],dep_id[,14],as.logical(dep_id[,20]),
as.logical(dep_id[,21]),!is.na(as.logical(dep_id[,24])))
,dim(sample)[1],5)
colnames(attributes)=c("Texture","Restriction","Calcic","Argillic","Durinodic")

#Texture
attributes[which(attributes[,1]==7,arr.ind=1),1]=6
attributes[which(attributes[,2]==3,arr.ind=1),2]=2
attributes[which(dep_id[,23]>0,arr.ind=1),2]=3

train_RF=cbind(attributes,sample[,5:25])
colnames(train_RF)=
        c("Texture","Restriction","Calcic","Argillic","Durinodic",
        "MIN4_3","MIN4_5","MIN3_7","MIN5_2","MIN5_1","MIN4_7",
        "Land1","Land2","Land3","Land4","Land5","Land6","CTI5x5","slope",
        "curve","slope11x11","CTI","lake","aspect","pj","elev")
train_RF[1,]
```

```
predict_RF=cbind(unknown[,5:25])
colnames(predict_RF)=
        c("MIN4_3","MIN4_5","MIN3_7","MIN5_2","MIN5_1","MIN4_7",
        "Land1","Land2","Land3","Land4","Land5","Land6","CTI5x5","slope",
        "curve","slope11x11","CTI","lake","aspect","pj","elev")


#Export in CSV format - best format to use in RF
write.csv(train_RF, file="att_train7_19.csv", row.names=FALSE)
write.csv(predict_RF, file="predict_RF6_7.csv", row.names=FALSE)

#exporting results
results=read.table("texture_results7_19.csv",header=1,sep=",")
results_p=cbind(unknown[,3:4],results[,2])
colnames(results_p)=c("X","Y","texture")
write.csv(results_p, file="texture_points7_19.csv", row.names=FALSE)
```

Appendix B: Prediction of Soil Attributes

I thought that predicting individual attributes could be an effective way to determine the taxonomic soil class, where predicted attributes of the soil lead to the determination of the soil class (specifically the soil series). Currently, soil attributes are inferred from class membership as predicted by the model. For example for example if Grove 2B predicts that a pixel is Garbo, we would then infer that it has a fine-loamy particle size family class, has both argillic and calcic horizons, does not have bedrock or petrocalcic contacts within 100 cm, and has durinodic properties.

Perhaps more ideally, if individual groves were grown to predict each one of these attributes for a pixel, we then could determine the class based on the predicted attributes. If it were predicted that a pixel was fine-loamy, has an argillic and calcic horizons, no bedrock or petrocalcic contacts, and has durinodic properties, it would be a fine-loamy, mixed, mesic Durinodic Xeric Calciargid or the Garbo soil series in this study area.

Another advantage is the concentration of the sample into fewer classes, e.g. five particle size family classes commonly occur throughout the survey area as opposed to 24 soil classes, or simply two classes to predict the presence or absence of argillic horizons.

The difficulty with this approach was deciding on appropriate class weights, if any, and ascertaining the accumulative error from the individual attribute groves. Still it is the author’s opinion that classification of soil class from the prediction of individual soil attributes is a valid approach and perhaps more desirable in some applications. Also, one could implement regression trees to predict continuous values such as depth to argillic or calcic horizon. The prediction of such would require much higher data resolution and denser sampling regime as it is a specific attribute as oppose to a more general thematic output.

Random Forests groves were grown predicting the presence of argillic horizon, presence of calcic horizon, presence of durinodic properties, presence of a contact with bedrock or petrocalcic (“limiting contact”), and soil particle size family classes. Only the actual field observations (561) were used to train these groves. Initial groves were grown where all independent variables were used. The strongest predictive variables based on Gini and standard measures were then selected for a final run (Table 15, Figures 33 and 34).

Table 15. The prediction summary of diagnostic features. The selected environmental covariates which were used for each diagnostic feature are indicated with an “X”.

Property	Argillic Horizon	Calcic Horizon	Durinodic Properties	Particle Size Family	Limiting Contact
Figure	37A	37B	37C	38	37D
OOB Error [%]	28	29.5	23.5	35.7	27.7
Weighted	Yes	Yes	Yes	Yes	Yes
Band 1	X				
Band 2	X	X			
Band 3	X	X	X	X	X
Band 4	X				
Band 5	X	X	X	X	X
Band 7	X	X	X	X	X
(4-3)/(4+3)	X	X	X	X	X
(3-7)/(3+7)	X	X		X	
(4-5)/(4+5)	X	X	X		
(5-2)/(5+2)		X	X	X	X
(5-1)/(5+1)		X			
(4-7)/(4+7)	X	X		X	X
CTI		X		X	X
CTI 5x5	X	X		X	
Slope	X	X	X	X	X
Slope 11x11	X	X	X	X	X
Curvature	X	X		X	X
Elevation	X	X	X	X	X
Lake	X		X	X	
Xeric SRM	X	X		X	
Aspect		X			

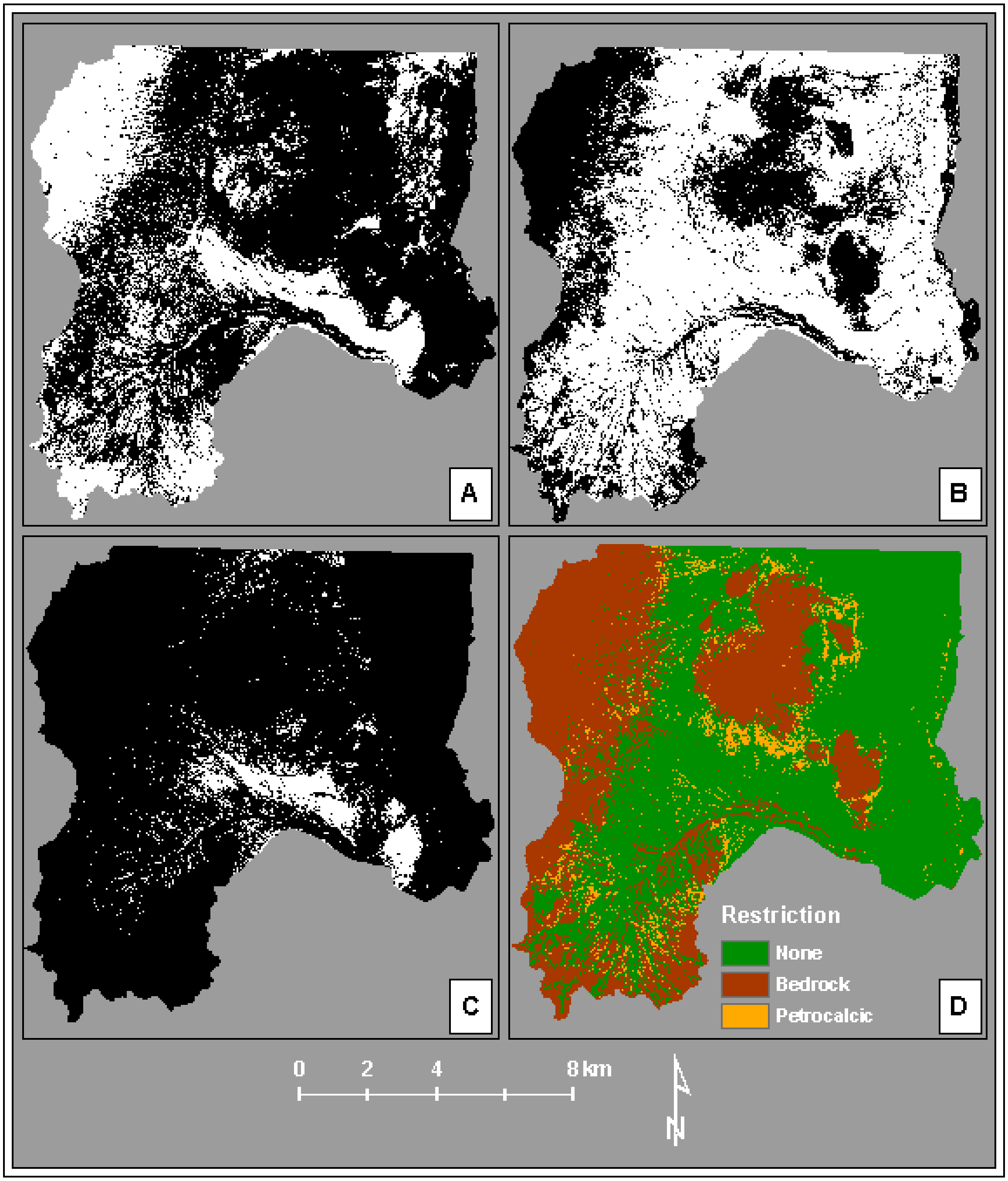


Figure 33. Prediction of diagnostic features. A: Presence of Argillic horizon. B: Presence of Calcic horizon. C: presence of Durinodic properties. D: Contact with bedrock or petrocalcic.

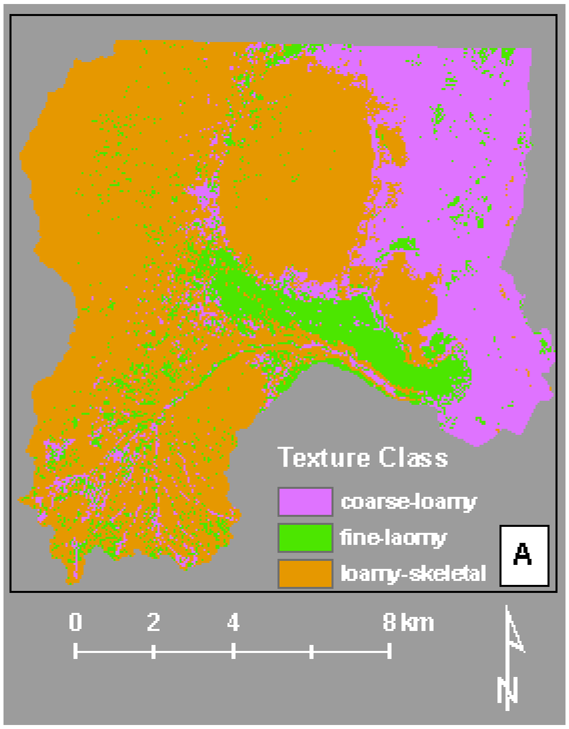


Figure 34. Prediction of the three most common particle size family classes (texture classes).

Appendix C: Spectral and Topographic Characteristics

In addition to the variable importance, the spectral and topographic characteristics were evaluated by class.

Table 16 shows the percentage of each class that was stratified into discrete pedogeomorphic units, e.g., below Lake Bonneville shoreline (“Below Shoreline) vs. above Lake Bonneville shoreline (not shown in Table 16), and Xeric SMR (“Pinyon”) vs. Aridic SMR wooded (“Juniper”) vs. Aridic SMR non-wooded (“Non Wooded”). For comparison, these percentages are given for the 671 observation points (“Sample”) and also for all pixels in the study area (“Area”) (Table 16).

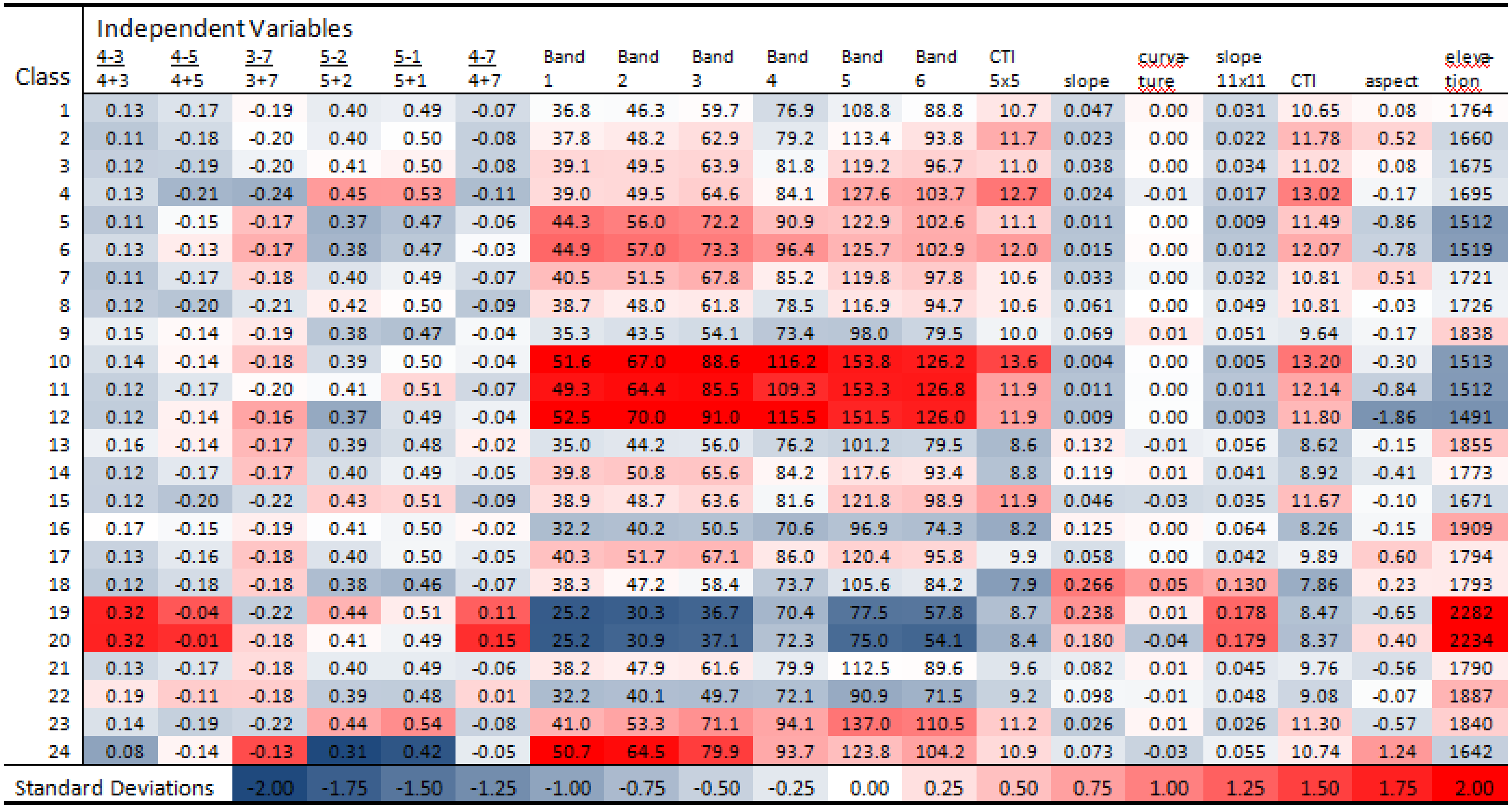
The unique characteristics of each class be evaluated from the mean values and standard deviations of the continuous environmental covariates. Table 17 displays the covariate mean values of all the 671 sample points. Table 18 displays the covariate mean values of the pixels classified by Grove 2B by class. Tables 17 and 18 were color coded by the standardized value of each covariate. Table 19 has the mean value and standard deviation for each continuous environmental covariate for all pixels in the entire study area.

Tables 17 and 18 clearly show that Class 24 (mine dumps), which was the best predicted class, stands out with the lowest average value in (5-2)/(5+2). The mine dump signature is the result of open pit mining in igneous rock which absorbs more electromagnetic radiation in the band 5 relative to band 2. Classes 19 (deep Haploxeralfs) and 20 (Haploxeralfs and Haploxerepts), which were observed in the San Francisco Mountains, stand out with high values of NDVI (a result of the greatest precipitation in the study area), elevation (highest mountain range in the study area), slope (900-m of relief), and very low values in all Landsat 7 bands (likely due to topography shading and vegetation). And classes 10, 11, and 12 have the highest reflectance values in all 6 Landsat bands, which is likely caused by the relatively high albedo of these typic aridic soils (driest soils) that have much lower vegetation cover than the other classes. Most classes, except for Classes 18, 19, and 20 which occurred in foothills and on mountains, had slopes lower to or equal to the mean, which was not particularly steep.

Table 16. The percentage of each soil class stratified into a categorical variable class for the 671 sample points ("Sample") and all pixels in the study area ("Area").

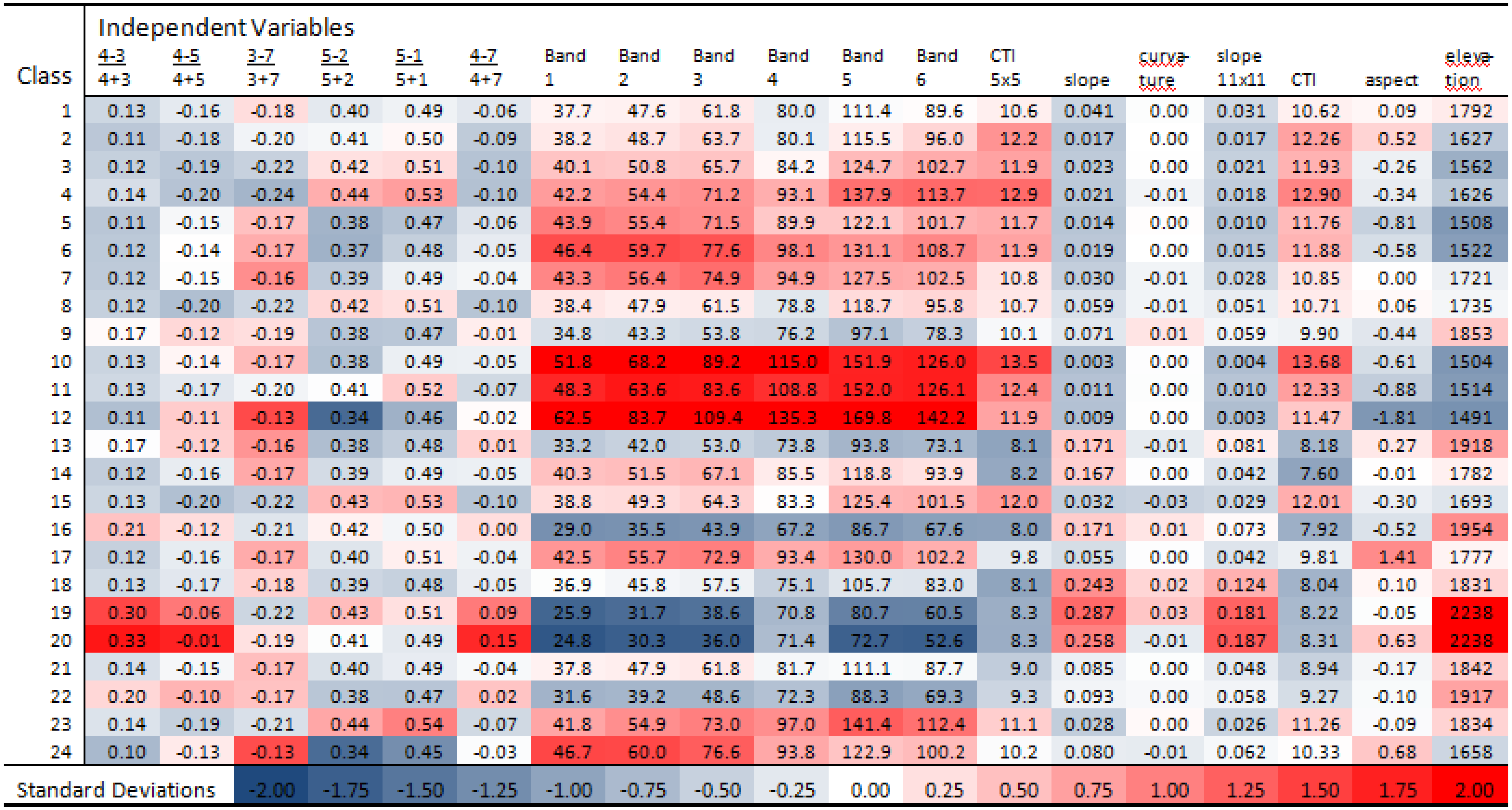
	Below Shoreline		Juniper		Pinyon		Non-Wooded	
Class	Sample	Area	Sample	Area	Sample	Area	Sample	Area
1	0.0	0.0	15.6	14.5	0.0	0.0	84.4	85.5
2	0.0	0.0	5.3	0.0	0.0	0.0	94.7	100.0
3	24.6	74.6	3.3	0.1	0.0	0.0	96.7	99.9
4	47.4	70.0	15.8	21.9	0.0	0.0	84.2	78.1
5	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0
6	100.0	99.0	0.0	0.0	0.0	0.0	100.0	100.0
7	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0
8	9.2	0.8	0.0	1.2	0.0	0.0	100.0	98.8
9	0.0	0.0	40.0	70.8	0.0	0.0	60.0	29.2
10	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0
11	100.0	99.7	0.0	0.0	0.0	0.0	100.0	100.0
12	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0
13	0.0	0.0	38.5	68.3	0.0	0.0	61.5	31.7
14	0.0	0.6	11.1	0.6	0.0	0.0	88.9	99.4
15	2.4	0.6	0.0	1.0	0.0	0.0	100.0	99.0
16	0.0	0.0	50.0	95.6	0.0	0.7	50.0	3.7
17	0.0	0.1	10.0	3.8	0.0	0.0	90.0	96.2
18	0.0	0.5	14.5	22.6	0.0	0.9	85.5	76.6
19	0.0	0.0	8.3	3.0	91.7	97.0	0.0	0.0
20	0.0	0.0	9.1	2.4	90.9	97.6	0.0	0.0
21	5.7	0.4	15.1	18.9	0.0	0.0	84.9	81.1
22	0.0	0.0	82.8	95.6	0.0	0.4	17.2	4.0
23	0.0	0.0	0.0	1.3	0.0	0.0	100.0	98.7
24	0.0	14.4	0.0	0.6	0.0	0.0	100.0	99.4

Table 17. The average value of the continuous environmental covariates for the sample data. Color coded according to the number standard deviations from the mean value of the entire study area.



Class	Independent Variables																		
	4-3 4+3	4-5 4+5	3-7 3+7	5-2 5+2	5-1 5+1	4-7 4+7	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	CTI 5x5	slope	curva- ture	slope 11x11	CTI	aspect	eleva- tion
1	0.13	-0.17	-0.19	0.40	0.49	-0.07	36.8	46.3	59.7	76.9	108.8	88.8	10.7	0.047	0.00	0.031	10.65	0.08	1764
2	0.11	-0.18	-0.20	0.40	0.50	-0.08	37.8	48.2	62.9	79.2	113.4	93.8	11.7	0.023	0.00	0.022	11.78	0.52	1660
3	0.12	-0.19	-0.20	0.41	0.50	-0.08	39.1	49.5	63.9	81.8	119.2	96.7	11.0	0.038	0.00	0.034	11.02	0.08	1675
4	0.13	-0.21	-0.24	0.45	0.53	-0.11	39.0	49.5	64.6	84.1	127.6	103.7	12.7	0.024	-0.01	0.017	13.02	-0.17	1695
5	0.11	-0.15	-0.17	0.37	0.47	-0.06	44.3	56.0	72.2	90.9	122.9	102.6	11.1	0.011	0.00	0.009	11.49	-0.86	1512
6	0.13	-0.13	-0.17	0.38	0.47	-0.03	44.9	57.0	73.3	96.4	125.7	102.9	12.0	0.015	0.00	0.012	12.07	-0.78	1519
7	0.11	-0.17	-0.18	0.40	0.49	-0.07	40.5	51.5	67.8	85.2	119.8	97.8	10.6	0.033	0.00	0.032	10.81	0.51	1721
8	0.12	-0.20	-0.21	0.42	0.50	-0.09	38.7	48.0	61.8	78.5	116.9	94.7	10.6	0.061	0.00	0.049	10.81	-0.03	1726
9	0.15	-0.14	-0.19	0.38	0.47	-0.04	35.3	43.5	54.1	73.4	98.0	79.5	10.0	0.069	0.01	0.051	9.64	-0.17	1838
10	0.14	-0.14	-0.18	0.39	0.50	-0.04	51.6	67.0	88.6	116.2	153.8	126.2	13.6	0.004	0.00	0.005	13.20	-0.30	1513
11	0.12	-0.17	-0.20	0.41	0.51	-0.07	49.3	64.4	85.5	109.3	153.3	126.8	11.9	0.011	0.00	0.011	12.14	-0.84	1512
12	0.12	-0.14	-0.16	0.37	0.49	-0.04	52.5	70.0	91.0	115.5	151.5	126.0	11.9	0.009	0.00	0.003	11.80	-1.86	1491
13	0.16	-0.14	-0.17	0.39	0.48	-0.02	35.0	44.2	56.0	76.2	101.2	79.5	8.6	0.132	-0.01	0.056	8.62	-0.15	1855
14	0.12	-0.17	-0.17	0.40	0.49	-0.05	39.8	50.8	65.6	84.2	117.6	93.4	8.8	0.119	0.01	0.041	8.92	-0.41	1773
15	0.12	-0.20	-0.22	0.43	0.51	-0.09	38.9	48.7	63.6	81.6	121.8	98.9	11.9	0.046	-0.03	0.035	11.67	-0.10	1671
16	0.17	-0.15	-0.19	0.41	0.50	-0.02	32.2	40.2	50.5	70.6	96.9	74.3	8.2	0.125	0.00	0.064	8.26	-0.15	1909
17	0.13	-0.16	-0.18	0.40	0.50	-0.05	40.3	51.7	67.1	86.0	120.4	95.8	9.9	0.058	0.00	0.042	9.89	0.60	1794
18	0.12	-0.18	-0.18	0.38	0.46	-0.07	38.3	47.2	58.4	73.7	105.6	84.2	7.9	0.266	0.05	0.130	7.86	0.23	1793
19	0.32	-0.04	-0.22	0.44	0.51	0.11	25.2	30.3	36.7	70.4	77.5	57.8	8.7	0.238	0.01	0.178	8.47	-0.65	2282
20	0.32	-0.01	-0.18	0.41	0.49	0.15	25.2	30.9	37.1	72.3	75.0	54.1	8.4	0.180	-0.04	0.179	8.37	0.40	2234
21	0.13	-0.17	-0.18	0.40	0.49	-0.06	38.2	47.9	61.6	79.9	112.5	89.6	9.6	0.082	0.01	0.045	9.76	-0.56	1790
22	0.19	-0.11	-0.18	0.39	0.48	0.01	32.2	40.1	49.7	72.1	90.9	71.5	9.2	0.098	-0.01	0.048	9.08	-0.07	1887
23	0.14	-0.19	-0.22	0.44	0.54	-0.08	41.0	53.3	71.1	94.1	137.0	110.5	11.2	0.026	0.01	0.026	11.30	-0.57	1840
24	0.08	-0.14	-0.13	0.31	0.42	-0.05	50.7	64.5	79.9	93.7	123.8	104.2	10.9	0.073	-0.03	0.055	10.74	1.24	1642
Standard Deviations			-2.00	-1.75	-1.50	-1.25	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00

Table 18. The average value of the continuous environmental covariates of each pixel according to their predicted class. Color coded according to the number standard deviations from the mean value of the entire study area.



Class	Independent Variables																		
	$\frac{4-3}{4+3}$	$\frac{4-5}{4+5}$	$\frac{3-7}{3+7}$	$\frac{5-2}{5+2}$	$\frac{5-1}{5+1}$	$\frac{4-7}{4+7}$	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	CTI 5x5	slope	curvature	slope 11x11	CTI	aspect	elevation
1	0.13	-0.16	-0.18	0.40	0.49	-0.06	37.7	47.6	61.8	80.0	111.4	89.6	10.6	0.041	0.00	0.031	10.62	0.09	1792
2	0.11	-0.18	-0.20	0.41	0.50	-0.09	38.2	48.7	63.7	80.1	115.5	96.0	12.2	0.017	0.00	0.017	12.26	0.52	1627
3	0.12	-0.19	-0.22	0.42	0.51	-0.10	40.1	50.8	65.7	84.2	124.7	102.7	11.9	0.023	0.00	0.021	11.93	-0.26	1562
4	0.14	-0.20	-0.24	0.44	0.53	-0.10	42.2	54.4	71.2	93.1	137.9	113.7	12.9	0.021	-0.01	0.018	12.90	-0.34	1626
5	0.11	-0.15	-0.17	0.38	0.47	-0.06	43.9	55.4	71.5	89.9	122.1	101.7	11.7	0.014	0.00	0.010	11.76	-0.81	1508
6	0.12	-0.14	-0.17	0.37	0.48	-0.05	46.4	59.7	77.6	98.1	131.1	108.7	11.9	0.019	0.00	0.015	11.88	-0.58	1522
7	0.12	-0.15	-0.16	0.39	0.49	-0.04	43.3	56.4	74.9	94.9	127.5	102.5	10.8	0.030	-0.01	0.028	10.85	0.00	1721
8	0.12	-0.20	-0.22	0.42	0.51	-0.10	38.4	47.9	61.5	78.8	118.7	95.8	10.7	0.059	-0.01	0.051	10.71	0.06	1735
9	0.17	-0.12	-0.19	0.38	0.47	-0.01	34.8	43.3	53.8	76.2	97.1	78.3	10.1	0.071	0.01	0.059	9.90	-0.44	1853
10	0.13	-0.14	-0.17	0.38	0.49	-0.05	51.8	68.2	89.2	115.0	151.9	126.0	13.5	0.003	0.00	0.004	13.68	-0.61	1504
11	0.13	-0.17	-0.20	0.41	0.52	-0.07	48.3	63.6	83.6	108.8	152.0	126.1	12.4	0.011	0.00	0.010	12.33	-0.88	1514
12	0.11	-0.11	-0.13	0.34	0.46	-0.02	62.5	83.7	109.4	135.3	169.8	142.2	11.9	0.009	0.00	0.003	11.47	-1.81	1491
13	0.17	-0.12	-0.16	0.38	0.48	0.01	33.2	42.0	53.0	73.8	93.8	73.1	8.1	0.171	-0.01	0.081	8.18	0.27	1918
14	0.12	-0.16	-0.17	0.39	0.49	-0.05	40.3	51.5	67.1	85.5	118.8	93.9	8.2	0.167	0.00	0.042	7.60	-0.01	1782
15	0.13	-0.20	-0.22	0.43	0.53	-0.10	38.8	49.3	64.3	83.3	125.4	101.5	12.0	0.032	-0.03	0.029	12.01	-0.30	1693
16	0.21	-0.12	-0.21	0.42	0.50	0.00	29.0	35.5	43.9	67.2	86.7	67.6	8.0	0.171	0.01	0.073	7.92	-0.52	1954
17	0.12	-0.16	-0.17	0.40	0.51	-0.04	42.5	55.7	72.9	93.4	130.0	102.2	9.8	0.055	0.00	0.042	9.81	1.41	1777
18	0.13	-0.17	-0.18	0.39	0.48	-0.05	36.9	45.8	57.5	75.1	105.7	83.0	8.1	0.243	0.02	0.124	8.04	0.10	1831
19	0.30	-0.06	-0.22	0.43	0.51	0.09	25.9	31.7	38.6	70.8	80.7	60.5	8.3	0.287	0.03	0.181	8.22	-0.05	2238
20	0.33	-0.01	-0.19	0.41	0.49	0.15	24.8	30.3	36.0	71.4	72.7	52.6	8.3	0.258	-0.01	0.187	8.31	0.63	2238
21	0.14	-0.15	-0.17	0.40	0.49	-0.04	37.8	47.9	61.8	81.7	111.1	87.7	9.0	0.085	0.00	0.048	8.94	-0.17	1842
22	0.20	-0.10	-0.17	0.38	0.47	0.02	31.6	39.2	48.6	72.3	88.3	69.3	9.3	0.093	0.00	0.058	9.27	-0.10	1917
23	0.14	-0.19	-0.21	0.44	0.54	-0.07	41.8	54.9	73.0	97.0	141.4	112.4	11.1	0.028	0.00	0.026	11.26	-0.09	1834
24	0.10	-0.13	-0.13	0.34	0.45	-0.03	46.7	60.0	76.6	93.8	122.9	100.2	10.2	0.080	-0.01	0.062	10.33	0.68	1658
Standard Deviations			-2.00	-1.75	-1.50	-1.25	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00

Table 19. The *Mean* and Standard Deviation of each continuous environmental covariate for all pixels in the study area.

Variable	**Mean**	**Standard Deviation**
NDVI	0.17	0.09
(4-5)/(4+5)	-0.15	0.08
(3-7)/(3+7)	-0.20	0.05
(5-2)/(5+2)	0.41	0.04
(5-1)/(5+1)	0.50	0.04
(4-7)/(4+7)	-0.03	0.11
band 1	37.04	6.63
band 2	46.62	9.29
band 3	59.41	13.46
band 4	82.98	15.79
band 5	112.16	22.87
band 6	89.05	20.83
CTI 5x5	10.3	2.2
slope	0.110	0.150
curvature	0.000	0.065
slope 11x11	0.068	0.091
CTI	10.3	2.7
transformed aspect	-0.020	1.57
elevation	1756	221

Appendix D: Natural Break - Overlay

The 1-m aerial photography has a higher resolution than the 30-m output from the RF classification. While much more precise than the 30-m and 10-m resolutions of the Landsat and DEM data layers, it lacks information (dimension) to predict soil classes. However, changes in tone in the aerial photography often demark meaningful breaks in the landscape. To make detailed class boundaries, aerial photography was segmented into five classes with unsupervised classification. The purpose of performing an unsupervised classification of the aerial photography was not to predict soil class but to segment the image into natural breaks in the landscape such as changes in the vegetation.

A smaller subset of the study area was chosen as this process required days of computation (Figure 35). The northeast corner of the study area was selected for the natural break overlay process as there were high contrasts and distinct vegetation-landform-soil breaks discernable in the aerial photography.

Below is an outline of the steps and software involved for this process.

- Aerial photography (1-m resolution)
    - Filtered aerial photography twice with a 3x3 low pass filter with Imagine (Figure 36A)
    - Unsupervised classification (5 classes) of aerial photography with Imagine (Figure 36B)
    - Clumped (8 neighbor) and eliminated all clusters of pixels less than 45 pixels (Imagine)
    - Vectorized the unsupervised classification with ArcGIS (Figure 37C)

- Overlay Random Forests classification (30-m resolution) over unsupervised classification (Figure 38)
    - Zonal attributes: majority (Imagine)
    - Dissolved neighboring polygons with the same Random Forests majority class (ArcGIS)
    - Eliminated all polygons <900 m^2 (ran twice to eliminate polygons within eliminated polygons) with ArcGIS (Figure 37D)

The aerial photography was filtered twice with a 3x3 low pass. An unsupervised classification was run where five classes were predicted (Figure 36B). The image was clumped, where contiguous pixels (8 neighbors) of the same class became a group of pixels. All clumps with <45 pixels were eliminated (assimilated by a majority filter) before being vectorized to reduce the number of potential polygons (if not the process may take several day of computation, see in Figure 37C). The resulting classified layer was vectorized. The final soil classification from the Random Forests model 2B was re-sampled to 1-meter resolution. The zonal statistics utility in Imagine (GIS Analysis) was used to determine the majority of pixels within each polygon (Figures 37D and 38B).

The natural breaks process using the aerial photography (Figure 36A) and the output of Grove 2B in the northeastern portion of the study area (Figure 38A) yielded fairly detailed boundaries, following the vegetation-landform breaks. However, some areas became too generalized. Areas with little contrast in the aerial photography failed to segment sufficiently, thus generalizing the RF output. For example, in an area that had both class 10 (Thermosprings) and class 11 (Typic Calciargids), RF identified a large areas of Thermosprings (class 10) (Figure 38B). These areas had little vegetation and high visible reflectance, so there was little contrast (variance) in the photography. When the unsupervised classification of the aerial photography was attributed with the majority soil class, many areas identified as class 10 (Thermosprings) disappeared and were now classified areas of class 11 (Typic Calciargids) (Figure 37D). The purpose of the natural breaks process was not to reclassify large areas but to define a more precise soil boundary. This issue, the unsupervised classification leading to the reclassification of large areas, may be addressed by splitting unsupervised classes of the aerial photography into an additional class.

The natural break overlay process was only applied to subsets of the study area as this process was computationally intense. This process was tried on fan remnants, but there was not enough variability in the aerial photography due to the homogenous vegetation pattern to produce a useful map. Many of the resulting breaks were from roads and other anthropogenic features while other topographic landform breaks were not pulled out.

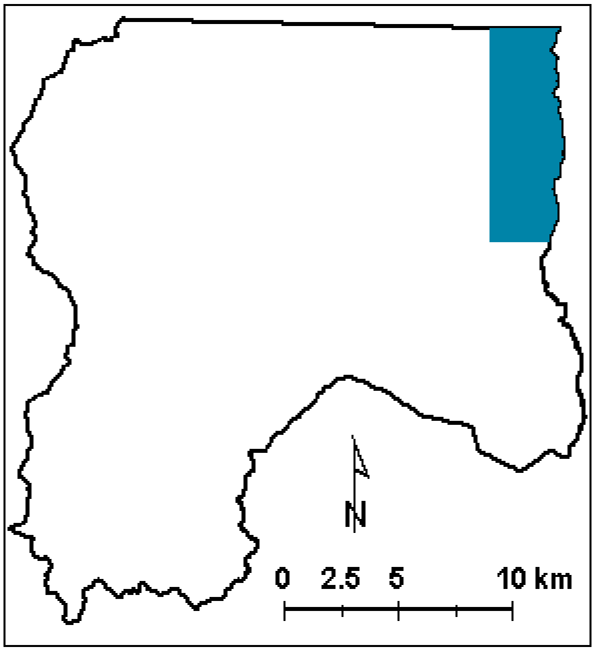


Figure 35. The area of interest selected for the natural breaks overlay process.

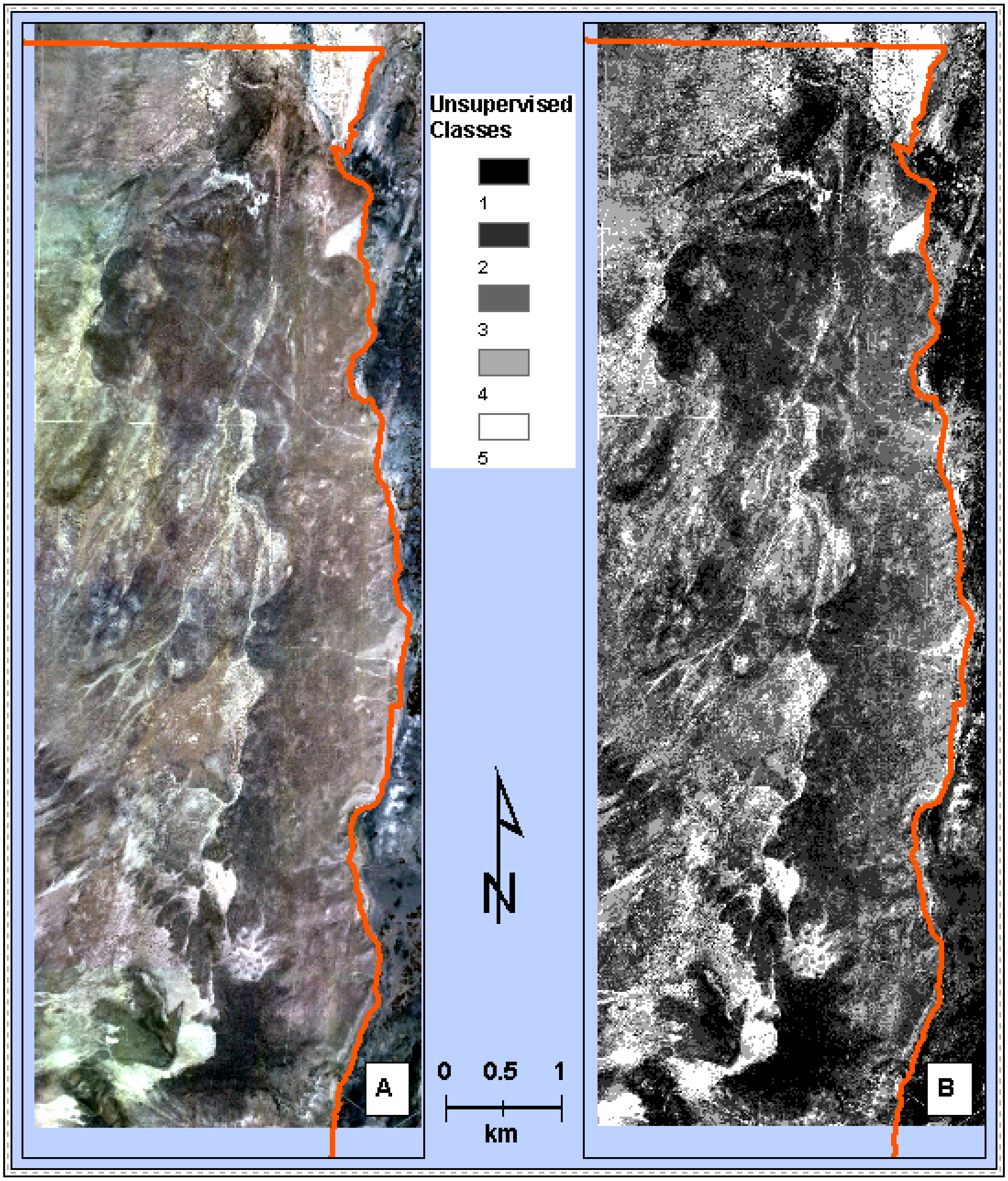


Figure 36. Aerial photography. A: The color aerial photography was subset and generalized with a 3x3 low pass filter. B: The filtered image to the left (A) was segmented into five unsupervised classes.

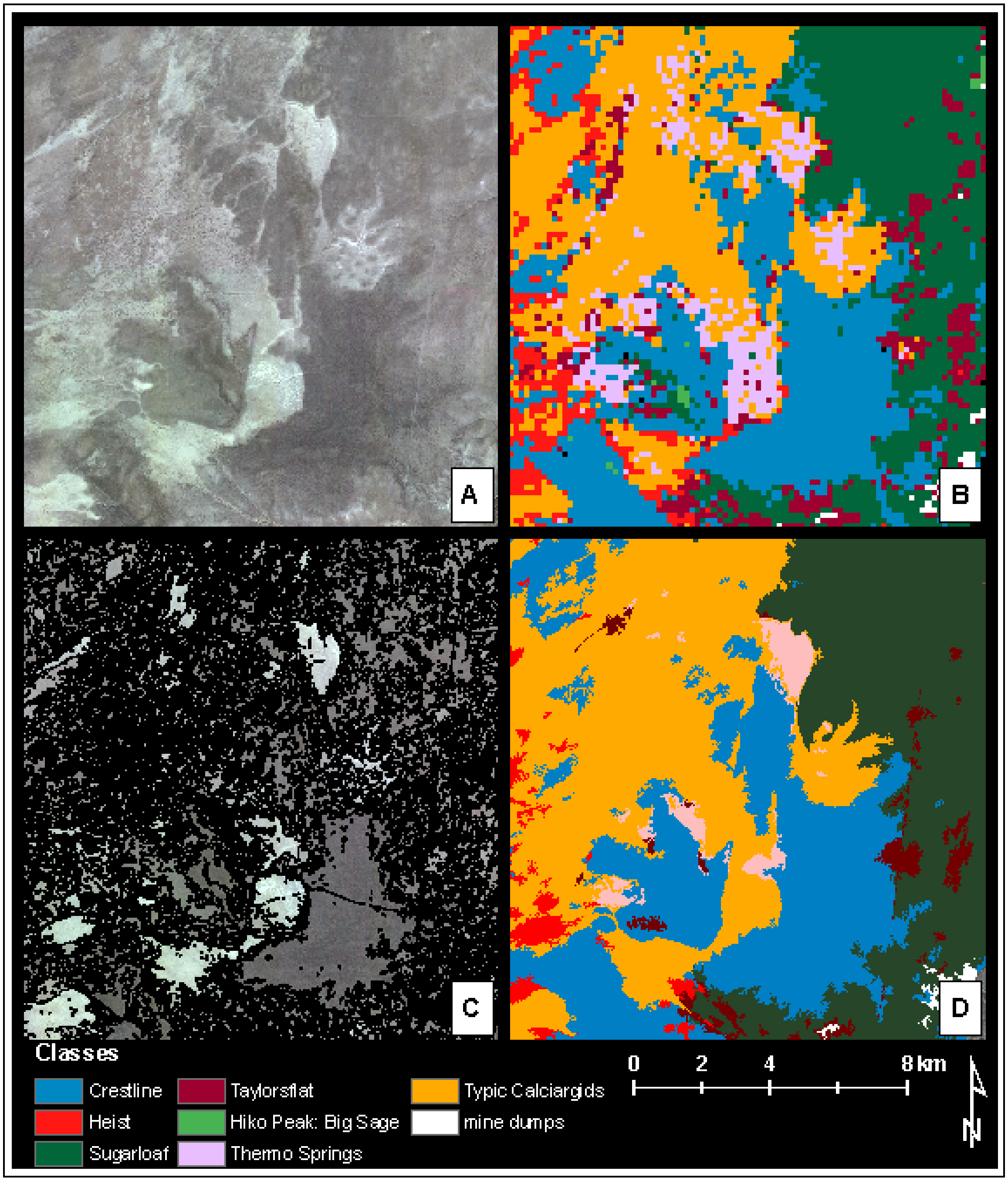


Figure 37. Results of the natural breaks exercise. A: 1-m Aerial photography that. B: The Random Forests output of Grove 2B (30-m resolution). C: The vectorized unsupervised classification of the aerial photography with clumps ≤45 pixels eliminated. D: The final result when the polygons in the vector layer (C) were attributed with the majority class from Grove 2B (B).

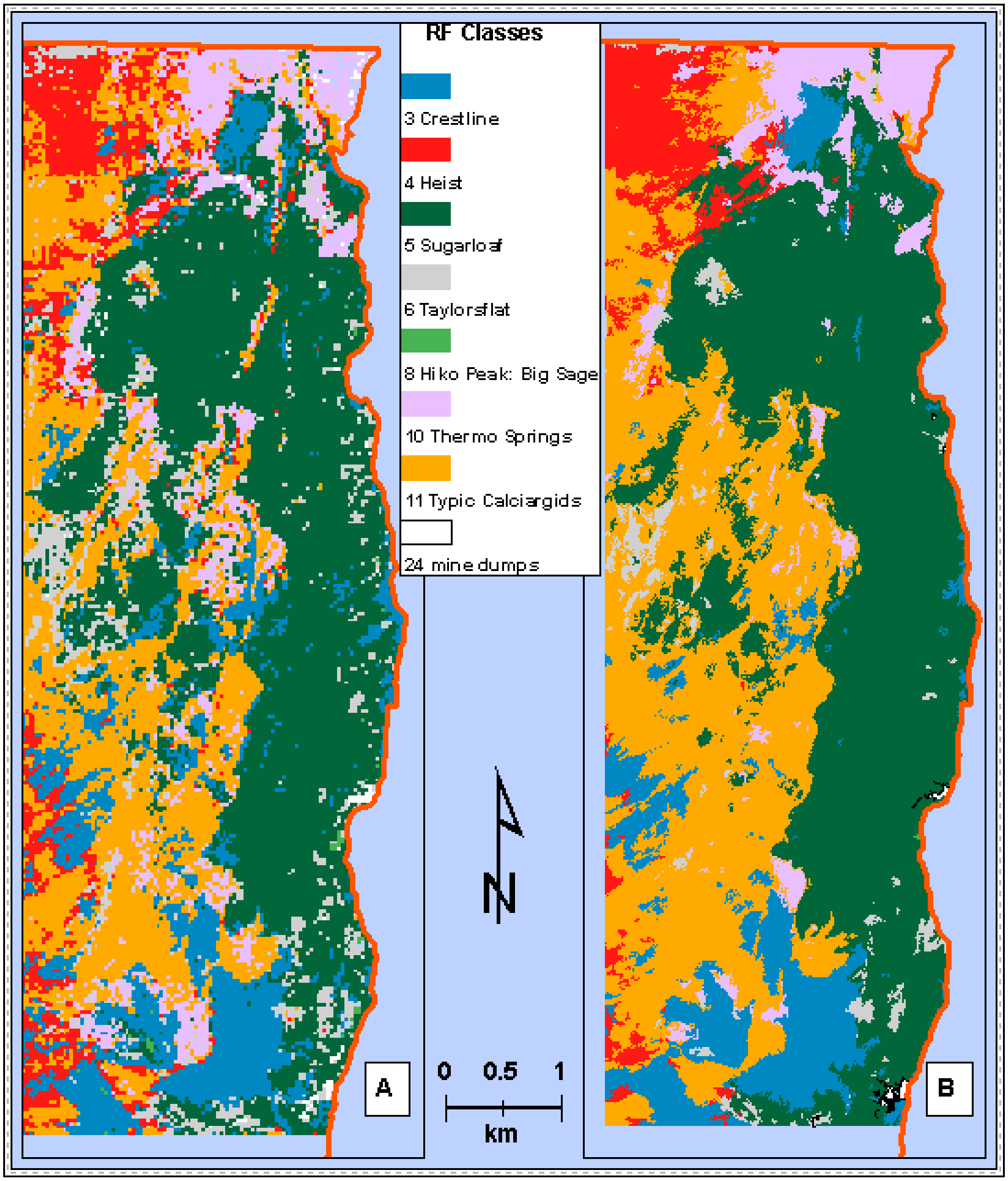


Figure 38. Comparison of the Random Forest classification (A) to Random Forests laid over the unsupervised classification (B). The final Random Forests classification (Grove 2B) had a 30-m resolution appears pixilated (A). The RF output at left (A) was laid over the unsupervised classification (B). All polygons less than 900 m^2 were removed (the area one 30-m pixel).

Appendix E: OOB Analysis

When Grove 2B was grown, optional reports with information about the OOB performance of each observation were not saved. A new grove which had the same parameters as Grove 2B was grown. The overall OOB error of this grove was 54.2% compared to 55.2% of the original Grove 2B. The OOB error by class is expected to be likewise similar to the original iteration of Grove 2B.

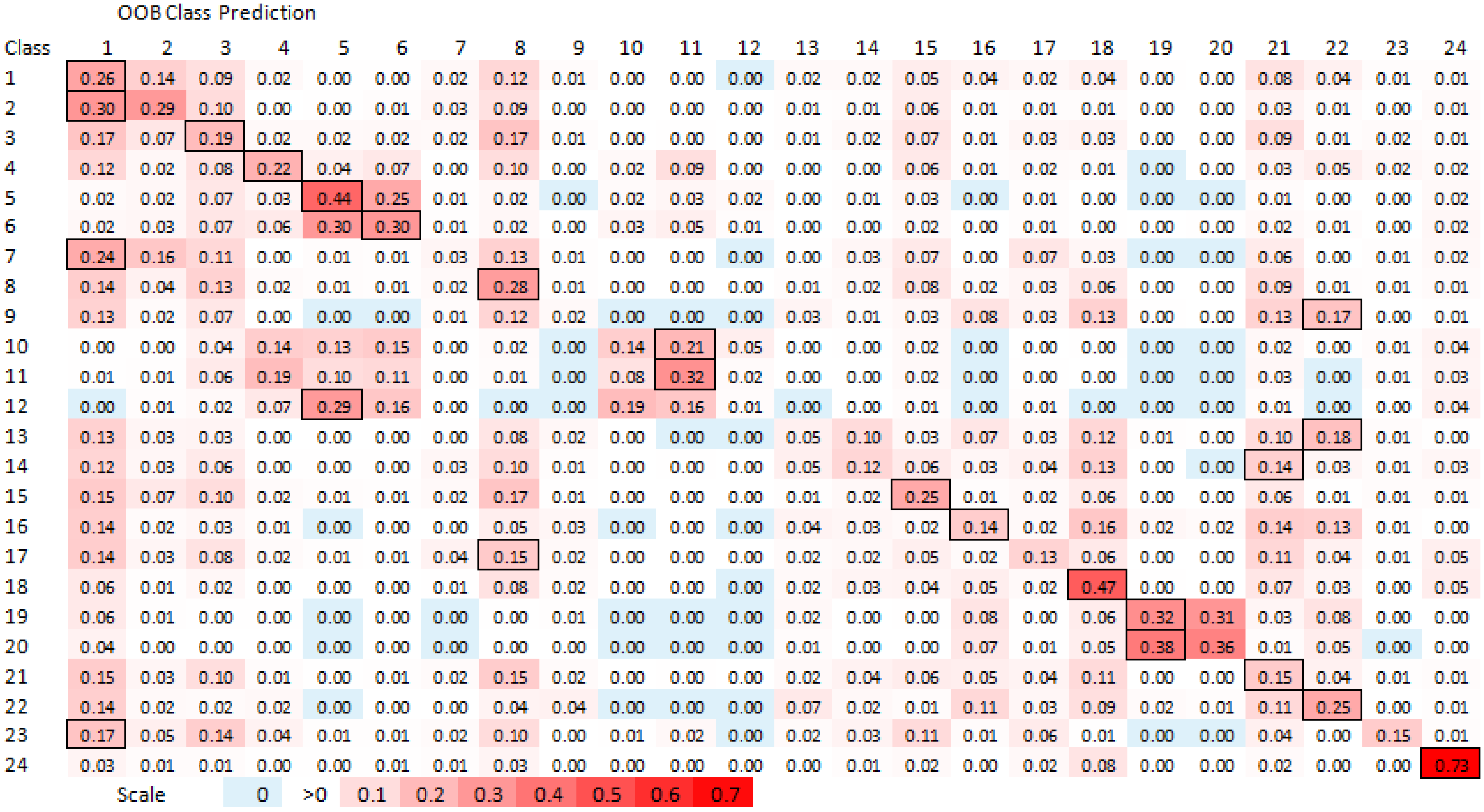
Table 20 is similar to the error matrix (Table 12) but is the proportion of all votes each individual observation received when left out of the bag. The error matrix is the result of all the times (mode) an observation was left out of the bag. Table 20 is a summary of all votes (predictions by each tree when OOB), e.g. popular vote; while the error matrix is a summary of the mode all trees for an observation (majority vote of all trees); e.g. an electoral college, where each observation is an elector.

Some interesting patterns related to soil catena were revealed in Table 20 that were not apparent in Table 12. For example, classes 10, 11, and 12 occurred together on the basin floors below the Lake Bonneville shoreline. These soils are morphologically similar, and support similar plant communities. Classes 10, 11, and 12 were frequently predicted to be classes 4, 5, and 6, which also occurred below the Bonneville shoreline. Classes 4 and 6 occurred on alluvial flats and basin floors and are more similar to each other than class 5, which differed the most morphologically, and was observed on a broad stream terrace. All of these soils were found on adjacent landforms, had low slope and lower vegetation density than the rest of the study area.

Table 21 summarizes of the OOB results by individual sample observations, and shows that the correct class was the first most predicted class for 46% of observations when left out of the bag (place 1), which is expected given that the overall OOB error of this grove was 54%. The correct class was the second most predicted class for 18% of observations when OOB (place 2); the correct class was the third most predicted class for 8% of observations when out of bag (place 3); etc.

For some observations when left out of the bag, the actual class is the 10th or 19th most predicted class. Looking through a table (not shown) of the 671 observations performance OOB, #216 stood out. It was observed to be Garbo (class 2), but of the 168 times it was OOB it was never predicted to be class 2. This observation was predicted to be 11 classes other than class 2. A review of the description of this observation reveals a poor class correlation. Garbo soils are typically found on fan remnants and support Wyoming Big Sage communities. Observation 216 was located in the hills, under a stand of Utah Juniper and had a possible duripan, among other differences. The soil had an argillic horizon, a fine-loamy particle size family class, and durinodic properties, therefore it was classified as Garbo, class 2, despite the differences in landform, soils morphology and ecological site. Analyzing the OOB performance of observations is useful for identifying outliers or possibly misclassified observations. It can also reveal linkages between classes and ambiguity between classes.

Table 20. The distribution of all OOB votes summarized by class. The class that the observations were most frequently predicted as are outlined. Cells that have a value of 0 and not highlighted blue have a value between 0 and 1%.



	OOB Class Prediction																							
Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0.26	0.14	0.09	0.02	0.00	0.00	0.02	0.12	0.01	0.00	0.00	0.00	0.02	0.02	0.05	0.04	0.02	0.04	0.00	0.00	0.08	0.04	0.01	0.01
2	0.30	0.29	0.10	0.00	0.00	0.01	0.03	0.09	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.01	0.01	0.01	0.00	0.00	0.03	0.01	0.00	0.01
3	0.17	0.07	0.19	0.02	0.02	0.02	0.02	0.17	0.01	0.00	0.00	0.00	0.01	0.02	0.07	0.01	0.03	0.03	0.00	0.00	0.09	0.01	0.02	0.01
4	0.12	0.02	0.08	0.22	0.04	0.07	0.00	0.10	0.00	0.02	0.09	0.00	0.00	0.00	0.06	0.01	0.02	0.01	0.00	0.00	0.03	0.05	0.02	0.02
5	0.02	0.02	0.07	0.03	0.44	0.25	0.01	0.02	0.00	0.02	0.03	0.02	0.00	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.02
6	0.02	0.03	0.07	0.06	0.30	0.30	0.01	0.02	0.00	0.03	0.05	0.01	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.02
7	0.24	0.16	0.11	0.00	0.01	0.01	0.03	0.13	0.01	0.00	0.00	0.00	0.00	0.03	0.07	0.00	0.07	0.03	0.00	0.00	0.06	0.00	0.01	0.02
8	0.14	0.04	0.13	0.02	0.01	0.01	0.02	0.28	0.01	0.00	0.00	0.00	0.01	0.02	0.08	0.02	0.03	0.06	0.00	0.00	0.09	0.01	0.01	0.01
9	0.13	0.02	0.07	0.00	0.00	0.00	0.01	0.12	0.02	0.00	0.00	0.00	0.03	0.01	0.03	0.08	0.03	0.13	0.00	0.00	0.13	0.17	0.00	0.01
10	0.00	0.00	0.04	0.14	0.13	0.15	0.00	0.02	0.00	0.14	0.21	0.05	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.04
11	0.01	0.01	0.06	0.19	0.10	0.11	0.00	0.01	0.00	0.08	0.32	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.03
12	0.00	0.01	0.02	0.07	0.29	0.16	0.00	0.00	0.00	0.19	0.16	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.04
13	0.13	0.03	0.03	0.00	0.00	0.00	0.00	0.08	0.02	0.00	0.00	0.00	0.05	0.10	0.03	0.07	0.03	0.12	0.01	0.00	0.10	0.18	0.01	0.00
14	0.12	0.03	0.06	0.00	0.00	0.00	0.03	0.10	0.01	0.00	0.00	0.00	0.05	0.12	0.06	0.03	0.04	0.13	0.00	0.00	0.14	0.03	0.01	0.03
15	0.15	0.07	0.10	0.02	0.01	0.01	0.02	0.17	0.01	0.00	0.00	0.00	0.01	0.02	0.25	0.01	0.02	0.06	0.00	0.00	0.06	0.01	0.01	0.01
16	0.14	0.02	0.03	0.01	0.00	0.00	0.00	0.05	0.03	0.00	0.00	0.00	0.04	0.03	0.02	0.14	0.02	0.16	0.02	0.02	0.14	0.13	0.01	0.00
17	0.14	0.03	0.08	0.02	0.01	0.01	0.04	0.15	0.02	0.00	0.00	0.00	0.02	0.02	0.05	0.02	0.13	0.06	0.00	0.00	0.11	0.04	0.01	0.05
18	0.06	0.01	0.02	0.00	0.00	0.00	0.01	0.08	0.02	0.00	0.00	0.00	0.02	0.03	0.04	0.05	0.02	0.47	0.00	0.00	0.07	0.03	0.00	0.05
19	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.08	0.00	0.06	0.32	0.31	0.03	0.08	0.00	0.00
20	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.07	0.01	0.05	0.38	0.36	0.01	0.05	0.00	0.00
21	0.15	0.03	0.10	0.01	0.00	0.01	0.02	0.15	0.02	0.00	0.00	0.00	0.02	0.04	0.06	0.05	0.04	0.11	0.00	0.00	0.15	0.04	0.01	0.01
22	0.14	0.02	0.02	0.02	0.00	0.00	0.00	0.04	0.04	0.00	0.00	0.00	0.07	0.02	0.01	0.11	0.03	0.09	0.02	0.01	0.11	0.25	0.00	0.01
23	0.17	0.05	0.14	0.04	0.01	0.01	0.02	0.10	0.00	0.01	0.02	0.00	0.02	0.03	0.11	0.01	0.06	0.01	0.00	0.00	0.04	0.00	0.15	0.01
24	0.03	0.01	0.01	0.00	0.00	0.01	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.02	0.08	0.00	0.00	0.02	0.00	0.00	0.73

Scale 0 >0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

Table 21. Summary of the OOB performance of individual observations.

Place	Number of Observations	Proportion of Observations
1	308	0.46
2	120	0.18
3	57	0.08
4	37	0.06
5	36	0.05
6	29	0.04
7	23	0.03
8	11	0.02
9	12	0.02
10	10	0.01
11	4	0.01
12	9	0.01
13	6	0.01
14	2	0.00
15	4	0.01
16	2	0.00
17	1	0.00
18	0	0.00
19	1	0.00
20	0	0.00
21	0	0.00
22	0	0.00
23	0	0.00
24	0	0.00