Semi-Automated Disaggregation of a Conventional Soil Map Using Knowledge Driven Data Mining and Random Forests in the Sonoran Desert, USA

Travis W. Nauman, James A. Thompson, and Craig Rasmussen

Abstract
Conventional soil maps (CSM) have provided baseline soil information for land use planning for over 100 years. Although CSM have been widely used, they are not suitable to meet growing demands for high resolution soil information at field scales. We present a repeatable method to disaggregate CSM data into ~30-meter resolution rasterized soil class maps that include continuous representation of probabilistic map uncertainty. Methods include training set creation for original CSM component soil classes from soil-landscape descriptions within the original survey database. Training sets are used to build a random forest predictive model utilizing environmental covariate maps derived from ASTER satellite imagery and the USGS National Elevation Dataset. Results showed agreement at 70 percent of independent field validation sites and equivalent accuracy between original CSM map units and the disaggregated map. Uncertainty of predictions was mapped by relating prediction frequencies of the random forest model and success rates at validation sites.

Introduction
The increased availability of both digital elevation data and remote sensing data have prompted many studies that use these data to improve soil property prediction and inventory in a field that has been coined “digital soil mapping” (DSM) (Grunwald et al., 2011; Grunwald, 2009; McBratney et al., 2003; Scull et al., 2003). Many of these studies use elevation data and remotely sensed imagery to represent one or more soil forming factors that include climate, organisms, relief, parent material, and time (Jenny, 1941). In this form, soil classes or properties are predicted from topographic or spectral indices derived from elevation and imagery.

Soil properties and functions influence many societal challenges particularly the response of ecosystem services such as carbon and nutrient cycling; water storage, purification and cycling; pollutant transport; and vegetation growth to climate change (Brady and Weil, 2008). However, our knowledge of soils is imprecise as demonstrated by estimates of global soil carbon stocks in the top meter of soil that range from 1,400 to 3,250 petagrams (Grunwald et al., 2011). In light of the projected challenges that climate change presents to ecosystem services (IPCC, 2007), high quality soil information is central to natural resource management and land use planning. Although many soil inventories in the form of CSM have been carried out around the world, the scope and coarse spatial resolution of many soil databases have been criticized as limitations to effective incorporation of soil information into models of ecosystem services and other earth surface processes (Burrough, 1989; Burrough et al., 1997; Grunwald, 2009; Grunwald et al., 2011; McBratney et al., 2003). The field of DSM has responded to this challenge with concerted efforts to quantitatively improve CSM soil information using a wide array of statistical, spatial, and information technology (Behrens et al., 2005; Bui et al., 2009; Bui et al., 2006; Bui et al., 1999; Bui and Moran, 2001; Cook et al., 1996a; Cook et al., 1996b; de Bruin et al., 1999; Häring et al., 2012; Kempen et al., 2009; Kerry et al., 2012; McBratney, 1998; Minasny and McBratney, 2010; Nauman and Thompson, 2014; Nauman et al., 2012; Thompson et al., 2010; Yang et al., 2011; Zhu, 1997; Zhu et al., 1997, 2001).

One of the main challenges to improving CSM data representation is that the original intent of CSM was management oriented, and properties attributed to soils were often estimates based on sparse data at representative locations and not quantifications based on rigorous statistical sampling and interpolation (USDA-NRCS, 2013). A large part of the goals of the original design of CSM was to provide somewhat qualitative interpretations intended to provide pragmatic initial guidance to developers, farmers, and other land management institutions (Soil Survey Staff, 1993). However, many current users of soil information, particularly those not familiar with CSM history and evolution, have attempted to use CSM data beyond their original purposes leading to the potential for spurious relationships and possible incorrect data and process interpretation.

Various models and analyses have been developed using spatial soil information from CSM (e.g., Gatzke et al., 2011; Lineback Gritzner et al., 2001; Thomas-Van Gundy et al., 2012; Thomas-Van Gundy and Strager, 2012). In the US, both the US General Soil Map (STATSGO2) and the Soil Survey...
Spatial disaggregation of CSM has been demonstrated in attempts to recreate soil maps without the original polygons, which we call universal soil map updates (Bui and Moran, 2001; Thompson et al., 2010a; Behrens et al., 2010b; Bui, 2004; Bui et al., 1999; Bui and Moran, 2001; de Bruin et al., 1999; McGratney, 1998; Wieland et al., 2001). Generally, approaches use new pedon data and/or environmental covariate data in a DSM framework to determine how soils within polygon map units vary spatially.

Spatial disaggregation of CSM has been demonstrated in attempts to recreate soil maps without the original polygons, which we call universal soil map updates (Bui and Moran, 2001; Hansen et al., 2009; Moran and Bui, 2002; Nauman and Thompson, 2014; Smith et al., 2012; Wei et al., 2010; Yang et al., 2011). Others have updated CSM within the bounds of original survey map units (Bui and Moran, 2001; Thompson et al., 2010; Häring et al., 2012). Other studies have looked at disaggregating CSM for specific soil properties (Goovaerts, 2011; Kerry et al., 2012; Nauman et al., 2010). Goovaerts (2011) evaluated geostatistical methods that can combine point data with choropleth data to look at intra-polygon variation of a specific variable, and Kerry et al. (2012) applied parts of these methods to soil organic carbon mapping in northern Ireland. Fuzzy logic has been used in disaggregation through applications like SoILM (Qi et al., 2006; Zhu, 1997; Zhu et al., 1996) to help organize and implement soil-landscape relationships for mapping soils. SoILM has been used in coordination with both expert knowledge (Smith et al., 2010) and statistical approaches (Yang et al., 2011) to implement discovered soil-landscape relationships for updating and disaggregating soil maps. Other fuzzy knowledge systems have leveraged landform element classifications to better disaggregate landscapes into units with similar soils (MacMillan et al., 2000). Classification and regression trees have also been a prominent technique used in disaggregation. Bui et al. (2001) and Wei et al. (2010) both used ensembles of decision trees and Haring et al. (2012) used random forests to refine soil and surficial geology classes. Tree-based models have also been used extensively in general DSM applications and seem to have the greatest flexibility of common modeling methods (Behrens et al., 2010a; Behrens et al., 2010b; Bui et al., 2009; Lemerrier et al., 2012; McKenzie and Ryan, 1999; Moran and Bui, 2002; Saunders and Boettinger, 2007; Schmidt et al., 2008; Scull et al., 2005).

The objective of this research was to address the common situation where an older CSM is available, but more detailed soil spatial data is needed and few soil observations are available. We also compare the usefulness of a variety of ASTER satellite imagery and USGS 1 arc-second National Elevation Dataset (NED) derived data layers for use in CSM disaggregation. We utilize soil-landscape rules that are usually present in soil survey database map unit descriptions in combination with a random forest to universally disaggregate a CSM to a ~30-meter resolution raster soil class map without collecting new field data. This approach leverages both the implicit information of the SSURGO spatial data (the standard CSM data product), and explicit expert knowledge about soil-geomorphology relationships attributed in the SSURGO database to create a training set. It pairs the training set with elevation and imagery in a random forest classification tree ensemble model. We selected methods and data sources based on repeatability, transparency, and manageability in an effort to make them more accessible to soil science professionals in government and consulting.

**Methods**

**Study Area**

Organ Pipe Cactus National Monument (ORPI) is located in the Basin and Range physiographic province of southern Arizona, USA. The area is characterized by alternating mountain ranges of diverse lithology and broad alluvial valleys with bajada and basin floor systems (Hendricks, 1985). The mountain ranges within ORPI include both intrusive and extrusive igneous rocks as well as sedimentary and meta-sedimentary materials with a wide variety of mineral assemblages (Figure 1) (Bezy et al., 2000; Eddy et al., 1991). The geologic history of the area includes four distinct periods of volcanism starting 1.6 billion years ago and ending with the last episode 26 to 14 million years before present (Bezy et al., 2000; Eddy et al., 1991). Tectonic uplift and erosion have worked and reworked the landscape during this time to produce a complicated arrangement of mountains rising up to 1,465 meters in elevation with an intricate assemblage of associated alluvial outwash landforms (Bezy et al., 2000). Current area geomorphology is a result of Pleistocene and Holocene aggradations and entrenched cycles leaving complex arrangements of deposits with varying dissection and escarpment patterns that differ between lithologic source materials (Bezy et al., 2000; McAuliffe, 1994; Parker, 1991; Parker, 1995; Simpson, 1991).

The area spans the transition from the Arizona Upland to Lower Colorado River Valley subdivision of the Sonoran Desert. This includes a variety of vegetation communities including juniper woodland in the high Ajo Mountains, desert scrub columnar cacti communities on Bajadas, and sparsely vegetated creosote flats in the Growler Valley (Figure 1) (Parker, 1991). Average annual precipitation at the ORPI headquarters is reported to be 251 mm with a strong bimodal distribution characterized by summer monsoon precipitation and a moderate winter rainy season (NOAA, 2004). The mean annual temperature from 1971 to 2000 was 21.6°C with a range of -10 to 47.8°C (NOAA, 2004). A precipitation gradient exists within ORPI decreasing from 342 mm in the high elevation Ajo Mountains to 190 mm in the western areas of the monument (Parker, 1991). Soils in ORPI were all classified as having an aridic soil moisture regime and hyperthermic soil temperature regime (USDA-NRCS, 1972). However, higher elevation areas characterized by juniper woodland may include ustic soil moisture and thermic soil temperature regimes.
Training Set Creation

The CSM used here includes a SSURGO dataset that consists of a polygon format vector map attributed with a map unit label and a relational database that attaches soil information to the map units. SSURGO includes multiple types of map units that generally have one to four named soil series components as well as “inclusions” of other soils or non-soil areas. Each of these component soil series can have different property distributions that must be generalized or aggregated somehow if a user wants to display a soil property using SSURGO polygons (e.g., Bliss et al., 1995; Thompson and Kolka, 2005).

In SSURGO, each component of every map unit has information regarding soil properties, as well as geologic and geomorphic characteristics, attributed to it (Table 1). After reviewing an extensive set of environmental rasters and SSURGO attributes for potential use in training models, we determined that a simple scheme that matches DEM derived layers to geomorphology attributes would be effective. By querying the geomorphic landform tables in SSURGO (cogeomorphdesc and cosurfmorphgc), soil-landscape relationship descriptions were extracted from the database to help determine where within a map unit a component is expected to occur, e.g., Growler series exists on the convex portions of valley floors. The geomorphic landform descriptions in these queries were then matched to values in environmental rasters that represent topographic wetness index (TWI) (Yang et al., 2007), created using Tarboton’s (1997) surface flow routing algorithm, and to relative elevation metrics within different neighborhoods to help distinguish components within map units (Table 1).

Rule sets were developed to match the descriptive language from the geomorphology queries as well as delineate small washes present in map units as inclusions. In most single component map units this only involved trying to eliminate small washes that were inclusions by setting a TWI cutoff of 17. This threshold was chosen by draping the TWI layer onto high resolution USGS 1-meter DOQQ imagery and matching a TWI value to delineate visible washes. Single component map units with younger soils that still flood, e.g., Gilman very fine sandy loam, were left alone for training selection because they were deemed acceptable as is.

For multi-component map units, each component geomorphic description was examined, and DEM variables were selected on the basis of which variable best distinguished the labeled differences. This was often difficult because the language in SSURGO can seem contradictory. For example, the Growler series in the Growler-Antho complex is described on “valley floors in convex portions,” and Antho is described on “flood plains in dips” or on “alluvial fans” in areas with “terrace tread” (Table 1). This seems to indicate that there are areas of flood plain, alluvial fan and valley floor in the map unit. However, when these map units were examined in the field and in aerial photography, they appeared to be alluvial fan shaped delineations that also have very subtle topography more like a valley floor with slightly lower areas that still flood and other slightly higher areas that do not receive much overland flow. Based on this observation, we decided to use the TWI raster to split these areas apart into lower wetter areas and higher convex locations. We also checked the original hardcopy soil survey manuscript for clarification which indicated that “Growler soils lie on slightly elevated convex areas” and are “easily recognized” by “varnished desert pavement and sparse vegetation.” Similarly, the report states that
<table>
<thead>
<tr>
<th>Map Unit Name</th>
<th>Component (% of map unit)</th>
<th>SSURGO Component Geomorphic Descriptions*</th>
<th>Raster rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajo very cobbly sandy loam, 2 to 5 percent slopes</td>
<td>Ajo (90%)</td>
<td>terrace tread on alluvial fans</td>
<td>TWI**</td>
</tr>
<tr>
<td>Ajo very gravelly loam, 1 to 5 percent slopes</td>
<td>Ajo (90%)</td>
<td>terrace tread on alluvial fans</td>
<td>H3**</td>
</tr>
<tr>
<td>Antho fine sandy loam</td>
<td>Antho (95%)</td>
<td>terrace tread on alluvial fans</td>
<td>R20**</td>
</tr>
<tr>
<td>Antho fine sandy loam</td>
<td>Antho (95%)</td>
<td>dips in flood plains</td>
<td></td>
</tr>
<tr>
<td>Antho soils, very gravelly variants</td>
<td>Antho (85%)</td>
<td>terrace tread on alluvial fans</td>
<td></td>
</tr>
<tr>
<td>Cherioni gravelly very fine sandy loam, 0 to 8 percent slopes</td>
<td>Cherioni (95%)</td>
<td>low beveled side slopes of hills</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Cipriano gravelly loam</td>
<td>Cipriano (90%)</td>
<td>terrace tread on bajadas</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Cipriano gravelly loam</td>
<td>Cipriano (90%)</td>
<td>terrace tread on alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Gachado extremely cobbly loam, 2 to 8 percent slopes</td>
<td>Gachado (75%)</td>
<td>toe slopes of hills</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Gachado extremely cobbly loam, 2 to 8 percent slopes</td>
<td>Gachado (75%)</td>
<td>mountainflank toe slopes</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Gilman very fine sandy loam</td>
<td>Gilman (90%)</td>
<td>alluvial fans, lower, terrace tread</td>
<td></td>
</tr>
<tr>
<td>Gilman very fine sandy loam</td>
<td>Gilman (90%)</td>
<td>dips in flood plains</td>
<td></td>
</tr>
<tr>
<td>Gilman very fine sandy loam, saline</td>
<td>Gilmansi (90%)</td>
<td>terrace tread on lower alluvial fans</td>
<td></td>
</tr>
<tr>
<td>Gilman very fine sandy loam, saline</td>
<td>Gilmansi (90%)</td>
<td>dips in flood plains</td>
<td></td>
</tr>
<tr>
<td>Growler-Antho complex</td>
<td>Antho (45%)</td>
<td>dips in flood plains</td>
<td>TWI &gt;= 14.5</td>
</tr>
<tr>
<td>Growler-Antho complex</td>
<td>Antho (45%)</td>
<td>terrace tread on alluvial fans</td>
<td>TWI &gt;= 14.5</td>
</tr>
<tr>
<td>Growler-Antho complex</td>
<td>Growler (35%)</td>
<td>convex portions of valley floors</td>
<td>TWI &lt; 14.5</td>
</tr>
<tr>
<td>Gunsight very gravelly loam, 0 to 2 percent slopes</td>
<td>Gunsight (75%)</td>
<td>terrace tread on lower portions of alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Gunsight very gravelly loam, 2 to 15 percent slopes</td>
<td>Gunsight (80%)</td>
<td>terrace tread on lower portions of alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Harqua very cobbly loam, 0 to 8 percent slopes</td>
<td>Harqua (90%)</td>
<td>terrace tread of degrading surface on plains</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Harqua very gravelly loam, 0 to 3 percent slopes</td>
<td>Harqua (90%)</td>
<td>terrace tread of degrading surfaces on plains</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Harqua-Gunsight complex</td>
<td>Gunsight (40%)</td>
<td>terrace tread on lower portions of alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Harqua-Gunsight complex</td>
<td>Harqua (40%)</td>
<td>terrace tread on degrading surfaces of plains</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Laveen loam</td>
<td>Laveen (85%)</td>
<td>tread of old terraces</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Lomitas extremely stony loam, 8 to 40 percent slopes</td>
<td>Lomitas (75%)</td>
<td>hills, Side Slope</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Lomitas extremely stony loam, 8 to 40 percent slopes</td>
<td>Lomitas (75%)</td>
<td>mountainflanks</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Perryville very cobbly fine sandy loam, 0 to 8 percent slopes</td>
<td>Perryville (80%)</td>
<td>terrace tread of old alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Rillito gravelly sandy loam</td>
<td>Rillito (75%)</td>
<td>terrace tread on alluvial fans</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Rock land</td>
<td>Rock land (90%)</td>
<td>mountain slopes</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Rock outcrop</td>
<td>Rock outcrop (90%)</td>
<td>mountain slopes and peaks</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Stony land-Rock outcrop association</td>
<td>Rock outcrop (30%)</td>
<td>mountain peaks</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Stony land-Rock outcrop association</td>
<td>Stony land (65%)</td>
<td>hill side slopes</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Stony land-Rock outcrop association</td>
<td>Stony land (65%)</td>
<td>mountainflanks</td>
<td>TWI &lt; 17</td>
</tr>
<tr>
<td>Torrifluvents</td>
<td>Torrifluents (90%)</td>
<td>dips with eroded overflow stream channels</td>
<td>original map units or twi &gt; 17 (any map unit)</td>
</tr>
</tbody>
</table>

* These are interpretations that take the original nouns in the SSURGO database and link them using prepositions to create meaningful context.
** twi = topographic wetness index.
*** No rules were included for these soils because they exist on floodplains effectively eliminating the need for exclusion of washes by twi.
Table 2. List of Raster Covariate Layers Used for Building Tree Models from Selected Training Sets

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Name</th>
<th>Original or Updated*</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_1$</td>
<td>$b_1$</td>
<td>ASTER Band 1</td>
<td>Original</td>
<td>VNIR Reflectance (0.52-0.6 μm)</td>
</tr>
<tr>
<td>$b_2$</td>
<td>$b_2$</td>
<td>ASTER Band 2</td>
<td>Original</td>
<td>VNIR Reflectance (0.63-0.69 μm)</td>
</tr>
<tr>
<td>$b_3$</td>
<td>$b_3$</td>
<td>ASTER Band 3</td>
<td>Original</td>
<td>VNIR Reflectance (0.76-0.86 μm)</td>
</tr>
<tr>
<td>$b_4$</td>
<td>$b_4$</td>
<td>ASTER Band 4</td>
<td>Original</td>
<td>SWIR Reflectance (1.6-1.7 μm)</td>
</tr>
<tr>
<td>$b_5$</td>
<td>$b_5$</td>
<td>ASTER Band 5</td>
<td>Original</td>
<td>SWIR Reflectance (2.145-2.185 μm)</td>
</tr>
<tr>
<td>$b_6$</td>
<td>$b_6$</td>
<td>ASTER Band 6</td>
<td>Original</td>
<td>SWIR Reflectance (2.185-2.225 μm)</td>
</tr>
<tr>
<td>$b_7$</td>
<td>$b_7$</td>
<td>ASTER Band 7</td>
<td>Original</td>
<td>SWIR Reflectance (2.235-2.285 μm)</td>
</tr>
<tr>
<td>$b_8$</td>
<td>$b_8$</td>
<td>ASTER Band 8</td>
<td>Original</td>
<td>SWIR Reflectance (2.295-2.365 μm)</td>
</tr>
<tr>
<td>$b_9$</td>
<td>$b_9$</td>
<td>ASTER Band 9</td>
<td>Original</td>
<td>SWIR Reflectance (2.360-2.430 μm)</td>
</tr>
<tr>
<td>$b_{2b1}$</td>
<td>$b_{2b1}$</td>
<td>ASTER Ratio 2/1</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/1</td>
</tr>
<tr>
<td>$b_{2b4}$</td>
<td>$b_{2b4}$</td>
<td>ASTER Ratio 2/4</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/4</td>
</tr>
<tr>
<td>$b_{2b5}$</td>
<td>$b_{2b5}$</td>
<td>ASTER Ratio 2/5</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/5</td>
</tr>
<tr>
<td>$b_{2b6}$</td>
<td>$b_{2b6}$</td>
<td>ASTER Ratio 2/6</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/6</td>
</tr>
<tr>
<td>$b_{2b7}$</td>
<td>$b_{2b7}$</td>
<td>ASTER Ratio 2/7</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/7</td>
</tr>
<tr>
<td>$b_{2b8}$</td>
<td>$b_{2b8}$</td>
<td>ASTER Ratio 2/8</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/8</td>
</tr>
<tr>
<td>$b_{2b9}$</td>
<td>$b_{2b9}$</td>
<td>ASTER Ratio 2/9</td>
<td>Original</td>
<td>Reflectance Ratio Bands 2/9</td>
</tr>
<tr>
<td>$b_{2sd14}$</td>
<td>$b_{2sd14}$</td>
<td>Band 2 Std. Dev - 14-pixel</td>
<td>Original</td>
<td>Std Dev of Band 2 in a 14-pixel radius</td>
</tr>
<tr>
<td>$b_{2sd3}$</td>
<td>$b_{2sd3}$</td>
<td>Band 2 Std. Dev - 3-pixel</td>
<td>Original</td>
<td>Std Dev of Band 2 in a 3-pixel radius</td>
</tr>
<tr>
<td>$b_{2sd5}$</td>
<td>$b_{2sd5}$</td>
<td>Band 2 Std. Dev - 5-pixel</td>
<td>Original</td>
<td>Std Dev of Band 2 in a 5-pixel radius</td>
</tr>
<tr>
<td>$b_{3b2}$</td>
<td>$b_{3b2}$</td>
<td>ASTER Ratio 3/2</td>
<td>Original</td>
<td>Reflectance Ratio Bands 3/2</td>
</tr>
<tr>
<td>$b_{4b3}$</td>
<td>$b_{4b3}$</td>
<td>ASTER Ratio 4/3</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/3</td>
</tr>
<tr>
<td>$b_{4b5}$</td>
<td>$b_{4b5}$</td>
<td>ASTER Ratio 4/5</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/5</td>
</tr>
<tr>
<td>$b_{4b6}$</td>
<td>$b_{4b6}$</td>
<td>ASTER Ratio 4/6</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/6</td>
</tr>
<tr>
<td>$b_{4b7}$</td>
<td>$b_{4b7}$</td>
<td>ASTER Ratio 4/7</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/7</td>
</tr>
<tr>
<td>$b_{4b8}$</td>
<td>$b_{4b8}$</td>
<td>ASTER Ratio 4/8</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/8</td>
</tr>
<tr>
<td>$b_{4b9}$</td>
<td>$b_{4b9}$</td>
<td>ASTER Ratio 4/9</td>
<td>Original</td>
<td>Reflectance Ratio Bands 4/9</td>
</tr>
<tr>
<td>$b_{4s414}$</td>
<td>$b_{4s414}$</td>
<td>Band 4 Std. Dev - 14-pixel</td>
<td>Original</td>
<td>Std Dev of Band 4 in a 14-pixel radius</td>
</tr>
<tr>
<td>$b_{4sd3}$</td>
<td>$b_{4sd3}$</td>
<td>Band 4 Std. Dev - 3-pixel</td>
<td>Original</td>
<td>Std Dev of Band 4 in a 3-pixel radius</td>
</tr>
<tr>
<td>$b_{4sd5}$</td>
<td>$b_{4sd5}$</td>
<td>Band 4 Std. Dev - 5-pixel</td>
<td>Original</td>
<td>Std Dev of Band 4 in a 5-pixel radius</td>
</tr>
<tr>
<td>$ca$</td>
<td>$ca$</td>
<td>Contributing Area</td>
<td>Original</td>
<td>Upstream surface area contributing flow to pixel</td>
</tr>
<tr>
<td>$twi$</td>
<td>$twi$</td>
<td>Topographic Wetness Index</td>
<td>Original</td>
<td>Calc: ln(ca / tan(slope))</td>
</tr>
<tr>
<td>$wetness_{tn}$</td>
<td>$wetness_{tn}$</td>
<td>Modified twi</td>
<td>Original</td>
<td>Calc: ln(ca / (meandiff25 / range of meandiff25))</td>
</tr>
<tr>
<td>$curv$</td>
<td>$curv$</td>
<td>Horizontal Curvature</td>
<td>Original</td>
<td>2nd-derivative across slope contour</td>
</tr>
<tr>
<td>$diff25$</td>
<td>$diff25$</td>
<td>Difference from Max - 25-pixel</td>
<td>Original</td>
<td>Max elevation in 25-pixel radius minus the cell value</td>
</tr>
<tr>
<td>$meandiff25$</td>
<td>$meandiff25$</td>
<td>Mean difference - 25-pixel</td>
<td>Original</td>
<td>Mean elevation in 25-pixel radius minus the cell value</td>
</tr>
<tr>
<td>$slopepos25$</td>
<td>$slopepos25$</td>
<td>Slope Position - 25-pixel</td>
<td>Original</td>
<td>(Max elevation in 25-pixel radius minus the cell value)/(range)</td>
</tr>
<tr>
<td>$swness$</td>
<td>$swness$</td>
<td>Southwesternness</td>
<td>Original</td>
<td>A -1 to 1 index of how southwest a slope faces: cos(aspect - 225°)</td>
</tr>
<tr>
<td>$plen$</td>
<td>$plen$</td>
<td>Longest Upslope Length</td>
<td>Original</td>
<td>Length of longest flow path above each cell</td>
</tr>
<tr>
<td>$tlen$</td>
<td>$tlen$</td>
<td>Total Upslope Length</td>
<td>Original</td>
<td>Additive length of all upslope flowpaths for each cell</td>
</tr>
<tr>
<td>$rel_{ht_3}$</td>
<td>$rel_{ht_3}$</td>
<td>Local Height - 3-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 3-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_5}$</td>
<td>$rel_{ht_5}$</td>
<td>Local Height - 5-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 5-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_10}$</td>
<td>$rel_{ht_10}$</td>
<td>Local Height - 10-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 10-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_20}$</td>
<td>$rel_{ht_20}$</td>
<td>Local Height - 20-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 20-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_30}$</td>
<td>$rel_{ht_30}$</td>
<td>Local Height - 30-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 30-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_50}$</td>
<td>$rel_{ht_50}$</td>
<td>Local Height - 50-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 50-pixel radius</td>
</tr>
<tr>
<td>$rel_{ht_70}$</td>
<td>$rel_{ht_70}$</td>
<td>Local Height - 70-pixel</td>
<td>Updated</td>
<td>Height of cell above the local minimum elevation in 70-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht3}$</td>
<td>$rel_{meanht3}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 3-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht5}$</td>
<td>$rel_{meanht5}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 5-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht10}$</td>
<td>$rel_{meanht10}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 10-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht20}$</td>
<td>$rel_{meanht20}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 20-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht30}$</td>
<td>$rel_{meanht30}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 30-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht50}$</td>
<td>$rel_{meanht50}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 50-pixel radius</td>
</tr>
<tr>
<td>$rel_{meanht70}$</td>
<td>$rel_{meanht70}$</td>
<td>Local Relief - 10-pixel</td>
<td>Updated</td>
<td>Height of cell relative to local mean elevation in 70-pixel radius</td>
</tr>
</tbody>
</table>

*Original refers to variables that were used directly from Nauman (2009), and Updated refers to variables that were added to the ‘Updated’ model beyond those from the 2009 study.
“Antho soils lie between the Growler areas and along shallow drainage ways” (USDA-NRCS, 1972). Based on this, we found that the TWI raster distinguished drainage patterns in Growler-Antho complex in such a way as to delineate subtle washes and variated surfaces based on comparing TWI breaks with visual inspection of 1-meter USGS DOQQ aerial photography. Following this process, we translated soil-landscape relationship records in the SSURGO database to environmental raster values. This rule translation attempted to emulate and streamline the approach Thompson et al. (2010) used to create spatial soil-landscape rule-sets. We simply employ these rule-sets as a means to create a model training set.

All rules were presumed to identify typical landscapes for respective component soils within each map unit. The areas selected by each component rule were added to a training set that was compiled for all soil series and other named components (e.g., rock outcrops or higher taxa such as Torrilluvents). In other words, areas from all map units that were “typical” of a given soil series were combined into one training set for each respective soil series. All environmental rasters used in rule-matching were derived from the 1-arc second USGS national elevation dataset (ned) (Gesch, 2007; Gesch et al., 2002). Final maps and other raster data used in later modeling steps were co-registered to the NED grid with the North American Datum of 1983 Universal Transverse Mercator projection in Zone 12N.

Model Building
Training areas for each soil series/component were randomly sampled proportionally to component areal extent in the original CSM (following Moran and Bui, 2002) to produce two ensemble models. Random forest classifications were built from the training sets using a more exhaustive set of environmental covariate rasters than were used in the original training rule-matching (Table 2). Variables used by Nauman (2009) for unsupervised soil-landscape classifications were used for one model (“Original” model) to compare with that study, and a set of new variables that highlight subtle topographical differences were added to that dataset to build an “Updated” model. The Updated model added a suite of relative elevation metrics to help with classification based on results from recent studies that show that varying neighborhood sizes of terrain indices can improve spatial prediction of soils (Behrens et al., 2010a; Behrens et al., 2010b).

Covariate Data Sources
ASTER satellite imagery (Abrams, 2000) and the 1 arc-second USGS National Elevation Dataset (NED) (Gesch, 2007; Gesch et al., 2002) were used for covariate layers. All terrain-based rasters were derived from NED after re-projection to NAD83-UTM12N using a bilinear interpolation (Table 2). An ASTER scene from 18 December 2001, was chosen for clear conditions and spatial coverage of ORPI. The ASTER On-Demand L3 Orthorectified imagery was acquired from LP-DAAC (http://edcimswww.cr.usgs.gov/pub/imswelcome/). Radiance at the sensor was calculated from the original imagery scaled radiance (DN values) based upon ASTER coefficients published at the LP-DAAC website (Abrahm et al., 2001). These radiance values were subsequently modified using a ground based correction coefficient supplied by the University of Arizona Optical Sciences Remote Sensing Group (Buchanan, 2007). Radiance images were then converted to reflectance values using the COST method (Chavez, 1996; instructions in Appendix A of Nauman, 2009). Inputs for this conversion included an average of two commonly used solar irradiation models for ASTER bands. World Radiation Center (WRC) and “MODTRAN” (Thome et al., 2001, p. 264). Earth-Sun distances were obtained online from the NASA horizons web-server (Giorgini et al., 1996). ASTER reflectance bands were used for all imagery variables (Table 2).

Decision Tree Classification
Tree-based machine learning techniques have shown great potential in the modeling of ecology and soil systems (Bui and Moran, 2001; Henderson et al., 2005; Bui et al., 2006; Minasy and McBratney, 2007; Schmidt et al., 2008; Behrens et al., 2010b). Generally, these algorithms recursively split a dataset by picking breaks in covariate data that help to purify or increase the information content of the model branches (Breiman, 1984; Pedregosa et al., 2011). The Scikit Machine Learning Tree module was used in Python for decision tree implementation and folows a Classification and Regression Tree (CART) implementation which allows for numerical and categorical variables to be used as inputs as well as for a target variable (Pedregosa et al., 2011; Scikit-learn.org, 2013). The algorithm as we implemented in the Tree module uses Gini’s impurity to measure the quality of splits for tree building and randomizes variable selection at each node to implement a Random Forest (Breiman, 2001). We conducted an informal sensitivity analysis with the parameters controlling maximum tree depth and minimum node split sample size to try and balance fit with tree complexity. Due to the large number of training classes and a complicated and geologically stratified study area, we felt that trees needed to be allowed to grow relatively large. A maximum tree depth of 20 splits and a 20-pixel minimum sample size to attempt a split were selected for simple tree pruning parameters. At each split in all trees, 18 variables were randomly chosen from the greater suite for possible use in rule creation. Fifty percent of the training set was randomly sampled with replacement for use in each tree. In each tree training sample, individual component class sizes were trimmed down so that all components retained the same relative proportions as in the original SSURGO survey. A 500-tree ensemble was generated for both models.

Validation of Disaggregated Maps
Disaggregated maps were validated with independent field data from ten USDA-NRCS pedon database locations (National Cooperative Soil Survey, 2012), and 53 field checks in 2006 and 2007 (Figure 1). Access to ORPI is very difficult due to the remoteness of the area and active smuggling issues along the international border with Mexico. Due to these logistical challenges, field checks were only allowed in limited areas along certain roads. As such, validation points were not randomly allocated and were located to best represent the units where access was granted. We were able to get points in 17 of 23 SSURGO map units and 16 of 18 disaggregated components (Updated model) given the field limitations (see Table 3). Field checks were determined by small hole and/or auger check of diagnostic soil features (e.g., argillic horizon) and basic soil morphology (i.e., rock content, texture, carbonates, surface rock, and color) to match with the nearest soil series. These field locations were intersected with disaggregation results to estimate overall classification accuracy, and to determine uncertainty based on the underlying ensemble model frequencies.

Results
The two disaggregation models performed well with training accuracies of 80 percent for the Original model and 85 percent for the Updated model. Corresponding validation accuracies were 66.7 percent for the Original model and 69.8 percent for the Updated model (Table 3). At validation sites, the original SSURGO map units listed one of the correct validation soils 74.6 percent of the time. Although this agreement is higher than the disaggregation models, the multi-component map units inflate the accuracy because they offer more than one possible class that can count for a match in a polygon. In contrast, the disaggregation models always predict one most-likely soil for one pixel, so a comparison to the validation of SSURGO map.
units with multinomial themes is not objective. To better compare performance between the original CSM and disaggregated maps, we looked at validation points that fell into single-component soil consociation SSURGO map units (52 total sites) to see if that affected agreement rates. In consociations, SSURGO matched at 73.1 percent of sites and the Updated disaggregated map at 75.0 percent of those same locations indicating very similar accuracies. The resulting disaggregated map presents a single consistent theme (i.e., one most-probable soil component per pixel; Plate 1b) whereas the SSURGO map units sometimes aggregate multiple components in map units and also sometimes delineate multiple map units with the same soil by breaking out general slope gradient classes (Plate 1a).

Uncertainty Map

Prediction frequencies of classes in the best performing random forest model (Updated) were compared between pixels that both matched and did not match validation site observations to create a simple field data derived uncertainty surface. Figure 2 shows how validation site prediction probabilities (derived from model prediction frequencies) were compared and translated to an uncertainty surface. Prediction probabilities were found to be higher at sites where validation matched predictions (Wilcoxon rank sum test with continuity correction, $W = 571.5$, 1-sided $p = 0.0187$, 2-sided $p = 0.0373$; Figure 2a). Although match rates seem to increase with prediction probabilities, there is a small drop in field data agreement rates in the highest bin (0.9 to 1.0) that does not follow the positive trend (Figure 2d). It is difficult to make any detailed conclusions beyond the overall positive relationship between field data probabilities and prediction probabilities because the actual sample sizes of the step function bins are all less than twenty and are variable in size. These small sample sizes mean that a change of just one validation match would influence any bin by more than 1/20, or 5 percent. In this case, the difference in bin sample size between the 0.8 to 0.9 bin ($n = 11$) and 0.9 to 1.0 bin ($n = 19$) makes it difficult to determine if the drop in the 0.9 to 1.0 probability (Figure 2d) is due to the difference in bin sample sizes or the predictive ability of the model. However, the overall field data accuracies still are all above 0.56, indicating a good deal of predictive ability across all prediction probabilities. It also appears clear that above a prediction probability of 0.70 the field data probability also goes up to above 0.70. Field data probabilities were mapped by translating prediction probabilities produced from the random forest using the step function shown in Figure 2d in order to create a map that can serve to represent uncertainty in predictions (Figure 3).

Examination of field data probability (uncertainty) values among original SSURGO map units and Updated disaggregation model components did show some variability between classes, but all class means were between 0.65 and 0.74 (Figure 4). Variation is evident within classes when SSURGO and Updated model maps are overlaid on the field data probability map. Visually, there were lower probability areas around some of the larger mountains (e.g., Rock land) or fan and wash scarps. Delineations of the Growler-Antho complex in the Valley of the Ajo (Figure 1) were a good example of lower alluvial fan units that had lower probability values (Figure 3). However, the large area of the Growler-Antho complex in the Growler Valley (Figure 1) had generally higher probabilities (Figure 3) suggesting that there might be some kind of difference between Growler-Antho units in the different valleys. Areas of valley floor (e.g., Gilman), Torrifluvents (in the Updated model), middle bajadas (e.g., Gunsight), upper alluvial fans (e.g., Ajo), and non-soil components of the larger mountains (e.g., Rock land) tended to have higher field data probabilities. We also observed that in both the Updated

### Table 3. Summary of the Distribution of Validation Points in Original SSURGO Map Units (Left), and the Updated Disaggregation Model (Right)

<table>
<thead>
<tr>
<th>SSURGO Map Unit (Plate 1a)</th>
<th>Validation Points</th>
<th># Correct</th>
<th>% Correct</th>
<th>Predicted Component in Updated Model (Plate 1b)</th>
<th>Validation Points</th>
<th># Correct</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajo very gravelly loam, 1% to 5% slopes</td>
<td>6</td>
<td>3</td>
<td>50.0%</td>
<td>Ajo</td>
<td>6</td>
<td>3</td>
<td>50.0%</td>
</tr>
<tr>
<td>Antho fine sandy loam</td>
<td>7</td>
<td>6</td>
<td>85.7%</td>
<td>Antho</td>
<td>7</td>
<td>6</td>
<td>85.7%</td>
</tr>
<tr>
<td>Cherioni gravelly very fine sandy loam, 0% to 8% slopes</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
<td>Cherioni</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
</tr>
<tr>
<td>Cipriano gravelly loam</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
<td>Cipriano</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
</tr>
<tr>
<td>Gachado extremely cobbly loam, 2% to 8% slopes</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>Gachado</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
</tr>
<tr>
<td>Gilman very fine sandy loam</td>
<td>4</td>
<td>4</td>
<td>100.0%</td>
<td>Gilman</td>
<td>6</td>
<td>5</td>
<td>83.3%</td>
</tr>
<tr>
<td>Gilman very fine sandy loam, saline</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>Gilman saline</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Growler-Antho complex</td>
<td>4</td>
<td>4</td>
<td>100.0%</td>
<td>Growler</td>
<td>2</td>
<td>2</td>
<td>100.0%</td>
</tr>
<tr>
<td>Gunsight very gravelly loam, 2% to 15% slopes</td>
<td>9</td>
<td>6</td>
<td>66.7%</td>
<td>Gunsight</td>
<td>18</td>
<td>10</td>
<td>55.6%</td>
</tr>
<tr>
<td>Harqua very cobbly loam, 0% to 8% slopes</td>
<td>3</td>
<td>2</td>
<td>66.7%</td>
<td>Harqua</td>
<td>4</td>
<td>2</td>
<td>50.0%</td>
</tr>
<tr>
<td>Harqua-Gunsight complex</td>
<td>7</td>
<td>5</td>
<td>71.4%</td>
<td>Laveen</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
</tr>
<tr>
<td>Laveen loam</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>Lomitas</td>
<td>4</td>
<td>2</td>
<td>50.0%</td>
</tr>
<tr>
<td>Lomitas extremely stony loam, 8% to 40% slopes</td>
<td>4</td>
<td>3</td>
<td>75.0%</td>
<td>Perryville</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
</tr>
<tr>
<td>Perryville very cobbly fine sandy loam, 0% to 8% slopes</td>
<td>1</td>
<td>1</td>
<td>100.0%</td>
<td>Rillito</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
</tr>
<tr>
<td>Rillito gravelly sandy loam</td>
<td>3</td>
<td>3</td>
<td>100.0%</td>
<td>Rock land</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Rock land</td>
<td>1</td>
<td>0</td>
<td>0.0%</td>
<td>Torrifluvents</td>
<td>2</td>
<td>2</td>
<td>100.0%</td>
</tr>
<tr>
<td>Torrifluvents</td>
<td>5</td>
<td>2</td>
<td>40.0%</td>
<td>Total</td>
<td>63</td>
<td>44</td>
<td>69.8%</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td>47</td>
<td>74.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
model and SSURGO that classes with less than 10,000 pixels (e.g., Gachado, Gilman-saline, Laveen, and Perryville) all had lower field data probability averages while the largest classes (e.g., Antho, Gilman, Gunsight, Lomitas, Rock land, and Stony land) with 100,000 or more pixels had generally higher means.

We are cautious about over-interpretation of this uncertainty data because the field data probability values are means of discretized probability classes created from 63 validation points, and this analysis is generalizing 1.6 million pixels. The large sample sizes (number of pixels) in individual classes would likely result in statistical differences between probability class means in both the SSURGO and Updated model maps. This would mainly be a result of the large class sample sizes that produce extremely low standard errors and hence greater statistical detectability that may not be meaningful. For example, the largest class standard error of field data probability values for all SSURGO maps units or Updated model components was 0.002.

Important Variables for Models

Variable usefulness was evaluated based on the relative frequency of each variable’s use in the random forests. In both the Original and Updated models, NED derived variables were generally used more than ASTER variables (Figure 5). The variables \( \text{dem} \) and \( \text{diff25} \) were in the top three used for tree building in both models, suggesting a strong influence from watershed-scale elevation gradients. The Original model, which included fewer NED derived variables, tended to rely more evenly on ASTER and NED variables, whereas the Updated model uses more NED based layers with 10 of the 14 newly added NED relative elevation variables being used quite frequently. This might suggest that the relative height metrics added to the Updated model provided more consistent predictive power than the ASTER variables they appear to replace in the Original model. However, considering that the Original model was only 3.1 percent less accurate at validation sites than the Updated model, the differences in variable use may not be that significant. Frequently used ASTER derived layers span a wide scope of types in the Original model including three reflectance layers, one band ratio, and three band neighborhood variation layers. In contrast, the Updated model only includes Band 4 and Band 2 neighborhood variation layers from ASTER. Overall, the \( \text{dem} \) variable and relative elevation surfaces that integrate more than 10 pixels seem to be used most often in trees, possibly indicating more predictive power than the other included variables.

In comparing the principal components analysis (PCA) unsupervised data reduction variable selection from Nauman (2009) to the random forest use of same data in the Original disaggregation model, eight of the ten variables used were selected in both of the studies as important (Figure 5, Table 4). This commonality drops to four of ten when comparing with the Updated random forest model. It is interesting that the \( \text{diff25} \) variable was not selected in Nauman (2009) as it is a more important variable in both of the random forest models. Of the ASTER-derived variables, neighborhood standard deviation of reflectance layers are the only variables that show up as important in both disaggregation models as well as the PCA unsupervised data reduction (Nauman, 2009).

Discussion and Conclusions

Our assessment is that both versions of random forest models worked well to disaggregate the CSM of ORPI, and that the near equivalent accuracy of SSURGO and the Updated model indicated that the models were able to reproduce much of the information captured by the survey. The accuracy at validation points in ORPI was higher than implementation of this same general methodology in the dissected Allegheny Plateau and Mountains of southern West Virginia (WV) (Nauman and Thompson, 2014) where classification validation accuracies ranged from 24 percent to 44 percent depending on spatial supports used in validation. However, similar to the ORPI results, the reported accuracy of the original WV CSM (27% to 41%) closely matched the disaggregation accuracy. It is encouraging that in both studies this method seems to produce...
a higher spatial resolution soil map at accuracies similar to the original soil surveys. The field validation accuracy in ORPI (69.8 percent) was also similar to results of similar studies in New Brunswick, Canada (64.9 percent to 67.6 percent, Yang et al., 2011), the African dambo (75.5 percent; Hansen et al., 2009), and in the Bavarian forests of Germany (70 percent; Haring et al., 2012).

The relationship between ensemble model prediction frequency (prediction probabilities) and validation accuracies allowed for a simple representation of classification uncertainty in both ORPI and WV (Nauman and Thompson, 2014) studies, indicating some degree of consistency across different physiographic regions using these approaches. This relationship might also prove useful in future studies for using prediction probabilities in tree ensembles for creating fuzzy membership classifications. This scenario would involve using the proportion of tree predictions as membership functions where multiple soils are predicted for the same pixel in different trees within the ensemble model. This should be investigated in future disaggregation attempts as it would allow for continuous fuzzy thematic representations of soil classes (e.g., Burrough, 1989; Burrough et al., 1997; De Gruijter et al., 1997; Hodza, 2010; Lagacherie et al., 1997; McBratney and Odeh, 1997; Qi et al., 2006; Zhu, 1997; Zhu et al., 1996, 2001, and

Figure 2. Validation site probabilities compared for Updated model showing the higher tendency of prediction probabilities at (a) matches or correctly predicted sites, (b) histograms of prediction probabilities at missed sites, (c) matched sites, and (d) the empirical relationship relating prediction probabilities and field data probabilities.

Figure 3. Field data probability representation of uncertainty for the Updated model based on step function in Figure 4d.
Thematic Issues in Disaggregation

Disaggregated maps create a singular and consistent theme of one most-probable soil class (or non-soil component, e.g., rock outcrops) per pixel. This is different than SSURGO map units which can have multiple soil classes, slope gradient modifiers, and soil taxonomic variants. Soil series variants can be dealt with in these disaggregation approaches if they are included as classes at the training stage, but care should be taken because they might be so closely related to the non-variant soil series that the environmental covariates used in modeling might not be able to discern the two. Slope gradient modifiers have also been used to help with management interpretations in SSURGO (Soil Survey Division Staff, 1993). However, because slope gradient maps can be made at such

Figure 4. Mean field data probabilities with standard deviation bars summarized by (a) Updated model disaggregation components, and (b) ssurgo map units.

Variables Most Used: Updated Model

Figure 5. (a) Average frequency of use of covariate layers (See Table 2 for definitions) used on average in more than 2 percent of tree nodes in the 500 trees in the Updated model; black brackets give standard deviations of the frequencies to show how these varied over all ensemble trees, and (b) Same as 5a, but for the variables used from Nauman (2009).
always, follow certain soil types as mapped in SSURGO and the
aries. We also observed relatively higher uncertainties that
representing transitions between soils along those bound-
ends at the Soil Taxonomy formal classification within US
series there is often inconsistent degrees of detail because
consistent at the soil series taxonomic level. Even among soils
offered with the SSURGO product. We aimed to produce these
Uncertainty Assessment
haps it is a better option to overlay a true soils themed map
display where accuracies were higher or lower. In analyzing
the ensemble trees predicted classes the most consistently
based on how well our models were able to match the original
survey concepts, but due to our lack of ability to validate this
and concerns about how different these soils truly are from
a soil genesis perspective, we chose not to separate variants of
the Ajo, Antho, Harqua, and Gilman soil series. Variants of Ajo and
Harqua were split out based on the presence of cobble sized
rock fragments at the surface and slightly different surface
structure (both only in the top 5 cm), and Antho variants were
distinguished based on variations of rock content at depth. A
saline variant of the Gilman series was also mapped. Among
these variants, we chose to only split out the Gilman saline
variant for disaggregation because it was the only series
where we had validation points in both the original series
and the variant. Saline variants of the Gilman series also have a very unique ecology with Atriplex (saltbrush) domi-
nated vegetation communities and a particular susceptibility
to erosion (USDA-NRCS, 1972). We do feel that all of these
variants could likely be identified in disaggregation models
based on how well our models were able to match the original
survey concepts, but due to our lack of ability to validate this
and concerns about how different these soils truly are from
a soil genesis perspective, we chose not to separate variants of
the Ajo, Antho, and Harqua soil series. Thematic choices at
the training stage of these models are difficult because soil
series variants are generally only locally defined; so if results
from ORPI were compared to other CSM, concepts will be more
consistent at the soil series taxonomic level. Even among soils
series there is often inconsistent degrees of detail because
formal classification within US Soil Taxonomy ends at the
family level (Soil Survey Staff, 2010 and 1999).

high resolutions with modern digital elevation models, per-
haps it is a better option to overlay a true soils themed map
with a slope map for such purposes in modern contexts.

In ORPI, variants of soil series were mapped for the Ajo,
Antho, Harqua, and Gilman soil series. Variants of Ajo and
Harqua were split out based on the presence of cobble sized
rock fragments at the surface and slightly different surface
structure (both only in the top 5 cm), and Antho variants were
distinguished based on variations of rock content at depth. A
saline variant of the Gilman series was also mapped. Among
these variants, we chose to only split out the Gilman saline
variant for disaggregation because it was the only series
where we had validation points in both the original series
and the variant. Saline variants of the Gilman series also have a very unique ecology with Atriplex (saltbrush) domi-
nated vegetation communities and a particular susceptibility
to erosion (USDA-NRCS, 1972). We do feel that all of these
variants could likely be identified in disaggregation models
based on how well our models were able to match the original
survey concepts, but due to our lack of ability to validate this
and concerns about how different these soils truly are from
a soil genesis perspective, we chose not to separate variants of
the Ajo, Antho, and Harqua soil series. Thematic choices at
the training stage of these models are difficult because soil
series variants are generally only locally defined; so if results
from ORPI were compared to other CSM, concepts will be more
consistent at the soil series taxonomic level. Even among soils
series there is often inconsistent degrees of detail because
formal classification within US Soil Taxonomy ends at the
family level (Soil Survey Staff, 2010 and 1999).

Fluvial Strata Layers  Mountain Strata Layers

dem  dem  
meanlndf25  swness  
b2sd3  b4sd14  
b2sd14  b4b5  
b2b1  b4b8  
b1  b1  

TABLE 4. LAYERS CHOSEN BY PCA DATA REDUCTION TO MOST EFFICIENTLY REPRESENT
SOIL-LANDSCAPE VARIABILITY FOR SOIL MAPPING (FROM NAUMAN, 2009)

Covariate Comparisons
Overall results indicate that both NED elevation data and
ASTER provide predictive power for soil survey disaggregation
modeling. Variable importance values showed that the layers
derived from the elevation data were likely more useful in
soil classifications in this environment. Based on the domi-
nance of the top four variables in both random forest models
by the dem and other topographic variables, it would seem
ASTER might not perform well without NED data to supplement
predictions. The presence of the dem variable as dominant in
both the fluvial and mountain areas in the PCA analysis (Table
4) (Nauman, 2009) also seems to support this. However, the
ASTER variable selection also selected more ASTER layers over-
all, which does support ASTER as a viable predictor. It would
be useful in the future to generate a model just using ASTER
or similar imagery and comparing that to models using only
DEM-derived covariates to better test the predictive power of
both data sources based on classification success as opposed to
variable importance values within models as presented here.

as a viable predictor. It would
be useful in the future to generate a model just using ASTER
or similar imagery and comparing that to models using only
DEM-derived covariates to better test the predictive power of
both data sources based on classification success as opposed to
variable importance values within models as presented here.

The dominance of elevation and relative height metrics
seem to relate to the relief-driven topographic sequence of
landforms in ORPI. These landforms include relatively young
mountains with rugged outcrops at summits and limited soil
development on side-slopes. Lower in the topographic se-
sequence sets of alluvial fans are arranged in step-wise patterns
moving away from the mountains, with basin floor deposits at
the valley bottom. Soils tend to follow these step patterns
because each riser between alluvial fans represents erosion
cutting into an older aged deposit with the oldest soils (i.e.,
Ajo - Argic Petrocalcids) being the closest to the mountains
and representing past basin base levels (USDA-NRCS, 1972;
Parker, 1995). These fans seem to have relatively distinct
drainage patterns with some having mainly deep gullies
whereas others include reticulating washes with deposition-
al areas, which were likely detected by the relative height
metrics (e.g., rel_ht_20 in Figure 5 and Table 2) based on the
height of fan treads above the drainage cuts.

The only ASTER layers used on average in more the 2 per-
cent of tree nodes in both models were neighborhood standard
deviation of reflectance layers of ASTER band 4 (lower wave-
length short-wave infrared) and band 2 (red-visible). Generally,
these layers highlight areas that have more active geomorphic
or hydrologic dynamics. They distinguish areas with higher
densities of washes; especially where drainages reticulate or
have greater vegetation cover. These layers also highlight areas where mountain and alluvial landforms adjoin. The lack of influential band ratios and reflectance bands in the Updated model was unexpected. We were expecting more of the ASTER covariates to possibly distinguish mineralogy or albedo differences based on the diverse lithology sources in ORPI. It was also perplexing that both band ratio and reflectance layers were highlighted in the Original disaggregation model and the PCA data reduction done by Nauman (2009), but not in the Updated model where more topography variables were introduced. However, examination of the original SSURGO map units reveals that soil series in ORPI were often mapped across areas sourced from multiple types of lithology, possibly indicating that mineralogy was not influential in distinguishing map units. This is supported by the SSURGO mineralogy classes, which is mixed for all soils in ORPI except for Perryville, which was attributed as carbonatic (USDA-NRCS, 1972).

We found that our model covariate importance values shared considerable similarities to the PCA-driven unsupervised soil-landscape classification in ORPI done previously (Nauman, 2009). Our Original disaggregation model used the same variables in the random forest classification as used by Nauman (2009) to see if similar variables would be used more frequently in the random forest. The previous study used a PCA-based approach to try to identify the most useful variables for soil mapping without any a priori knowledge of an area from a large suite of possible DEM and ASTER variables. The similar selection of variables by our Original random forest disaggregation model and the PCA-based method used by Nauman (2009) seems to confirm that the PCA-based approach can help select covariates from large datasets effectively for initial soil mapping at a site.

Future Efforts
The success with these general methods for disaggregating CSM at ORPI and in WV (Nauman and Thompson, 2014) seem to demonstrate that a consistent general approach can be taken to updating CSM around the United States. The key to this method is finding suitable initial variables in raster format to match with soil-landscape descriptions published in soil surveys to properly train a model. We would point out that this does not need to be limited to the terrain metrics used in ORPI and WV. There are vegetation and geologic attributes in SSURGO that could also be matched to imagery or other data sources for training. The main differences between the WV and ORPI studies included: (a) using different initial rule matching variables, (b) inclusion of hillslope position descriptions (e.g., footslope, backslope, shoulder, etc.) in addition to the geomorphic table in the WV study, (c) implementation of a full random forest algorithm in ORPI rather than just a classification tree ensemble, and (d) inclusion of a larger set of covariate rasters in ORPI. The use of the random forest model in ORPI was more appropriate given the larger number of covariates being used (Breiman, 2001).

Based on the higher accuracies at ORPI, updating the work in WV to incorporate more variables and a random forest framework might help results there. However, both disaggregation studies showed similar accuracies to the original CSM from which they were derived; which might indicate that the slight differences in methods were not as important as the original CSM accuracy in the reported differences in disaggregation accuracy. There are many factors that might influence these original CSM accuracies, but the scale of soil variation and the actual mapping scale are likely responsible for this in large part. The dominant soils in ORPI follow more contiguous patterns of alluvial sediments that might have less intrinsic variability than the forested and highly dissected plateaus and mountains in the WV study. We think the general workflow presented here and in Nauman and Thompson (2014) offers an opportunity to both improve and harmonize large CSM databases into more useful modern data products.

Acknowledgments
We wish to acknowledge Sue Rutman and the rest of the ORPI staff for their insight and help, as well as Chris Bertrand and Corrie Hannah for dedicated assistance in the field. Portions of this research were supported by USDA-NRCS Cooperative Agreements No. 68-9457-8-466 and No. 68-7482-12-535 with Dr. Craig Rasmussen, as well as Arizona Agricultural Experiment Station ARZT-1367190-H21-155. Funding supporting Travis Nauman at the University of Arizona came through USDA-NRCS Cooperative Agreement No. 68-3A75-4-100. Additional funding was provided by USDA-NRCS Cooperative Agreements No. 68-7482-11-527 and No. 68-7482-9-511 with Dr. James Thompson. Scientific contribution No. 3193 from the West Virginia Agricultural and Forestry Experiment Station, Morgantown, WV.

References


