Fuzzy Disaggregation of Conventional Soil Maps using Database Knowledge Extraction to Produce Soil Property Maps

T.W. Nauman  
West Virginia University, Morgantown, West Virginia, USA; US Department of Agriculture – Natural Resources Conservation Service - Geospatial Research Unit, Morgantown, West Virginia, USA

J.A. Thompson  
West Virginia University, Morgantown, West Virginia, USA

N. Odgers  
West Virginia University, Morgantown, West Virginia, USA

Z. Libohova  
US Department of Agriculture – Natural Resources Conservation Service – National Soil Survey Center, Lincoln, NE, USA

ABSTRACT: In efforts to create more detailed raster soil maps in areas with sparse point data, legacy conventional soil maps are often the main source of soil-landscape information. In the United States, the Soil Survey Geographic Database (SSURGO) is the most detailed conventional soil map. SSURGO often combines multiple soil types into one polygon, and lacks statistical design desirable for modeling. However, SSURGO does relate soil type ‘components’ to terrain positions, parent materials, and other soil forming factors which can provide a framework to disaggregate polygons into soil type maps. Expert rule sets and fuzzy membership methods were used to model SSURGO components spatially to produce continuous soil property maps. Efforts in West Virginia, USA resulted in seamless maps that produced more accurate watershed estimates of soil organic carbon, but lacked prediction ability in field scale maps of rock content and soil organic carbon.

1 INTRODUCTION

Efforts to produce high resolution soil property data have included attempts to enhance or disaggregate conventional soil maps to produce new more detailed raster products (Bui and Moran, 2001; Yang et al., 2011). Conventional soil maps in the United States (US) differ from new modern needs in that they are: 1) in a vector format, and 2) were originally geared towards land management rather than quantitative assessment of specific soil properties, and 3) have inconsistencies in data produced on different dates. In the US, SSURGO is the most detailed conventional soil map available widely. This database has become a source for soil property data inputs into various models (e.g. Causarano et al., 2008). Concerns have been raised about the use of SSURGO to estimate properties for use in models. This paper examines using a semi-automated map-unit disaggregation for producing more model-appropriate soil property raster maps from SSURGO data. We hypothesized that information in the SSURGO database and the spatial data could be translated into a model that matched environmental raster layers (e.g. digital elevation map, slope, curvatures, etc.) to soil type for use in making improved raster versions of SSURGO data. The premise of the model created was to emulate SSURGO data to the greatest detail available in the database in a seamless raster format.

2 METHODS

2.1 SSURGO Disaggregation

The SSURGO dataset consists of a polygon format vector map attributed with a map-unit label, a relational database with information that connects the map-units to information about the soils and survey area, and a hardcopy survey manuscript that is published for survey project areas (usually counties). The first challenge of this data is the need to ‘disaggregate’ the polygon map-units into the component ‘soil series’ types that exist within map-unit delineations. The mapping infrastructure in SSURGO includes multiple types of map-units that can have generally one to four soil series components as well as ‘inclusions’ of other non-named soils or non-soil areas. Components within a map unit can have different property distributions and thus must be mapped as separate entities to represent the true theoretical spatial distribution of soil properties within delineations.
In SSURGO, each component of every map unit has information regarding soil properties, from expert estimates or representative pedons depending on age, and environmental context attributed to it. By querying the geomorphic and landform related attributes in SSURGO, soil-landscape relationship descriptions were extracted from the database to help determine where in a map unit a component should occur (e.g. series x exists on the upper third of mountain flanks). The language in these queried descriptions was then matched to values in environmental rasters that represent hillslope position (Hatfield, 1996), landforms (Schmidt and Hewitt, 2004), terrace height (relative elevation with reference to local minimum in alluvial map-units), percent slope, elevation, aspect, and catchment areas (Tarboton, 1997). As an example, the descriptor ‘upper third of mountain flanks’ was linked with a zero to one-hundred hillslope position index raster by specifying that the soil exists on hillslope indexes between 66 and 95. This essentially translates database soil-landscape relationship records to environmental raster values. The rules were constrained within each map unit for each component. No lithology spatial data was available for this study to use for matching with parent materials listed in SSURGO.

Once a list of rules was created for all the components in the model area, training areas were created for all components. Rulesets for all components recorded as the same soil series were combined to create a soil series training area. Each soil series training area was created separately and allowed to overlap other soil series’ training areas because it was assumed that soils change in continuum, and thus the domain of soil classes may overlap through transitions. Each soil series training area was then used in a single-class maximum likelihood supervised classification (i.e. logistic framework) in SAGA GIS. A more exhaustive set of continuous environmental rasters were used for the classification. These consisted of slope position (Hatfield, 1996), percent slope, aspect, planar curvature, profile curvature, base-ten logarithmic scaled catchment area using the ‘dinf’ algorithm (Tarboton, 1997), elevation, relative height (relative elevation with reference to local minimum), and UTM x-coordinates and y-coordinates as rasters to help constrain spatial patterns.

The supervised classification module in SAGA-GIS was used to create a single-class maximum likelihood probability surface for each soil series separately. This created a probability score for every pixel indicating degree of membership that a pixel has in a given soil series. These membership rasters for each soil series were then used as weights for calculating soil property rasters based on soil property profiles for each soil series.

2.2 Modeled Soil Property Calculations

Soil series property values were queried from the SSURGO database (http://soildatamart.nrcs.usda.gov) by horizon and then standardized by an equal area depth spline (Bishop et al., 1999). The percent volume of rock particles greater than two-millimeters in size and kilograms of soil organic carbon per square meter in the top 100 centimeters were chosen as properties to examine. Rock content was adjusted by the spline function to the GlobalSoilMap.net specified depths of 0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 centimeter bins (see www.globalsoilmap.net: Specifications). SSURGO queries of volumetric rock content were found not to match the hard copy manuscripts. The SSURGO values were found to be much lower than the representative profile values found in the manuscripts, and after consulting local studies and scientists, the manuscript values were used. Soil organic carbon was calculated from SSURGO organic matter values by converting using a divisor of 1.724 (Soil Survey Laboratory Staff, 1996). Organic carbon values were then converted to mass using the SSURGO dry bulk density representative estimate and also adjusted for rock content using the manuscript volumetric rock content values.

Rasters of both rock and organic carbon were made using the top three probability weighted soil series property profiles from the disaggregation for all the depth increments specified above. All the mapping was done snapped to the 1-arc second national elevation dataset (NED) published by the US Geologic Survey after projection into UTM coordinates for modeling purposes. All modeling steps were done to a study area including Webster and Pocahontas counties in West Virginia (two survey project areas). A raster was also made using original SSURGO data to calculate organic carbon (kg/m² for top 100 cm) for comparison purposes based on the methods of Bliss et al. (1995). Map unit property values were calculated by taking an area weighted mean of component property values. The percent of a map unit occupied by a component was used as the weight value for the mean.

2.3 Study Area and Validation Data

Webster and Pocahontas counties in West Virginia, US were chosen as a trial area for these procedures. These two counties comprise a 388,000 hectare area of east central West Virginia (Fig. 1). These span the eastern Allegheny Plateau and Mountain as well as Northern Appalachian Ridge and Valley Major Land Resource Area as defined by the US Department of Agriculture Natural Resource Conservation Service (USDA-NRCS). The area consists of a highly dissected mountainous plateau on the west side of the Appalachians and part of the folded long
ridge-valley complexes of the central Appalachians on the east.

A 91 pedon dataset in a sub-watershed was also used to validate part of the raster model (Fig. 1). These pedons were allocated in a stratified random sample in the Upper Gauley River Watershed and were characterized for an independent project not related to soil survey (Roeker, 2011). Pedon soil organic carbon values were analyzed by a dry combustion Leco Truspec CHN elemental analyzer. Pedon rock values were estimated in the field by horizon and bulk densities were calculated by a pedo-transfer function created in a study of an adjacent watershed (Roeker, 2011).

Figure 1. Top frame illustrates position of the combined Webster-Pocahontas counties disaggregation model area in West Virginia. Bottom frame shows the spatial distribution of the

3 RESULTS

The updated maps produced by the disaggregation model (DM) created a much smoother surface than the original SSURGO maps (Fig. 2). However, when compared with the ground truth pedon data used for validation, the different property models fit the pedon values poorly. This general result shows that the concept of the model emulates SSURGO well, but the data content emulated from SSURGO may be questionable for the purposes of detailed property estimation.

Figure 2. Maps showing the disaggregated model of SOC (top) and the SSURGO map unit based estimates (bottom).
mates were -2.40 to -5.44 kg/m² different than the validation pedons (95% Confidence Interval of the difference between SSURGO estimates and pedon values). In contrast, the DM estimate differences were centered close to zero (mean of 0.44) and averaged from -2.06 to 1.18 (95% Confidence Interval), indicating that on average, the population means were not different (Fig. 3). Both the DM and SSURGO estimates of SOC showed no fit with the pedon estimates (Fig. 4), which means that neither of the maps were capturing the landscape-level spatial variance of SOC.

Table 1. Root Mean Squared Error (RMSE) values for various predicted properties when compared with pedon values.

<table>
<thead>
<tr>
<th>Variables</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOC kg/m²</td>
<td>Disaggregation Model</td>
</tr>
<tr>
<td></td>
<td>SSURGO Model</td>
</tr>
<tr>
<td>0-5 cm</td>
<td></td>
</tr>
<tr>
<td>5-15 cm</td>
<td></td>
</tr>
<tr>
<td>15-30 cm</td>
<td></td>
</tr>
<tr>
<td>30-60 cm</td>
<td></td>
</tr>
<tr>
<td>60-100 cm</td>
<td></td>
</tr>
<tr>
<td>100-200 cm</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Mean and 95% CI intervals of error differences (predicted – pedon values) for various predicted properties.

Rock estimates made with the DM model showed a tendency to systematically overestimate volumetric rock content. This tendency increases with depth, and is statistically significant for all depths except the 0-5 centimeter depths (95% Confidence Interval of the differences between DM model and pedon rock values). Figure 3 shows this trend and how the depth ranges greater than 60 centimeters overestimate rock noticeably more than the shallower depths. This might be due to an over estimation of bedrock at these depths because bedrock was recorded as 100% rock content for modeling purposes. Rock predictions with the DM also showed no fit with respect to the actual pedon data indicating a lack of ability in the DM model to capture landscape-level spatial variance of rock content.

Figure 4. Scatterplots of Disaggregation Model (DM) versus pedon 0-100 cm SOC (top) and SSURGO versus pedon data (bottom).

4 DISCUSSION

From the SOC and volumetric rock comparisons, the SSURGO data and SSURGO derived DM data both seem to lack the ability to represent property variability on the ~28 meter grid scale used for modeling. However, we do see that with the DM model the distribution of SOC estimates become centered around the pedon estimates which might indicate that this model could be generalized to a coarser spatial scale and predict SOC accurately. The integration of multiple SSURGO survey areas (i.e.
Webster and Pocahontas Counties) could be improving the population estimates of SOC in the Gauley watershed (where the validation pedons were located) when compared with conventional SSURGO data. The Gauley validation pedons spanned across the two surveys which supports this premise. More work with an expanded area and additional datasets would be required to test these possibilities systematically.

These results indicate that trying to emulate the SSURGO mapping concepts with disaggregation might not produce accurate soil property maps at a landscape scale, but perhaps might support more regional estimates of organic carbon stocks. However, more trials in different regions must be attempted before this can be generalized beyond the small validation area used in this survey. Even if these results are shown to be a systematic pattern across SSURGO, opportunities for using SSURGO along with other training data still need to be further explored. The DM model presented here might offer a more automated approach to making SSURGO consistent or more ‘harmonized’ across the vast areas it covers in the US.

The expert rulesets created in the training stage of these models also represent a substantial body of information to be explored. The specificity of soil series in both the geographic and soil property domains need to be examined more thoroughly and systematically at this stage in any attempt to harmonize SSURGO to produce seamless products. An analysis of these training areas to reveal possible needs to switch soil classifications around to make soil-environment relationships and property maps more consistent is a subject for future work. The large size of SSURGO makes this task daunting, and even just this two-county area requires thousands of conditional logic statements strung together in code to select training areas.

It must also be stressed that SSURGO was not designed to be a quantitative soil property map, and so these results were somewhat expected. SSURGO was created as a management and land planning aid, and it has served that purpose well. However, SSURGO has also been used much more in recent years by scientists in quantitative models that require accurate soil property data. The results presented here reinforce the premise that SSURGO was not designed for quantitative modeling, and scientists using SSURGO data in models need to be extremely thorough in validating results with independent data sources.

5 REFERENCES


