Geoderma 213 (2014) 385–399

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees

Travis W. Nauman, James A. Thompson *

Division of Plant and Soil Sciences, West Virginia University, United States

ARTICLE INFO

Article history:
Received 21 February 2013
Received in revised form 31 July 2013
Accepted 16 August 2013
Available online xxxx

Keywords:
Polygon disaggregation
Spatial disaggregation
Soil–landscape models
Classification trees
Uncertainty
Variable importance

ABSTRACT

Disaggregation of conventional soil surveys has been identified as a potential source for much of the next generation of model-ready digital soil spatial data. This process aims to apportion vector soil surveys into raster (gridded) representations of the component soils that are often aggregated together in map unit designs. Most soil surveys are published with some description of the soil–landscape relationships that distinguish component soils within map units. We used these descriptions found in the Soil Survey Geographic (SSURGO) database of Webster and Pocahontas Counties in West Virginia, USA, to build a set of representative training areas for all soil components by using 1-arc second digital elevation data and derived geomorphic indices. These training areas were then used in classification tree ensembles with a more extensive environmental database to transform the original SSURGO map into a gridded soil series map. We created underlying prediction frequency surfaces from the models that can be used for creating continuous representations of soil class and property distributions.

Disaggregation models matched training sets in 71%–74% of pixels and matched components in original SSURGO map units in 56%–65% of the study area. We evaluated both the original SSURGO data and our models using 87 independent pedons not used in model building. Validation pedons matched components in SSURGO map units at 39% of sites, but in map units that only included one named component (as opposed to multiple soils that could be matched to validation pedons) only 27% of the sites matched. Disaggregation predictions matched validation pedon classes 22–24% of the time using nearest neighbor spatial matches, and these rates increased to 39–44% for correct predictions within a 60-meter radius of the pedon. To characterize uncertainty, we compared relative ensemble prediction frequency (probability) of final hardened model classes at validation sites. Sites with correct predictions had generally higher prediction frequencies; which lead us to use them to create an uncertainty model. Uncertainty was calculated by determining the rate of correct predictions at validation sites within different intervals of prediction frequencies using nearest neighbor validation results. We were able to discern four uncertainty classes with values of 7%, 18%, 20% and 43%, which we called "ground truth probabilities". We present the methods to create these models as a specific example of how disaggregation techniques may be used to aid in updating national soil survey inventories.

1. Introduction

Soil properties and soil functions influence many of the problems facing society today. Soil is a primary storage mechanism for carbon and nutrients that control how vegetation grows and how our climate is changing. However, our knowledge of soils is imprecise, with estimates of global soil carbon stocks in the top meter of soil that range from 1400 to 3250 petagrams (Grunwald et al., 2011). In light of the projected challenges of global warming and maintaining natural resource services such as crops and clean water (IPCC, 2007), high quality soil information is key to making sustainable decisions. Although many soil inventories have been carried out around the world, the scope and spatial structure of these have been criticized for having, among other things, relatively coarse resolution and map legends that are difficult to interpret (Burrough, 1989; Burrough et al., 1997; Grunwald, 2009; Grunwald et al., 2011; McBratney et al., 2003). These issues are exacerbated as more and more researchers use soil data in environmental, agricultural, hydrological, and engineering models. Many studies have attempted to improve on past soil inventories using digital soil mapping and related methods (Bui and Moran, 2001; Bui et al., 1999, 2006, 2009; Cook et al., 1996; de Bruin et al., 1999; Hansen et al., 2009; Häring et al., 2012; Kempen et al., 2009; Kerry et al., 2012; Moran and Bui, 2002; Nauman et al., 2012; Thompson et al., 2010; Yang et al., 2011; Zhu, 1997; Zhu et al., 1997, 2001). The *GlobalSoilMap* project (www.globalsoilmap.net) is a recent effort to help produce standard functional soil property maps for the whole world that can be used in more

* Corresponding author at: Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV 26506-6108, United States. Tel.: +1 304 293 2921; fax: +1 304 293 2960.
E-mail address: james.thompson@mail.wvu.edu (J.A. Thompson).

0016-7061/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2013.08.024

modern contexts (Hartemink et al., 2010; Sanchez et al., 2009). The *GlobalSoilMap* consortium has recognized that methods to best utilize old maps for production of new digital models is one of the best ways to begin creating new and more detailed soil maps (Minasny and McBratney, 2010).

One of the main challenges to providing appropriate data is that the classic paradigm of soil survey is management based, and properties attributed to soils are most often estimates based on sparse data at representative locations, not quantifications based on statistics. A large part of the goals of the original design of these maps was to determine suitability or hazards to human activities. These interpretations provide pragmatic initial guidance to developers, farmers, and other land management institutions for issues like road building, septic tank evaluations, and many other uses (Soil Survey Division Staff, 1993). The soil survey was supposed to be a starting point in planning and general management, but more current users have stretched far beyond these original concepts (Bouma, 1989; Soil Survey Division Staff, 1993).

Many studies have used soil survey spatial data with property estimates as inputs into models (e.g., Bandaragoda et al., 2004; Causarano et al., 2008; Gatzke et al., 2011; Lineback Gritzner et al., 2001; Wilson et al., 1993; Zhang et al., 2011). In the U.S., both the U.S. General Soil Map (STATSGO2: 1:250,000 scale) and the Soil Survey Geographic (SSURGO: most commonly 1:24,000 scale, but varies between 1:12,000 and 1:125,000) databases often aggregate multiple soil classes into spatial polygon delineations used in maps (Soil Survey Division Staff, 1993; Thompson et al., 2012). The data model for SSURGO, which is the primary high resolution soil inventory for the US, includes polygonal map units with generally one to four named soil series (soil taxonomic class) per map unit, plus minor inclusions of soils or non-soil areas, which are sometimes but not always fully documented. This aggregation, and the inherently crisp breaks that polygon mapping imposes on spatial data, make spatial representation of estimated soil properties (e.g., soil texture, organic matter, and pH) somewhat convoluted and predisposed to artifacts. For example, there are often distinct changes in property values between polygons or at survey project boundaries that do not make logical sense (Loerch, 2012; Thompson et al., 2012). The problem that now emerges is how to use the wealth of information in legacy soil surveys in an appropriate way. Part of the answer might be to restructure the data to more appropriately address current applications, and one basic step to doing that is to spatially disaggregate the information into its component parts in a manner that better represents how soils truly occur in the field. This paper illustrates a technique to use widely available elevation, lithology, and remote sensing data to disaggregate two existing adjacent soil surveys in West Virginia, USA, into one continuous soil series class map using no new soil field data. This process potentially reveals much more information about spatial soil distribution and spatially harmonizes somewhat disjoint mapping projects that have artifacts along their boundaries (Nauman et al., 2012; Thompson et al., 2010, 2012).

1.1. Soil survey spatial disaggregation

The primary focus of soil survey disaggregation is to express the spatial distribution of soil individuals in cases where older soil maps have lumped them into one spatial unit (Table 1). Another way to describe it would be the enhancement of a prior generalized soil map to produce a more detailed map that spatially distinguishes soil properties or types at a greater level of detail. Generally these techniques also tend to translate the data from object-based polygon maps to grid-based raster formats by using new point or environmental maps (e.g. DEM or satellite imagery) as predictors to map within polygons. Disaggregation has been identified as a conceptual approach to translate current data into formats compatible with modern needs and with pedologic concepts of soil formation (Bui, 2004; Bui and Moran, 2001; Bui et al., 1999; de Bruin et al., 1999; McBratney, 1998; Wielemaker et al., 2001). Generally,

Table 1
Two multi-component map units recorded in the Webster County soil survey, West Virginia (Delp, 1998).

Map unit (MU) name	MU kind	Components	Parent material	% of MU
Gilpin–Laidig association, very steep, extremely stony	Association	Gilpin	Residuum	45
		Laidig	Colluvium	35
		Included soils	n/a	20
Pineville–Gilpin–Guyandotte association, very steep, extremely stony	Association	Pineville	Colluvium	35
		Gilpin	Residuum	25
		Guyandotte	Colluvium	15
		Included soils	n/a	25

approaches use new pedon data and/or environmental covariate data to determine how soils within polygon map units vary spatially. Approaches tend to draw from digital soil mapping frameworks (Grunwald, 2009; Grunwald et al., 2011; McBratney et al., 2003; Scull et al., 2003) that employ a state-factor theory of soil formation summarized by Jenny (1941).

Spatial disaggregation of multi-component soil map polygons into individual component soil classes has been demonstrated in attempts to universally update soil maps (Bui and Moran, 2001; Hansen et al., 2009; Smith et al., 2012; Wei et al., 2010), and to create class distinctions within the bounds of original survey map units (e.g. Bui and Moran, 2001; Häring et al., 2012; Thompson et al., 2010). Other studies have looked at disaggregating polygons for specific soil properties using conventional soil survey. Goovaerts (2011) evaluated geostatistical methods that can combine point data with choropleth data to look at within-polygon variation in 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 SoLIM (Qi et al., 2006; Zhu, 1997; Zhu et al., 1996, 2010) to help organize and implement soil–landscape relationships for mapping soils. SoLIM 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 used in 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), and combined landform element maps with other ecological and environmental covariate maps to create ecosystem maps that incorporate soil information (MacMillan et al., 2007). Classification and regression trees have been a prominent technique used in disaggregation. Bui and Moran (2001) and Wei et al. (2010) both used ensembles of decision trees and Häring et al. (2012) used random forests to refine soil and surficial geology classes. Tree-based models have also been used extensively in general digital soil mapping applications and seem to have the greatest flexibility of common modeling methods (Behrens et al., 2010a,b; Moran and Bui, 2002; Bui et al., 2009; Lemercier et al., 2011; McKenzie and Ryan, 1999; Saunders and Boettinger, 2007; Schmidt et al., 2008; Scull et al., 2005).

The objective of this research was to identify a pragmatic and repeatable method for systematic disaggregation of legacy soil maps. This technique addresses the common situation where an older soil map is available, but more detailed soil spatial data is needed, and too few new soil observations are available to use in geostatistical approaches or for building empirical models. We utilize soil–landscape rules that are usually present in soil survey database map unit descriptions in combination with a classification tree ensemble with different randomization schemes to universally disaggregate two adjacent soil survey projects into one harmonized soil series map. This approach captures both implicit and explicit expert knowledge about soil–landscape relationships in SSURGO and pairs that with available elevation, imagery, and geology data in a classification tree ensemble model. We selected methods and data sources based on repeatability, transparency, and