Geoderma 214 (2014) 91–100

Contents lists available at ScienceDirect

Geoderma

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

Disaggregating and harmonising soil map units through resampled classification trees

Nathan P. Odgers [a,*], Wei Sun [b], Alex B. McBratney [a], Budiman Minasny [a], David Clifford [c]

[a] Department of Environmental Sciences, Faculty of Agriculture and Environment, C81 Biomedical Building, The University of Sydney, New South Wales 2006, Australia
[b] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[c] CSIRO, Division of Computational Informatics, CSIRO EcoSciences Precinct, 41 Boggo Road, Dutton Park, Queensland 4102, Australia

ARTICLE INFO

Article history:
Received 21 March 2013
Received in revised form 21 August 2013
Accepted 25 September 2013
Available online 21 October 2013

Keywords:
Spatial disaggregation
Soil map units
Classification trees
Legacy soil maps
Soil class mapping
Digital soil mapping

ABSTRACT

Legacy soil maps typically consist of a tessellation of polygon soil map unit delineations where the map units consist of a defined assemblage of soil classes assumed to exist in more-or-less fixed proportions. There are several limitations in this kind of mapping approach that relate to the original intent of the soil survey, the effect of mapping scale, and the nature of soil polygon boundaries. Yet perhaps a more fundamental limitation is the fact that most of the time, the soil classes that comprise the soil map units are not mapped individually: in effect their spatial distributions are unknown beyond the qualitative indications given in the accompanying soil map unit report. Spatial disaggregation of soil map units attempts to map the spatial distribution of the individual soil classes that comprise a legacy soil map. We developed an approach called "Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees" (DSMART). DSMART samples the polygons of a legacy soil map and uses classification trees to generate a number of realisations of the potential soil class distribution. The realisations are then used to estimate the probability of occurrence of the individual soil classes. These estimates are mapped as raster grids, which can overcome some of the limitations of mapping scale and polygon boundaries inherent in the original legacy soil map.

We demonstrate the DSMART approach on a legacy soil map from the former Dalrymple Shire in central Queensland, Australia. We mapped the estimated probability of occurrence of the 71 soil classes in the legacy soil map, as well as the most probable soil class, second-most-probable soil class and the degree of confusion between them as determined by a confusion index. Validation on 285 observed soil profiles indicated that for 48.4% of the validation profiles, the observed soil class was identified in the top three most probable soil classes.

1. Introduction

Many jurisdictions have a great store of soil class maps, which in some cases have been compiled since about the turn of the twentieth century (Kellogg, 1938). Such maps typically depict the spatial distribution of soil as a set of tessellated polygons with crisp boundaries. These polygons delineate areas of land that belong to a specific soil map unit, and each map unit is composed of a prescribed assemblage of soil classes that occur within the map unit in more-or-less fixed proportions.

Over time, rules have been developed in various soil survey programs (for example Powell, 2008; Soil Survey Staff, 1993) that govern the design of soil map units. Different kinds of map unit have been conceived; we are familiar with consociations, complexes, associations and undifferentiated groups, for example (Avery, 1973; Simonson, 1968). The choice of mapping scale affects the kind of map units that can be delineated (Valentine, 1981). Coarser scale maps require more generalised map units, which necessarily depict broader-scale patterns (Hewitt, 1993).

There are a few difficulties with a polygon representation of the soil distribution. For example, soil survey has focused more on assessing the soil resource in terms of appropriate land use than on really understanding the spatial variation of the soil distribution for its own sake. This is a pragmatic objective (Kellogg, 1950), but means that eventually such maps may be proven inflexible or inadequate in the face of new or unexpected applications.

The degree of spatial detail that can be depicted in a traditional soil map is highly dependent upon mapping scale. For example, the minimum area of land that can be legibly delineated at a mapping scale of 1:24,000 is about 2.3 ha but at a mapping scale of 1:250,000 is about 252 ha (Soil Survey Staff, 1993). A mapping scale of 1:250,000 is just too coarse to show a really fine-scale pattern with any degree of legibility, but finer-scale soil maps are more expensive and time-consuming to produce. As a result, spatial variation is often unavoidably obscured.

Due to limitations of the mapping scale and the traditional desire to map assemblages of soil fit for a common purpose, the constituent soil

* Corresponding author. Tel.: +61 2 8627 1137.
E-mail addresses: nathan.odgers@sydney.edu.au (N.P. Odgers), sunwei1028@126.com (W. Sun), alex.mcbratney@sydney.edu.au (A.B. McBratney), budiman.minasny@sydney.edu.au (B. Minasny), david.clifford@csiro.au (D. Clifford).

0016-7061/$ – see front matter
http://dx.doi.org/10.1016/j.geoderma.2013.09.024

classes of soil map units are not usually mapped individually. Often the accompanying map report includes block diagrams of modal landscapes or qualitative descriptions of the typical landscape setting of the individual soil classes; however—even when standard terminology or style is used—descriptions can be vague or open to interpretation. Modellers criticise these kinds of maps because of the difficulty in parameterising the soil property distribution for a particular area.

A further difficulty with the polygon representation of the soil distribution is the oft-cited fact that the soil does not usually change abruptly at the polygon boundaries (Burrough et al., 1997; Greve and Greve, 2004). This is somewhat *cliché* but nevertheless true. Nowadays other technologies allow us to represent the spatial variation more faithfully.

So from a scientific point of view, there seems to be a few ways that we can improve upon this traditional representation of soil variability. The technology is now available to enable us to map individual soil classes at a high level of detail. In addition the raster data structure allows us to model a numerical spatially continuous variation more flexibly than polygons can.

1.1. Spatial disaggregation

There is a small yet growing corpus of soil science literature concerning itself with the spatial disaggregation of soil map units. Disaggregation involves the downscaling of information to produce new information at a finer scale than the original source (McBratney, 1998). In (digital) soil mapping, the aim of spatial disaggregation of soil map units is to map the constituent soil classes of soil map units individually (Thompson et al., 2010). Several techniques have been demonstrated in the last couple of decades. Many of them seek to emulate the kind of logic behind traditional approaches to soil survey.

Lagacherie et al. (1995) did not concern themselves with downscaling per se but their ideas were not dissimilar to contemporary disaggregation techniques. They quantitatively and probabilistically extrapolated soil pattern rules from a reference area to a small physiographic region in southern France.

Thompson et al. (2010) manually recovered soil-landscape rules from a soil map report. They obtained good results but the technique is likely time-consuming for large areas. Smith et al. (2012) used a weights-of-evidence approach coupled with fuzzy inference using SIE (Shi et al., 2009). Other techniques include supervised classification (Nauman et al., 2012), decision trees (Bui and Moran, 2001; Bui et al., 1999; Hansen et al., 2009; Häring et al., 2012; Subburayalu et al., 2014), hybrid Boolean- and fuzzy-logic-based classification rule techniques (MacMillan et al., 2007), Bayesian techniques (Bui et al., 1999), multinomial logistic regression (Kempen et al., 2009), pycnophylactic splines (McBratney, 1998) and area-to-point kriging (Kerry et al., 2012).

The decision tree approaches are favourable because of their interpretability and their ability to utilise both discrete-valued and continuous-valued covariates. In addition decision trees mimic the mental processes involved in creating the legacy products in the first place (Bui and Moran, 2001; Bui et al., 1999). Yet a single decision tree applied to a raster grid of the study area yields only one result: one, hard, realisation of the disaggregated soil class distribution. Such a map effectively tells us nothing about the probability of existence of *other* soil classes in the study area for a given location.

A better approach is to be able to say how probable *any* soil class may be at a given location. One way to do this is to generate a number of realisations of what the soil class distribution *could* be. Then we could summarise these predictions to produce a probability surface for each soil class. Some researchers (for example Bui and Moran, 2001; Bui et al., 1999; Häring et al., 2012) have employed multiple models but only as a means to the end of getting at the modal or most probable soil class for each location. Other workers have only disaggregated within existing map unit boundaries (Häring et al., 2012), or have not demonstrated the ability to disaggregate all soil classes in the area of interest (Subburayalu et al., 2014).

We argue that an effective spatial disaggregation technique should at least be able to disaggregate all soil classes in the area of interest simultaneously. In most circumstances this means not limiting the predictive exercise by disaggregating only within polygon boundaries, which are drawn subjectively at a specific mapping scale. Soil classes may be found as inclusions in map unit polygons where they are not prescribed, or be absent when they are prescribed. On the other hand, it may make sense to disaggregate within map unit boundaries in circumstances where the soil-landscape relationships between map units are known to be very different. The reasons for making this choice might be dependent on the scale of the map to be disaggregated and the nature of the landscape in the area of interest.

Of course outputs such as the most probable soil class are useful; however, we argue that it is of great benefit for an effective spatial disaggregation algorithm to also explicitly report the probability or possibility or membership surfaces associated with each soil class in the area of interest. This not only provides a richer picture of the potential spatial variation of the soil class distribution but may open other, hitherto unexplored, avenues for research or application of these predictive outputs.

1.2. Aim

In this paper we present a new algorithm called DSMART which stands for Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees. We demonstrate the use of DSMART on a case study from central Queensland, Australia. In this paper we focus exclusively on disaggregation of soil map units, whereas our earlier work demonstrated the ability of spatial disaggregation to harmonise the depiction of the soil distribution across adjacent soil survey areas (Sun et al., 2010).

2. Methods

2.1. Study area

The study area comprises most of the former Dalrymple Shire in central Queensland, Australia (Fig. 1). It has an area of about 68,000 km^2 and is approximately 1000 km north of Brisbane. It comprises a large part of the northern Burdekin River catchment. The study area is bounded on the east by the Seaview and Leichhardt Ranges, the Great Dividing Range in the west, and the Suttor and Belyando Rivers in the south-east. Most of the area is flat to gently undulating and elevation generally decreases towards the south-east. It is drained by the Burdekin River and its tributaries (Rogers et al., 1999).

2.2. Overview of the DSMART algorithm

The aim of DSMART is to predict the spatial distribution of soil classes by disaggregating the soil map units of a soil polygon map. For the purpose of this analysis, some definitions are necessary. A *soil map unit* is an entity consisting of a defined set of soil classes which occur together in a certain spatial pattern and in an assumed set of proportions. Mapped instances of soil map units are often called *delineations* (Soil Survey Staff, 1993) and are the polygons of a soil polygon map. An entire soil map consists of a universe of s soil classes $S_1,\ldots,S_s$ and any S_i may be a component of many soil map units.

One way of representing the disaggregated soil class distribution is as a set of numerical raster surfaces—one raster per soil class, where the surfaces represent the estimated probability of occurrence of each soil class.

In order to generate the probability surfaces, a straightforward approach is to generate n realisations of the potential soil class distribution. Then at each grid cell, the probability of occurrence of each soil class is estimated by the proportion of times the grid cell is predicted as each soil class across the set of realisations.