

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.

© 2013 Elsevier B.V. All rights reserved.

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).

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 km² 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 $S_1, \dots, S_s$ 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.

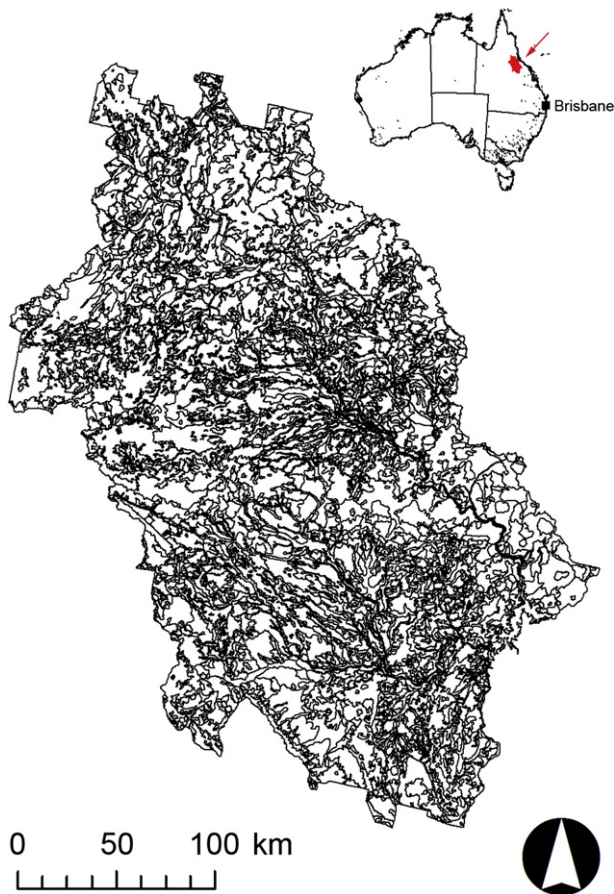


Fig. 1. Dalrymple Shire soil polygon map and location within Australia.

A schematic representation of the process carried out by DSMART is presented in Fig. 2. The individual steps are described in more detail below. DSMART is implemented in Python 2.7.3. It uses the Geospatial Data Abstraction Library (GDAL) for spatial data processing and the See5 algorithm (Rulequest Research, 2008) for generating decision trees.

2.3. Input data

2.3.1. Soil polygon map

Along with the soil polygons that are to be disaggregated, we assembled a database containing information on the soil classes that are

assumed to occur in each polygon and the proportions in which they are assumed to occur (Step 1 in Fig. 2).

The soil polygon map that we used is the 1:250,000 soil map of the Dalrymple Shire by Rogers et al. (1999); Fig. 2. It is comprised of 3058 polygons that belong to 1110 soil map units. The map units are soil associations consisting of two or more soil classes and are named after the most dominant soil class occurring within them. 72 soil classes are recognised in the study area. At the most detailed taxonomic level, each soil class has a name such as *Two Creek*, *Conolly* or *Burdekin*. The soil classes are grouped hierarchically into Orders of the Australian Soil Classification system (Isbell, 1996) and then into geological landscapes (Rogers et al., 1999).

2.3.2. *Scorpan* covariates

We assembled a dataset of georeferenced *scorpan* covariates (McBratney et al., 2003) which consisted of terrain attributes, Landsat 5 bands and gamma radiometric surfaces (Step 1 in Fig. 2). The terrain attributes were derived from the 1-arcsecond Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) and included elevation, slope, topographic wetness index (TWI) and multi-resolution valley bottom flatness (MRVBF; Gallant and Dowling, 2003). The gamma-radiometric K, Th and U layers were included because of their utility in inferring parent material and soil distribution (Wilford and Minty, 2007). The radiometric survey was flown at greater than 200-m line spacing and resampled to approximately 1-arcsecond resolution. This is a scale mismatch but the radiometrics was judged as too valuable not to be included. The 30-m-resolution Landsat 5 imagery consisted of a mosaic of 7 Landsat 5 Thematic Mapper tiles captured about July 2009. We converted the raw digital numbers of bands 1, 2, 3, 4, 5 and 7 to planetary reflectance values and calculated the Normalised Difference Vegetation Index (NDVI; e.g. Tucker et al., 1981) using bands 3 and 4. All covariates were projected to an Albers equal area projection based on the Geocentric Datum of Australia 1994. The covariates were gridded at a 30 m grid resolution. The gamma radiometric surfaces, Landsat reflectance bands and NDVI surface were reregistered to the DEM grid.

We did not use covariates that represented positional information such as easting or northing. The map unit polygons were only used to guide the sampling and not as a covariate in the mapping, for several reasons. First, by way of trial and error, when we *did* include the map unit polygons, they tended to feature very prominently in the decision trees and led to sharp discontinuities in the probability surfaces. Second, the map unit polygons were drawn at a relatively coarse mapping scale, which can have the effect of generalising the variability in the landscape. Third, there is uncertainty in the location of the map unit polygon boundaries (Rogowski and Wolf, 1994). Boundaries are drawn according to expert judgement (Webster, 1973). Misalignment errors can creep in to the digitisation process.

2.4. Generating the realisations of the soil class distribution

Each realisation is generated by calibrating a decision tree with a set of sampling points drawn from the soil polygons. The decision trees have been used in the digital soil mapping literature to predict soil class distribution (Bui et al., 2006; Qi and Zhu, 2003; Scull et al., 2005). We used decision trees because of the flexibility they allow in choosing covariates—both continuous-valued and discrete-valued components can be used (Scull et al., 2003)—and because of their ease of interpretation, but any model capable of explaining categorical variables could be used.

2.4.1. Sampling the map unit polygons

Each decision tree required its own calibration dataset. We derived each calibration dataset by drawing m random sampling points from each soil polygon (Step 2 in Fig. 2). Therefore the total number of sampling points in any calibration dataset is essentially m multiplied by the number of soil polygons in the soil map to be disaggregated.

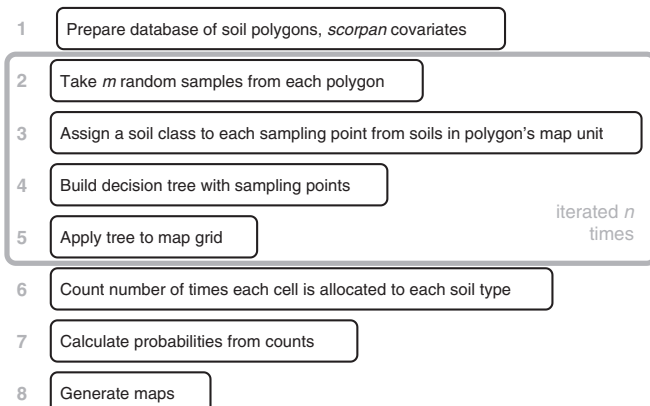


Fig. 2. Schematic overview of the DSMART algorithm.

This approach differed from our original implementation (Sun et al., 2010) where we sampled the entire study area systematically on a moving regular grid. Choosing equal numbers of points from each soil polygon places equal weight on the information associated with the polygons whereas the regular grid approach placed more weight on larger polygons than smaller ones (owing to their size, larger polygons were sampled more densely). We could also have randomly sampled within polygons differentially depending on the size of the polygon.

Although we took the same number of samples from each polygon, the map units are sampled differentially depending on the number of polygons there are per map unit (i.e. map units with many polygons contribute more samples than map units with few polygons). This is similar to the approach Moran and Bui (2002) demonstrated was superior to taking the same number of samples per map unit, although in our case the sampling points are labelled with a soil class from within the respective map units (see Section 2.4.2) rather than the map unit itself.

The acceptable sampling points were those that did not contain any null values in the covariates. The algorithm was allowed ten tries to find an acceptable sampling point to fill each of the m sampling point “slots” within each polygon. The majority of the time, an acceptable sampling point will be found on the first try; however, null values may exist if the covariate rasters do not completely overlap the soil map boundaries or if water bodies have been masked. In such cases it is difficult or even impossible to find acceptable sampling points, particularly if the areas of the polygon and internal null values are similar. If m acceptable sampling points cannot be found, the size of the calibration dataset for that polygon for this particular realisation will be less than m .

2.4.2. Soil class assignment

Step 3 (Fig. 2) involved allocating a soil class to each sampling point. It was necessary to allocate a soil class to each sampling point arbitrarily because the true soil class at each sampling point is unknown. There are several possible soil class allocation approaches:

1. A completely random allocation where all soil classes in the relevant soil map unit are weighted equally (i.e. all soil classes in the relevant map unit have an equal chance of being allocated to a given sampling point);
2. A weighted random allocation where we weight each soil class in the relevant map unit with its assumed proportion of occurrence in the map unit;
3. A targeted allocation where we seek to allocate the most likely soil class from the relevant map unit based on some kind of similarity of the landscape characteristics at the sampling point to assumed modal landscape characteristics of the relevant soil classes.

We determined that we could do better than Approach 1 because in most cases we have some idea of the expected in-map-unit proportions of occurrence of the relevant soil classes. For example, if a soil class is assumed to occupy 80% of the given map unit then, in the absence of any other information, each sampling point in that map unit should have an 80% chance of being assigned to that soil class. Hence we implemented Approach 2 in DSMART.

On the other hand if our sampling point had a slope of 20% and we knew that one or two of the soil classes in the relevant map unit could occur on a 20% slope then we could narrow down the soil class allocation a bit further, as in Approach 3. This would be suitable if we could translate some characteristics of the typical landscape setting of the soil classes into covariate values for slope, elevation and so-on. It might be especially valuable if we do not really have an idea of what the in-map-unit proportions are, such as may be the case in very coarse scale soil surveys.

In this work we use Approach 2. For each realisation, each sampling point was assigned to one of the soil classes assumed to occur in the map unit polygon in which the sampling point was located. The allocation procedure was a weighted random choice from the polygon's pool of soil classes where the weights were the areal proportion of the map unit polygon that the soil classes are assumed to occupy as reported in

the soil polygon map. However, it is well known that the within-map-unit variability can be large (Young et al., 1997), and areal proportions are often reported to the nearest 5% or 10%. As a result, we reflect the uncertainty in the reported areal proportions by employing the Dirichlet distribution, which is commonly used as the prior distribution for probabilities associated with the multinomial distributions in Bayesian mixture models. It is often denoted $\text{Dir}(\alpha)$ where α is a vector of concentration parameters $\alpha_1, \dots, \alpha_K$, where $\alpha_i > 0$ and $K \geq 2$. A draw from the Dirichlet distribution returns a vector of positive numbers $x_1, \dots, x_K$ where $x_i \in [0, 1]$ and $\sum x_i = 1$. Mathematically these can be viewed as the probabilities of a K -way categorical event.

Within each soil polygon, K soil classes are prescribed according to the polygon's map unit definition. For any polygon we set the α_i to $c_{S_i} \times T$, where c_{S_i} is the reported within-map-unit proportion for soil class S_i in the set of soil classes in the given polygon and T is a parameter that controls the variability of the x_i about the α_i . If the soil class areal proportions sum to 1 then the effect of drawing from the Dirichlet distribution is to vary probabilities about the reported soil class proportions by a small amount, such that the average x_i over a large set of draws should respect the reported c_{S_i} .

Larger values of T (e.g. 1000 or 10,000) result in less variability (effectively less uncertainty) about the c_{S_i} over many draws from the Dirichlet distribution, compared with smaller values of T (e.g. 1 or 10). Setting an appropriate value of T lets us incorporate the degree of certainty we have in the soil class proportions. For a particular choice of T , the values of x_i will be more variable when c_{S_i} is close to 0.5 than when c_{S_i} is close to 0 or 1. We set T to 10,000.

We performed a power analysis to find the value of m that gives rare soil classes a specific chance of occurring at least once in each realisation. For example, if a soil class occurs in 5% of a soil map unit then at least $m = 13$ sampling points are required to have a 50% chance of that class being present in a realisation. Fig. 3 illustrates the relationship between the chance of a soil class appearing and m for different values of c_{S_i} .

2.4.3. Generating the decision tree

For each realisation, the full set of sampling points were intersected with the georeferenced *scorpan* covariates and then used as a calibration dataset to build the decision tree relating the covariates to the soil

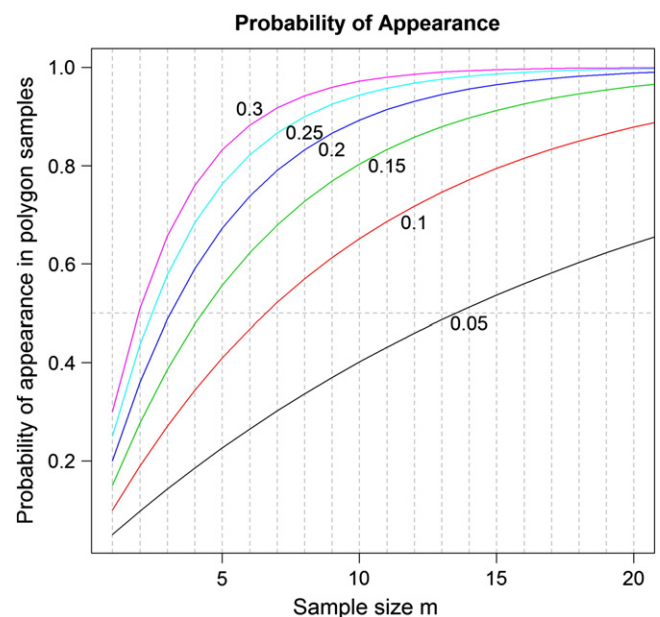


Fig. 3. The estimated probability of appearance in a set of m samples drawn randomly from within a single map unit polygon of soil classes with the given in-map-unit areal proportion.

classes in the polygon map (Step 4 in Fig. 2). We used the classification tree software See5 (Rulequest Research, 2008) to build the decision trees and then we applied each tree to the covariate rasters to generate the realisations of the soil class distribution over the study area at a 30-m resolution (Step 5 in Fig. 2).

2.5. Probability surface calculation

Steps 6–8 (Fig. 2) involved the calculation of the probability surfaces from the n realisations of the soil class distribution. At each grid cell we counted the number of times the grid cell was predicted as each soil class in the study area. The probability of occurrence of each soil class was estimated by dividing its frequency by n .

2.6. Validation profiles

285 observed profiles from throughout the study area were used as a validation dataset (Fig. 4). They were randomly selected from a larger dataset of about 2500 observed soil profiles collected from across the study area. The profiles were sampled during 1990–1994 as part of a free survey (Hewitt et al., 2008; Steur, 1961) in support of the creation of the polygon map. The profiles were located via Global Positioning System with an accuracy of about 100m (Rogers et al., 1999). All profiles were allocated to a local soil type.

3. Results and discussion

We ran DSMART on the 3058 Dalrymple Shire soil polygons for $n = 100$ iterations at $m = 15$ sampling points per polygon and produced probability surfaces for the 72 soil classes. We also calculated a map of the most probable and second-most-probable soil classes, their probabilities, and the confusion index between them. Examples of the probability maps are shown in Fig. 5.

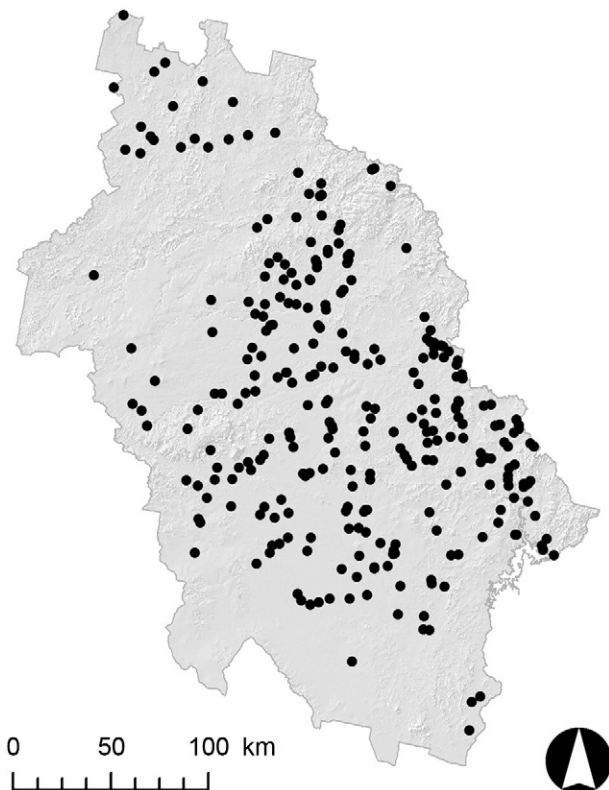


Fig. 4. Locations of the 285 validation sites.

3.1. Decision tree performance

A summary of the covariate usage across all 100 decision trees is presented in Table 1. The covariate usage for a given covariate is the proportion of leaves in a given decision tree which require the covariate at some point in one of the if-then rules descending from the root of the tree. Elevation and gamma radiometric K were used 100% of the time in all the decision trees. Gamma radiometric Th and U, and MRVBF, SAGA wetness index and Landsat 5 NDVI were also heavily used on average.

The SAGA wetness index had the largest standard deviation of usage meaning some trees used it heavily but others did not, but other covariates were used more frequently (e.g. elevation, gamma radiometric K) or less frequently (e.g. slope aspect, plan curvature).

Table 2 summarises the performance of the 100 decision trees. The trees were all very similar in their size and error statistics. The mean error rate was 19.67%. The number of sampling points varied very slightly due to nodata in the covariates at the edges of the study area (see Section 2.4.1). This was an artefact of clipping the covariate rasters to the polygon boundaries and could have been avoided if the covariate rasters were not clipped so tightly.

3.2. Probability surfaces

A selection of probability surfaces for soil classes is presented in Fig. 5. They demonstrate that for a given soil class, P_i was only greater than 0 over limited areas of the map. This was contrary to our expectations, as we thought that the soil classes might have a very small probability of occurrence across most of the area and larger values only in limited areas. Most grid cells typically had $P_i > 0$ for about 8 to 12 soil classes and $P_i = 0$ for all the other soil classes.

3.3. Most probable soil class

Each location where predictions are made has a *probability series* consisting of the most probable soil class and its probability, the second-most-probable soil class and its probability, and the third-most-probable soil class and its probability, and so-on until all soil classes in the study area are exhausted or the cumulative sum of probabilities equals 1. An example of a probability series is presented in Fig. 6.

We mapped the most probable and second-most-probable soil classes (Fig. 7), their probabilities (Fig. 8), and the confusion index between their probabilities of occurrence (Fig. 8). The confusion index (CI) was calculated as:

$$CI = 1 - (P_{\max} - P_{\max-1})$$

where $P_{\max}$ is the probability of the most probable soil class and $P_{\max-1}$ is the probability of the second-most-probable soil class (after Burrough et al., 1997). The CI encapsulates our degree of confusion about what is the “correct” soil class given—in this case—the most probable and second-most-probable soil classes: if the difference in probability between the most probable and second-most-probable soil class is small (say 0.1), we are more confused about which is the “correct” class than if the difference in probability between the most probable and second-most-probable soil class is large (say 0.9).

The probability of the most probable soil class was highest in the central west of the study area, being often above 0.75 and occasionally above 0.95. The areas of highest probability coincided with the lowest degree of confusion for the most probable soil. In these areas the value of the CI was usually less than 0.40 and occasionally less than 0.1. Throughout most of the rest of the study area the probability of the most probable soil was fairly low (often less than 0.3) and confusion was correspondingly high (often greater than 0.9). The probability of the second-most-probable soil class was usually less than 0.15 but did get as high as 0.5.

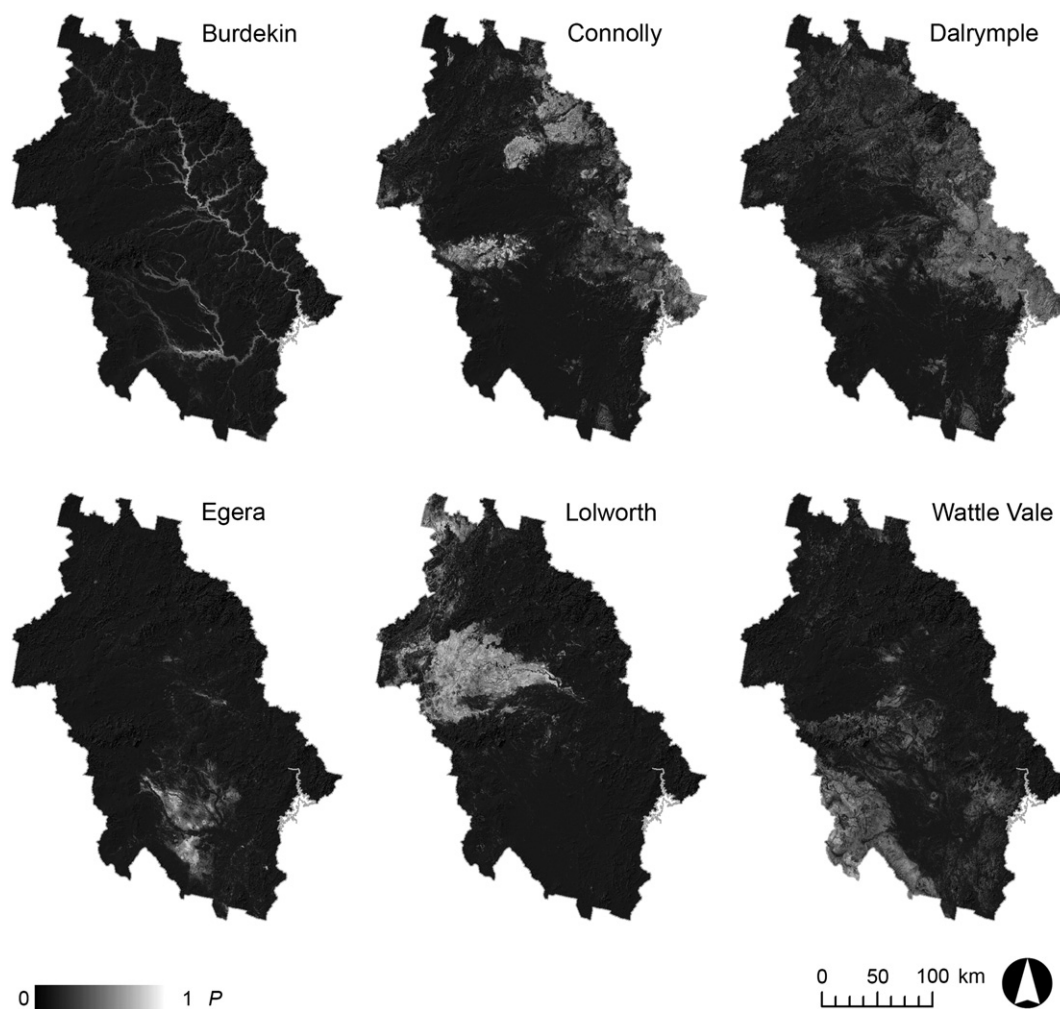


Fig. 5. Probability surfaces for several of the Dalrymple Shire soil classes.

It is evident that there are relationships between what is predicted as most probable and what is predicted as second-most-probable, although the map of second-most-probable soil class is considerably

noisier throughout than the most probable soil class map. The dominance of the most probable Lolworth (LL) and Hillgrove (HG) soils on the basalt geology in the central west of the study area is particularly

Table 1
Summary of covariate usage across 100 decision trees.

Covariate	<i>n</i>	Mean usage (%)	Std. dev. usage (%)	Min usage (%)	Max usage (%)
Elevation	100	100	0	100	100
Gamma radiometric K	100	100	0	100	100
Gamma radiometric Th	100	88.66	4.68	79	98
MRVBF	100	77.92	9.16	65	90
SAGA wetness index ($t = 10$)	100	66.74	11.64	42	85
Gamma radiometric U	100	65.99	5.2	51	80
Landsat 5 TM NDVI	100	52.14	3.94	44	65
SAGA modified catchment area ($t = 10$)	100	49.85	7.45	36	79
Valley depth	100	46.4	3.84	36	59
Slope height	100	42.15	3.35	35	50
MRRTF	100	41.95	2.64	36	50
Mid slope position	100	36.53	3.42	29	46
Landsat 5 TM Band 5	100	31.6	5.7	20	43
Terrain ruggedness index	100	31.4	4.42	23	41
Landsat 5 TM Band 1	100	30.43	3.01	25	39
Landsat 5 TM Band 4	100	28.84	3.17	22	38
Landsat 5 TM Band 7	100	25.67	4.82	17	43
Landsat 5 TM Band 3	100	24.24	5.06	15	45
Profile curvature	100	23.89	2.42	19	32
Slope aspect	100	23.71	1.26	20	27
Plan curvature	100	21.7	2.42	17	29
Landsat 5 TM Band 2	100	20.51	3.42	14	34
Slope gradient	100	18.15	3.77	11	30

Table 2
Summary of decision tree performance.

	<i>n</i>	Mean	Std. Dev.	Min.	Max.
Number of calibration samples	100	45,853.9	0.9	45,851	45,855
Tree size	100	11,815.24	55.0	11,627	11,956
Errors	100	9018.4	77.2	8771	9169
Error rate (%)	100	19.67	0.2	19	20

noticeable. The most probable Lolworth soil is associated with the second-most-probable Conjuboy soil (CB) and Maryvale soil (MA). In the north-east, the most probable Conolly soil (CO) is associated with the second-most-probable Carse O'Gowrie (CG) soil. In the south-east, most probable Dalrymple soil (DA) is associated with Conolly, Bluff soil (BL), Charters Towers soil (CT) and Mingela soil (ML). In the south-west relationships are a little less clear, but the five principal soils in the most probable or second-most-probable position are Flagstone (FL), Pentland (PE), Rishton (RI), Rangeside (RS) and Wattle Vale (WV).

As we move down the probability series it is not uncommon for several soils to become equally probable (Fig. 6). Consequently the degree of confusion is high in such circumstances and the choice for the next-most-probable soil class is essentially arbitrary. In Fig. 6 the equi-probable soil classes are identified in alphabetical order. The degree of confusion increases as we move down the probability series.

That large swaths of the study area were predicted as the same most probable soil was initially surprising as we assumed that there would be more variability in the most probable soil class distribution. However, when we overlaid the soil polygons onto the most probable soil classes it was clear that the spatial distribution of the most probable soil classes was indeed more detailed than that of the original polygons (Fig. 9).

In many cases when the map unit polygons were overlaid, the within-polygon variation was not dissimilar to that prescribed by the polygons' map unit definitions. This is nice to see but should not be the main criterion determining a good or bad result: the soil class proportions defined for a particular map unit are uncertain and it has been shown that the within-map-unit variability can be large (Young et al., 1997). It is also possible that not all polygons belonging to a given map unit will possess all the soils prescribed by its map unit definition. For the landscape delineated by a given polygon, minor soils may be absent or non-prescribed soils (inclusions) present, and the proportions in which the soils occur will vary depending on the characteristics of that landscape.

3.4. Validation

Of the 285 validation points, only 22.5% were predicted correctly as the most probable soil, but this does not take into account the degree of confusion or whether the true soil was misclassified as being further down the probability series. 16.8% were predicted incorrectly as the second-most-probable soil, and 9.1% were predicted incorrectly as the third-most-probable soil for their location. Therefore the true soil class of nearly 50% of the validation sites was predicted as being in the top three-most-probable soil classes at their particular location. The

remainder of the validation profiles (51.6%) were incorrectly classified as the fourth-most-probable soil or worse.

The validation statistics may improve if locational uncertainty in the validation profiles was taken into account, given that the locational accuracy was reported as being around 100 m (Rogers et al., 1999).

The validation statistics would also improve if the decision tree models were better. A number of factors may contribute to the quality of the decision trees. For example, one of the most important contributors to the quality of the predictive models is probably the soil class allocation procedure for training samples, described in Section 2.4.2. Allocation approach 2 is not completely random but is not likely to adhere as closely to reality as Approach 3 potentially could. Depending on *T*, Approach 2 more or less respects the map unit's reported areal proportions at each sampling point but the variability within the *m* samples drawn from a particular polygon as a whole may not. As the reported proportions may not respect the true variability within the polygon anyway, Approach 3 may give a better picture of the true within-polygon variability.

Despite these considerations, it is clear from the results that the data mining process was able to detect real patterns in the data. A visual validation by comparing the probability surfaces of the soil classes with their respective extent maps in Rogers et al. (1999) indicated general agreement in the spatial pattern.

It is always possible that at least some of the covariates chosen to represent the *scorpan* factors do not actually contribute to the predictive models very much. That is, they are not able to explain much of the variation in the soil class distribution. As a result we culled the most poorly-used covariates from the covariate pool using the See5 performance statistics as a guide.

The spatial resolution at which the disaggregation is carried out affects the degree to which the covariates are able to capture the variation in the landscape. For example a coarser-resolution DEM may not adequately capture small terrain features which may be the niche of a particular soil class.

3.5. Unanswered questions and future work

It remains to be seen how effective DSMART is at disaggregating soil polygon maps at scales other than 1:250,000. Clearly the grid resolution at which the analysis is carried out is also important (Cavazzi et al., 2013). The choice of covariates is important too: a set of covariates that are effective in strongly dissected landscapes may not be as effective in flat plains. On the other hand strongly dissected landscapes may be disaggregated more effectively at finer scales than coarse if the detail in the topography is more adequately captured at the finer resolution. Just as in traditional soil mapping, disaggregation at coarser grid resolutions may mean that soils of small spatial extent may not be mappable.

DSMART does not yet take into account the taxonomic similarity between soil classes when making predictions. Minasny and McBratney (2007) demonstrated how taxonomic distance can be incorporated as a loss function indecision trees in order to lower the misclassification error.

The feedback from a number of testers indicated that it would be useful to be able to incorporate more expert knowledge into the soil

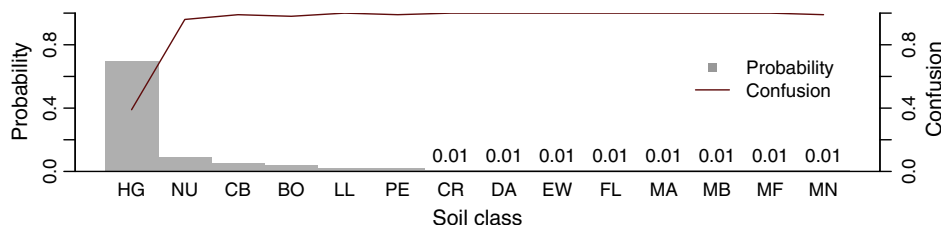


Fig. 6. Probability series for one of the validation sites. Only soil classes with $P > 0$ are shown. Equi-probable soil classes are depicted in alphabetical order. Confusion index for a given soil class is the degree of confusion between the soil class and the soil class immediately less probable in the probability series.

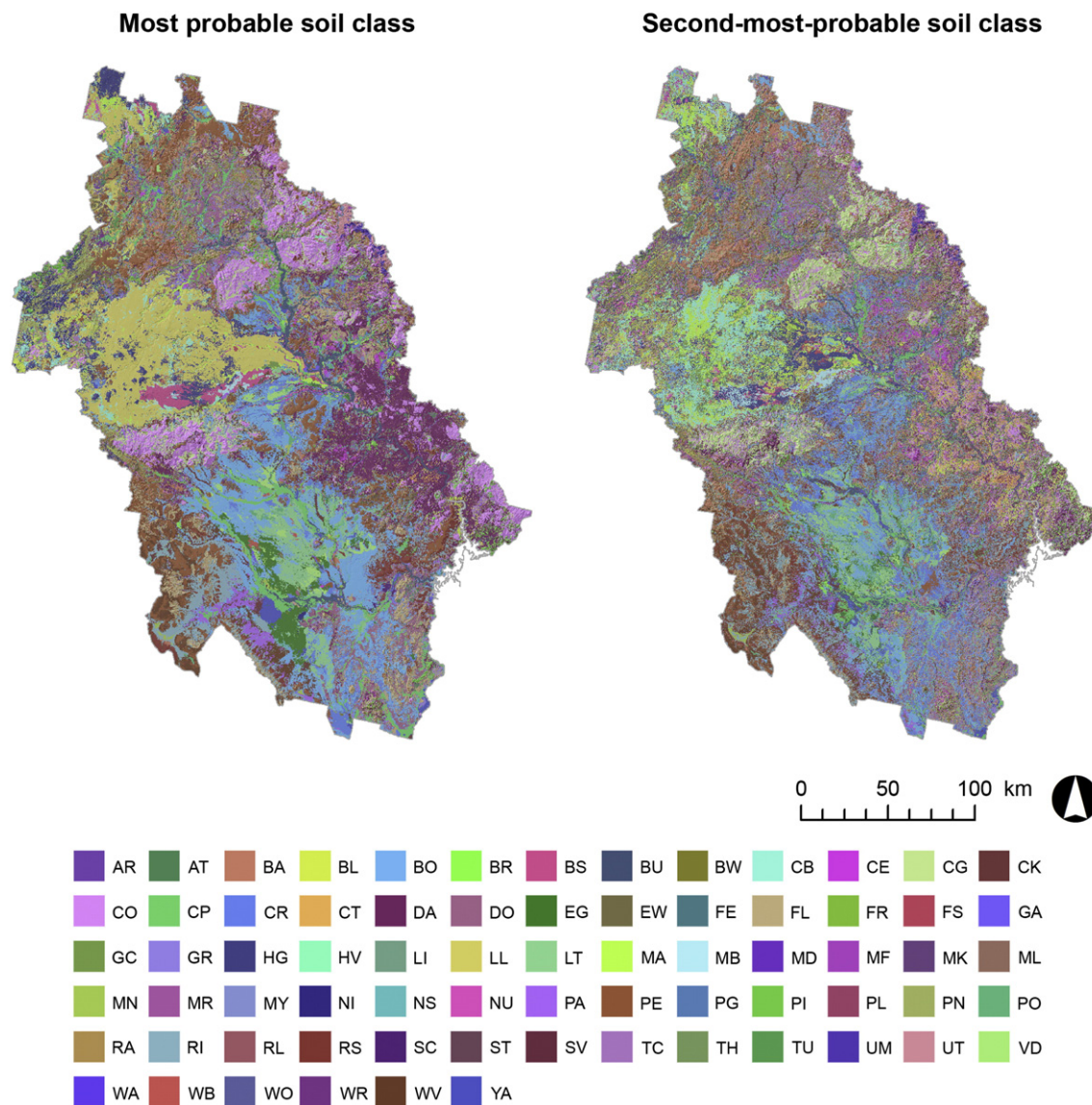


Fig. 7. Most probable and second-most-probable soil classes.

class allocation procedure, perhaps by the mechanism outlined in Approach 3 in Section 2.4.2. Soil survey reports often have much contextual information about the type location of soil classes that could be mined.

It remains to be seen whether there is an optimal number of iterations to run DSMART at and the spatial extent over which a DSMART application can be extrapolated. The predecessor to DSMART demonstrated the ability to harmonise the representation of the soil class distribution of adjacent soil polygon maps (Sun et al., 2010) but we have not yet attempted to do this with DSMART.

The effect of varying the uncertainty in the within-map-unit soil class proportions in the weighted random soil class allocation procedure by adjusting the T factor has not been extensively examined. If it turns out that there is little difference in the output then the validity of the proportions as reported in the soil survey report would come into question.

This may have implications for the disaggregation of soil polygon maps where the soil class proportions are only given qualitatively, although typical toposquences are often described in detail, indicating the spatial context of the relevant soil classes. Hird (1991) is a good example: in this report, soil class dominance within each map unit is given in terms like “dominant”, “common” and “minor” although there is a

description of the typical toposquence found in each map unit. Bui (2004) pointed out that the syntactic structure of such topossequence descriptions allows soil-landscape rules to be parsed and extracted; DSMART cannot do this. In order to apply DSMART we would have to translate these qualitative dominance terms into ordered though necessarily uncertain numerical proportions. If the value of T strongly affected the result then we would have more reason to be careful about the translation into numerical proportions than if T did not make much of a difference to the result.

Still other work relates to the usage of the output from DSMART. This is so far relatively unexplored. It should be possible to reaggregate the spatial distribution of the disaggregated soil classes into new map units. One of the most promising avenues of research concerns the generation of continuous soil property maps from the disaggregated output. The generation of maps of soil properties and their uncertainties based on DSMART output is the subject of a forthcoming paper.

3.6. Conclusions

We presented a new spatial disaggregation algorithm called DSMART and demonstrated its ability to disaggregate the polygons of a traditional soil map automatically. It is possible to disaggregate very large areas of

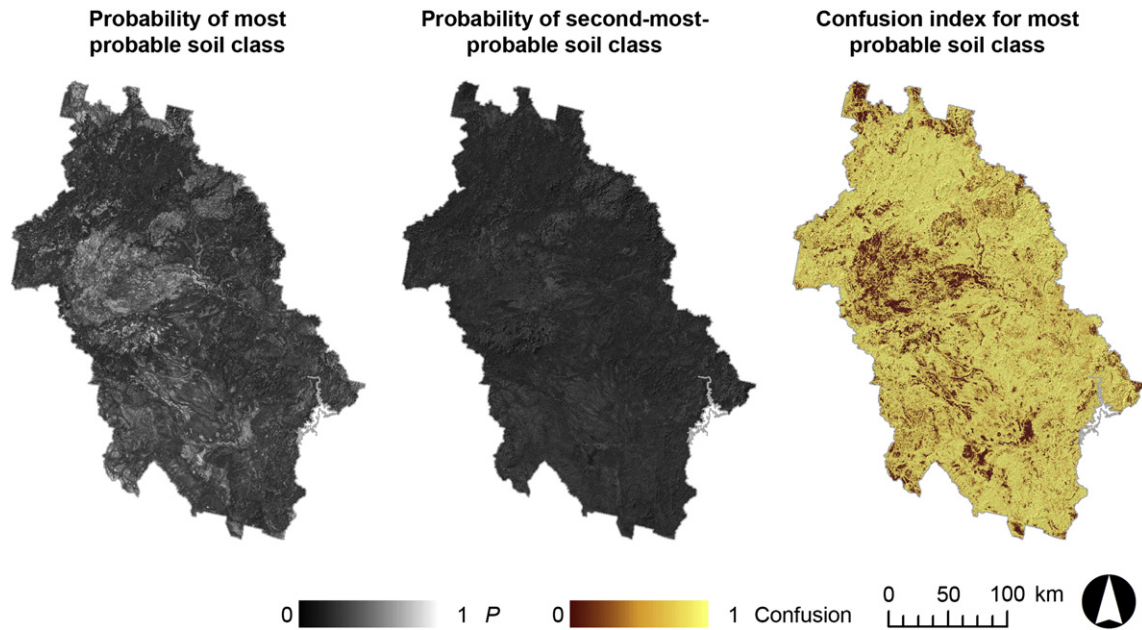


Fig. 8. Probabilities of the most probable and second-most-probable soil classes and the degree of confusion between them.

interest relatively quickly—currently in the order of a couple of days. We were able to disaggregate all the soil classes in the Dalrymple Shire area and produce probability surfaces for all soil classes. We demonstrated the advantage of being able to map the n most probable soil classes and their respective degrees of confusion. Such output may enhance our understanding of soil-landscape relationships. There is much work left to do and the opportunities for applying the outputs of an algorithm like DSMART are largely unexplored.

Acknowledgements

The authors gratefully acknowledge the provision of the soil data for this work from the Queensland Department of Science, Information Technology, Innovation and the Arts. We would like to acknowledge

Melissa Dobbie, Linda Gregory, Karen Holmes and Ross Searle at CSIRO for their feedback and help in road-testing DSMART. This work was funded by the Australian Government's Terrestrial Ecology Research Network (TERN) Soil and Landscape Grid of Australia Facility.

References

- Avery, B.W., 1973. Soil classification in the soil survey of England and Wales. *J. Soil Sci.* 24 (3), 324–338. <http://dx.doi.org/10.1111/j.1365-2389.1973.tb00769.x>.
- Bui, E.N., 2004. Soil survey as a knowledge system. *Geoderma* 120, 17–26. <http://dx.doi.org/10.1016/j.geoderma.2003.07.006>.
- Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. *Geoderma* 103, 79–94. [http://dx.doi.org/10.1016/S0016-7061\(01\)00070-2](http://dx.doi.org/10.1016/S0016-7061(01)00070-2).
- Bui, E.N., Loughhead, A., Corner, R., 1999. Extracting soil-landscape rules from previous soil surveys. *Aust. J. Soil Res.* 37, 495–508. <http://dx.doi.org/10.1071/S98047>.
- Bui, E.N., Henderson, B.L., Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. *Ecol. Model.* 191, 431–446. <http://dx.doi.org/10.1016/j.ecolmodel.2005.05.021>.
- Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma* 77, 115–135. [http://dx.doi.org/10.1016/S0016-7061\(97\)00018-9](http://dx.doi.org/10.1016/S0016-7061(97)00018-9).
- Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma* 195–196, 111–121.
- Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resour. Res.* 39 (12), 1347–1360. <http://dx.doi.org/10.1029/2002WR001426>.
- Greve, M.H., Greve, M.B., 2004. Determining and representing width of soil boundaries using electrical conductivity and MultiGrid. *Comput. Geosci.* 30 (6), 569–578.
- Hansen, M.K., Brown, D.J., Dennison, P.E., Graves, S.A., Bricklemeyer, R.S., 2009. Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data. *Geoderma* 150 (1–2), 72–84. <http://dx.doi.org/10.1016/j.geoderma.2009.01.013>.
- Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., Schröder, B., 2012. Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils. *Geoderma* 185–186, 37–47. <http://dx.doi.org/10.1016/j.geoderma.2012.04.001>.
- Hewitt, A.E., 1993. Predictive modelling in soil survey. *Soils Fertil.* 56 (3), 305–314.
- Hewitt, A.E., McKenzie, N.J., Grundy, M.J., Slater, B.K., 2008. Qualitative survey. In: McKenzie, N.J., Grundy, M.J., Webster, R., Ringrose-Voase, A.J. (Eds.), *Guidelines for Surveying Soil and Land Resources*. CSIRO Publishing, Melbourne, pp. 285–306.
- Hird, C., 1991. Soil Landscapes of the Goulburn 1:250,000 Sheet. Soil Conservation Service of NSW, Sydney, Australia.
- Isbell, R.F., 1996. *The Australian Soil Classification*. CSIRO Publishing.
- Kellogg, C.E., 1938. *Soil Survey Manual*. United States Department of Agriculture Miscellaneous Publication 274. United States Department of Agriculture, Washington, D. C.
- Kellogg, C.E., 1950. The future of the soil survey. *Soil Sci. Soc. Am. Proc.* 14, 8–13.
- Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. *Geoderma* 151 (3–4), 311–326. <http://dx.doi.org/10.1016/j.geoderma.2009.04.023>.

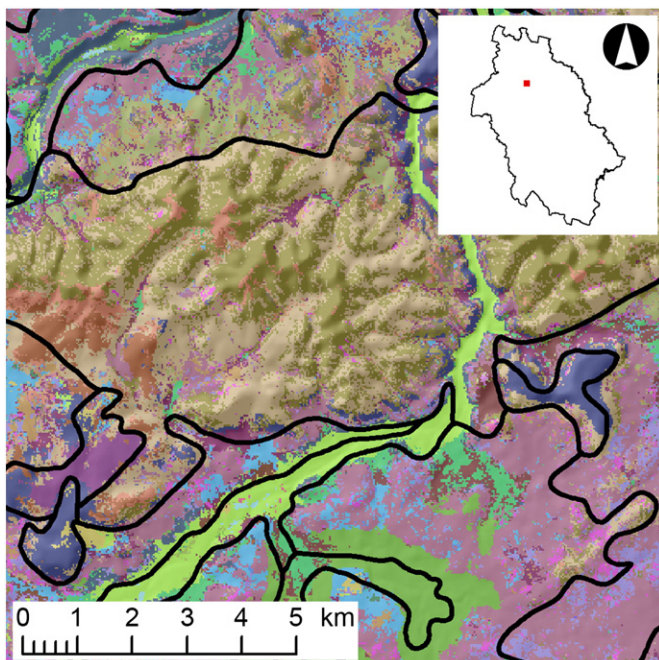


Fig. 9. Most probable soil class with the original soil polygons overlaid.

- Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* 170, 347–358. <http://dx.doi.org/10.1016/j.geoderma.2011.10.007>.
- Lagacherie, P., Legros, J.P., Burrough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. *Geoderma* 65, 283–301. [http://dx.doi.org/10.1016/0016-7061\(94\)00040-H](http://dx.doi.org/10.1016/0016-7061(94)00040-H).
- MacMillan, R.A., Moon, D.E., Coupé, R.A., 2007. Automated predictive ecological mapping in a forest region of B.C., Canada, 2001–2005. *Geoderma* 140, 353–373. <http://dx.doi.org/10.1016/j.geoderma.2007.04.027>.
- McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. *Nutr. Cycl. Agroecosyst.* 50, 51–62.
- McBratney, A.B., Mendonça Santos, M.d.L., Minasny, B., 2003. On digital soil mapping. *Geoderma* 117, 3–52. [http://dx.doi.org/10.1016/S0016-7061\(03\)00223-4](http://dx.doi.org/10.1016/S0016-7061(03)00223-4).
- Minasny, B., McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma* 142, 285–293. <http://dx.doi.org/10.1016/j.geoderma.2007.08.022>.
- Moran, C.J., Bui, E.N., 2002. Spatial data mining for enhanced soil map modelling. *Int. J. Geogr. Inf. Sci.* 16 (6), 533–549. <http://dx.doi.org/10.1080/13658810210138715>.
- Nauman, T.W., Thompson, J.A., Odgers, N.P., Libohova, Z., 2012. Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), *Digital Soil Assessments and Beyond: Proceedings of the Fifth Global Workshop on Digital Soil Mapping 2012*, Sydney, Australia. CRC Press, London, United Kingdom, pp. 203–207.
- Powell, B., 2008. Classifying soil and land. In: McKenzie, N.J., Grundy, M.J., Webster, R., Ringrose-Voase, A.J. (Eds.), *Guidelines for Surveying Soil and Land Resources*. Australian Soil and Land Survey Handbook Series. CSIRO Publishing, Melbourne, pp. 307–315.
- Qi, F., Zhu, A.X., 2003. Knowledge discovery from soil maps using inductive learning. *Int. J. Geogr. Inf. Sci.* 17, 771–795. <http://dx.doi.org/10.1080/13658810310001596049>.
- Rogers, L.G., Cannon, M.G., Barry, E.V., 1999. *Land Resources of the Dalrymple Shire, 1*. Land Resources Bulletin DNRQ980090. Queensland Department of Natural Resources, Brisbane, Queensland.
- Rogowski, A.S., Wolf, J.K., 1994. Incorporating variability into soil map unit delineations. *Soil Sci. Soc. Am. J.* 58 (1), 163–174.
- Rulequest Research, 2008. *See5 Version 2.04*. Rulequest Research Pty. Ltd., Sydney.
- Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. *Prog. Phys. Geogr.* 27 (2), 171–197. <http://dx.doi.org/10.1191/0309133303pp366ra>.
- Scull, P., Franklin, J., Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. *Ecol. Model.* 181, 1–15. <http://dx.doi.org/10.1016/j.ecolmodel.2004.06.036>.
- Shi, X., Long, R., Dekett, R., Philippe, J., 2009. Integrating different types of knowledge for digital soil mapping. *Soil Sci. Soc. Am. J.* 73, 1682–1692. <http://dx.doi.org/10.2136/sssaj2007.0158>.
- Simonson, R.W., 1968. Concept of soil. *Adv. Agron.* 20, 1–47.
- Smith, C.A.S., Daneshfar, B., Frank, G., Flager, E., Bulmer, C., 2012. Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), *Digital Soil Assessments and Beyond: Proceedings of the Fifth Global Workshop on Digital Soil Mapping*, Sydney, Australia. CRC Press/Balkema, Leiden, pp. 215–220 (10–13 April).
- Soil Survey Staff, 1993. *Soil survey manual*. U. S. Department of Agriculture Handbook 18. United States Department of Agriculture Soil Conservation Service.
- Steuer, G.G.L., 1961. *Methods of soil surveying in use at the Netherlands Soil Survey Institute*. Boor en Spade 11, 59–77.
- Subburayalu, S., Jenhani, I., Slater, B.K., 2014. Disaggregation of component soil series using possibilistic decision trees from an Ohio County soil survey map. *Geoderma* 213, 334–345. <http://dx.doi.org/10.1016/j.geoderma.2013.08.018>.
- Sun, W., McBratney, A.B., Hempel, J.W., Minasny, B., Malone, B.P., D'Avella, T., Burras, L., Thompson, J.A., 2010. Digital harmonisation of adjacent analogue soil survey areas—4 Iowa counties. *Proceedings of the 4th Global Workshop on Digital Soil Mapping*, Rome.
- Thompson, J.A., Prescott, T., Moore, A.C., Bell, J., Kautz, D.R., Hempel, J.W., Waltman, S.W., Perry, C.H., 2010. Regional approach to soil property mapping using legacy data and spatial disaggregation techniques. 19th World Congress of Soil Science, Brisbane, Queensland.
- Tucker, C.J., Holben, B.N., Elgin Jr., J.H., McMurtrey III, J.E., 1981. Remote sensing of soil dry-matter accumulation in winter wheat. *Remote Sens. Environ.* 11, 171–189.
- Valentine, K.W.G., 1981. How soil map units and delineations change with survey intensity and map scale. *Can. J. Soil Sci.* 61 (4), 535–551.
- Webster, R., 1973. Automatic soil-boundary location from transect data. *Math. Geol.* 5, 27–37.
- Wilford, J.R., Minty, B.R.S., 2007. The use of airborne gamma-ray imagery for mapping soils and understanding landscape processes. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), *Digital Soil Mapping: An Introductory Perspective*. Developments in Soil Science 31. Elsevier, Amsterdam, pp. 207–218.
- Young, F.J., Hammer, R.D., Williams, F., 1997. Estimation of map unit composition from transect data. *Soil Sci. Soc. Am. J.* 61, 854–861.