Geoderma 237–238 (2015) 190–198

Contents lists available at ScienceDirect

Geoderma

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

Digital soil property mapping and uncertainty estimation using soil class probability rasters

Nathan P. Odgers *, Alex B. McBratney, Budiman Minasny

Department of Environmental Sciences, Faculty of Agriculture and Environment, C81 Biomedical Building, The University of Sydney, New South Wales 2006, Australia

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form 3 September 2014
Accepted 6 September 2014
Available online 16 September 2014

Keywords:
Digital soil property mapping
Weighted means
Prediction interval
Uncertainty estimation
Probability rasters

ABSTRACT

The objective of the work presented in this paper was to investigate how rasters of the probabilities of occurrence of soil classes may be used to create digital soil property maps and maps of their associated uncertainties. The approach we present is formalised in an algorithm we developed called “Digital Soil Property Mapping Using Soil Class Probability Rasters” (PROPR).

The soil class probability rasters were derived previously from a spatial disaggregation of the 1:250,000-scale Dalrymple Shire legacy soil polygon map from central Queensland, Australia.

We created digital soil property maps of soil pH 1:5 H_2O and their uncertainties (as indicated by estimates of the limits of the 90% prediction interval) at six depth increments down the soil profile (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, 100–200 cm). The calculation of the weighted mean soil pH value for each depth increment at each grid cell was based on reference pH values for each soil class and used the probabilities of occurrence at each grid cell as weights.

The calculation of the prediction interval limits for each depth increment involved sampling from the triangular distribution of the soil pH of each soil class using the soil class probabilities at each grid cell as weights in order to identify the number of samples to draw from each distribution. The 90% prediction interval limits were then estimated as the 5th and 95th percentiles of the distribution of samples drawn from the soil classes' triangular distributions.

The maps of soil pH displayed strong spatial patterns. Soil pH generally increased with depth. Uncertainty generally increased with depth. Validation on 300 randomly-selected soil profiles returned a Lin's concordance correlation coefficient of 0.193 at the surface increasing to 0.266 at depth. RMSE increased with depth from about 0.75 pH units at the surface to 1.15 at depth.

Soil class probability rasters are useful for generating digital soil property maps and maps of the associated uncertainties. Validation left room for improvement but the quality of the results is probably strongly affected by the quality of the spatial disaggregation that produced the soil class probability rasters. The PROPR approach may be useful in situations where profile observations are limiting but where legacy soil maps are available.

Generation of the soil class probability rasters to use in PROPR is a predictive exercise in itself and so is also subject to uncertainty. The probability rasters likely can be derived by several methods including logistic regression and data mining; the probability rasters we used were derived via spatial disaggregation of a legacy soil polygon map.

PROPR may be useful in situations where profile observations are limiting but where legacy soil maps are available. The soil class probability rasters need to be produced separately. Information on the within-soil-class variability of the target soil property at each depth increment must be known in order to establish the triangular distributions for the uncertainty estimation.

PROPR may reduce reliance on having sufficient soil profile observations in areas where such data is limiting. We used available profile observations to estimate the within-soil-class variability in order to establish triangular distributions for the soil classes in our study area, but the information required to do so could also be derived from legacy reports or expert knowledge.

1. Introduction

Digital soil property mapping has come a long way in the past few decades. There are several core methods in the literature, and their use has been reviewed comprehensively by others (McBratney et al.,

* Corresponding author.
E-mail addresses: nathan.odgers@sydney.edu.au (N.P. Odgers), alex.mcbratney@sydney.edu.au (A.B. McBratney), budiman.minasny@sydney.edu.au (B. Minasny).

http://dx.doi.org/10.1016/j.geoderma.2014.09.009
0016-7061/

2003; Scull et al., 2003). They commonly rely on having sufficient point data. We will take a brief look at some of the more common techniques. Of the geostatistical techniques, there are now several kinds of kriging, including ordinary kriging (Burgess and Webster, 1980a), block kriging (Burgess and Webster, 1980b), co-kriging (McBratney and Webster, 1983) and others. Statistical techniques usually involve linear regression (for example Moore et al., 1993) to establish relationships between soil classes or soil properties and environmental covariates. Hybrid techniques such as regression-kriging have been available since the mid-1990s (Odeh et al., 1994, 1995). In this case the target soil property is predicted using linear regression and summed with the kriged regression residuals. This is akin to the formal statistical approach of linear mixed model. Data mining techniques such as regression trees have now been widely used to map soil properties (for example Henderson et al., 2005; McKenzie and Ryan, 1999).

As we pointed out, these approaches all rely on having sufficient point data. In the case of kriging, Webster and Oliver (1992) concluded that 150 to 200 point observations are required to estimate the variogram reliably in areas where the spatial variation is isotropic, and potentially 300 to 400 observations where the spatial variation is anisotropic (Voltz and Webster, 1990). Data mining or regression techniques also rely on enough soil observations that cover the whole variation of the covariates so that the spatial prediction function can be applied to the area of interest.

For mapping large regions, to assemble a dataset of this size via sampling anew is often impractical or even prohibitive for reasons of cost or time. At the same time it may not be feasible to rely on legacy observations for reasons of data availability (do we have enough data?); data accessibility (can we get our hands on it?); or because of questions over data provenance, which covers a host of issues including the quality and completeness of the legacy data, positional accuracy, representivity, sampling scheme and so-on (Lagacherie, 2008).

Some digital soil class mapping techniques are able to produce spatially exhaustive probability distributions of the soil classes in an area of interest. One example is our DSMART algorithm (Odgers et al., 2014), which spatially disaggregates legacy soil polygon maps to produce raster maps of the estimated probabilities of occurrence of the soil classes defined in the legacy map. Other disaggregation techniques produce similar output (for example Kempen et al., 2009) or could potentially be adapted to do so (for example Bui and Moran, 2001; Häring et al., 2012; Subburayalu et al., 2014).

Others have demonstrated how the probability rasters can be used in a weighted-mean setting to generate maps of soil properties (Kempen et al., 2010, 2011). Still others have demonstrated similar approaches using soil class similarity rasters (Zhu and Band, 1994; Zhu et al., 1997). The theme of this approach is not new and has commonly been applied at the map unit level using areal proportions of the soil classes as weights (for example Davidson and Lefebvre, 1993; Galbraith et al., 2003; Homann et al., 1998; Odgers et al., 2012); the disadvantage in doing so at the map unit level is that the weighted mean property value is invariant within the respective map unit polygon. On the other hand, using a raster data model where the soil class probabilities vary from grid cell to grid cell allows the predicted soil property value to vary cellwise also, which more closely approximates the reality of soil variation across the land surface.

It is also desirable to have an estimate of the uncertainty associated with the weighted mean prediction at each grid cell. Zhu et al. (1997) did not do so, and the approach of Kempen et al. (2010) heavily relied on having sufficient point observations. Therefore given a set of soil class probability rasters and some reference soil property data, we can easily make predictions of the target soil property that vary from grid cell to grid cell, and in this paper we suggest a method of estimating the uncertainty associated with these predictions that does not rely on having abundant point observations.

2. Aim

In this paper we introduce an algorithm we developed called PROPR (for "digital soil property mapping using soil class probability rasters"). The aim of this paper is to demonstrate how soil class probability rasters can be used to generate predictions of soil properties and their associated uncertainties.

3. Methods

3.1. 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).

3.2. Disaggregated soil map

In an earlier work (Odgers et al., 2014) we spatially disaggregated a soil polygon map from the Dalrymple Shire which yielded maps of the estimates of probability of occurrence of the 71 soil classes in the original soil polygon map. The probabilities of occurrence were mapped onto a 30-m resolution raster grid covering the study area (some examples

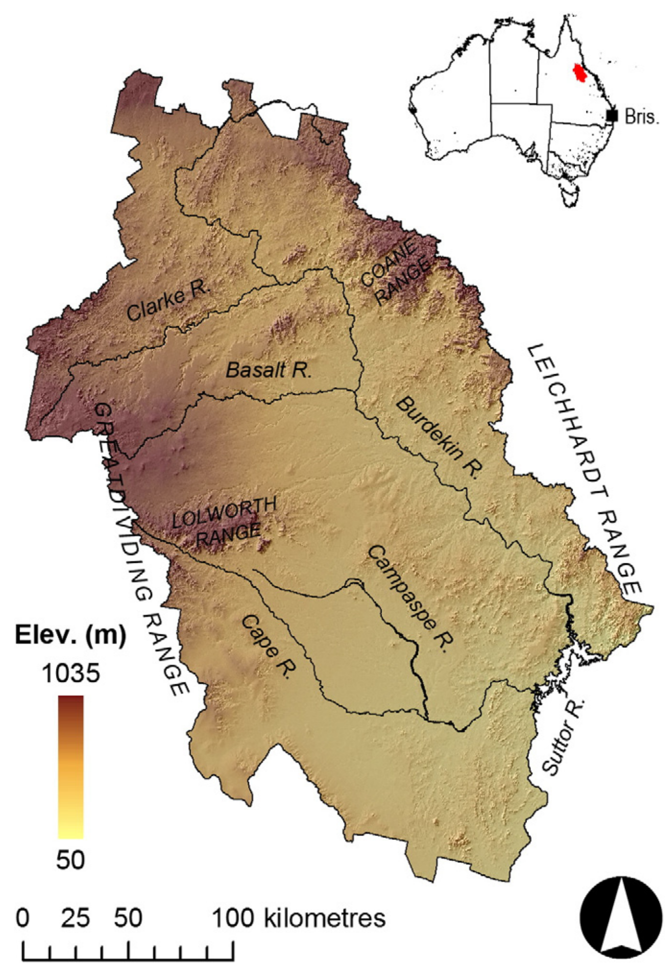

Fig. 1. Overview of the study area in central Queensland, Australia.