Catena 164 (2018) 125–134

Contents lists available at ScienceDirect

Catena

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

Updating a national soil classification with spectroscopic predictions and digital soil mapping

Hongfen Teng[a,b], Raphael A. Viscarra Rossel[a,*], Zhou Shi[b], Thorsten Behrens[c]

[a] Bruce E. Butler Laboratory, CSIRO Land & Water, PO Box 1700, Canberra, ACT 2601, Australia
[b] Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, 310058 Hangzhou, China
[c] Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Rümelinstraße 19-23, 72074 Tübingen, Germany

ARTICLE INFO

Keywords:
Soil mapping
Digital soil mapping
Soil classification
Random forests
Visible–near infrared spectroscopy

ABSTRACT

Traditional soil maps have helped us to better understand soil, to form our concepts and to teach and transfer our ideas about it, and so they have been used for many purposes. Although, soil maps are available in many countries, there is a need for them to be updated because they are often deficient in that their spatial delineations and their descriptions are subjective and lack assessments of uncertainty. Updating them is a priority for federal soil surveys worldwide as well as for research, teaching and communication. New data from sensors and quantitative 'digital' methods provide us with the tools to do so. Here, we present an approach to update large scale, national soil maps with data derived from a combination of traditional soil profile classifications, classifications made with visible–near infrared (vis–NIR) spectroscopy, and digital soil class mapping (DSM). Our results present an update of the Australian Soil Classification (ASC) orders map. The overall error rate of the DSM model, tested on an independent validation set, was 55.6%, and a few of the orders were poorly classified. We discuss the possible reasons for these errors, but argue that compared to the previous ASC maps, our classification was derived objectively, using currently best available data sets and methods, the classification model was interpretable in terms of the factors of soil formation, the modelling produced a 1 × 1 km resolution soil map with estimates of spatial uncertainty for each soil order and our map has no artefacts at state and territory borders.

1. Introduction

Traditional soil maps are the outcome of many years' experience and investigation by pedologists, who have made meticulous descriptions of easily-observable morphological characteristics of soil profiles combined with often sparsely gathered laboratory measurements. They aggregated and simplified the information by dividing the variation into more manageable soil classes, which are assumed to exist in fixed proportions and that can be interpreted for different uses. These soil classes are expected to be similar in terms of their intrinsic chemical and physical characteristics, and representative of other soil types in other similar bio-climatic regions and landscapes.

There are two international systems for soil classification that have produced soil maps for the world: the Soil Taxonomy (Soil Survey Staff, 2014) and the World Reference Base for Soil Resources (WRB) (FAO, 2014). But there are also many national systems and maps that are used more locally for *e.g.* land use planning, evaluations of hydrology and agricultural land. Examples include the soil classification systems of France (Baize and Girard, 1995), Russia (Lebedeva and Gerasimova, 2012), Germany (Ad-hoc-AG Boden, 2005), Brazil (EMBRAPA, 2006), and China (Gerasimova, 2010; Shi et al., 2006). All reflect local pedological descriptions. The most widely used soil map in Australia was derived from a general purpose hierarchical classification system that consists of five levels: order, suborder, great group, sub-group and family (Isbell, 2002). At the top level there are fourteen soil orders that reflect the arid, strongly weathered nature of Australia.

Soil maps derived from such classifications have helped us to better understand soil, to form our concepts and to teach and transfer our ideas about it. They are valuable because of the expertise that has been used to create them and, which is inherently contained in the maps. Because soil is directly related to climate, vegetation, parent material and relief, soil maps have been useful in many soil and environmental applications, such as land management, ecosystem assessments and modelling (Yang et al., 2011). However, traditional soil maps are limited in terms of both their spatial delineations and their representations of the soil attributes within the classes (Bui and Moran, 2001; Wilson, 2005). There is a need to update traditional soil maps with modern methods and technologies to provide more objective and accurate

* Corresponding author.
E-mail address: raphael.viscarra-rossel@csiro.au (R.A. Viscarra Rossel).

https://doi.org/10.1016/j.catena.2018.01.015
Received 20 October 2017; Received in revised form 9 January 2018; Accepted 15 January 2018
Available online 20 February 2018
0341-8162/

classifications of the spatial distribution of soil types.

Digital soil mapping (DSM) (McBratney et al., 2003) can help to derive improved versions of soil maps by combining new spatially explicit data collected with new technologies and the traditional soil maps. Several authors have used DSM to update existing maps, either the descriptions of the soil (Nelson and Odeh, 2009; Vaysse and Lagacherie, 2015; Werban et al., 2013; Yang et al., 2011), the delineations (Behrens et al., 2008), or both (Behrens and Scholten, 2006), but regardless of how good the updating models might be, success largely depends on having accurate and sufficient new data.

Those data can come primarily from soil survey and measurement, which is often costly and time-consuming. One technology, which can help to improve the efficiency of soil survey is diffuse reflectance spectroscopy in the visible and near infrared range (vis–NIR, 400–2500 nm) (Viscarra Rossel et al., 2016). It enables us to extract soil information on colour, iron oxide, clay and carbonate mineralogy, organic matter content and composition, the amount of water present and its particle-size distribution, quickly and cheaply. The integration of vis–NIR spectroscopy, remote sensing and DSM is enabling soil mapping over large and sparsely sampled regions of the world. In Australia, this approach has been used to map soil properties such as clay and iron mineralogy (Viscarra Rossel, 2011; Viscarra Rossel et al., 2010), carbon stocks (Viscarra Rossel et al., 2014), phosphorus stocks (Viscarra Rossel and Bui, 2016), clay, sand, silt contents, pH, cation exchange capacity, bulk density, organic carbon content, total nitrogen (Viscarra Rossel et al., 2015) and soil erosion (Teng et al., 2016).

Although digital mapping has been used for updating soil maps in small sample regions (Cambule et al., 2013; Grimm et al., 2008; Guo et al., 2013; Kempen et al., 2012), few pedologists have investigated digital soil class mapping over large scales, and none in combination with vis–NIR spectroscopy. Thus, we have produced an updated quantitative version of the ASC orders map using DSM with random forests and derived estimates of spatial uncertainty for each class. Here we describe our procedure and the results.

2. Material and methods

2.1. The data set

We used data from two sources originating from 38 756 unique sites across Australia (Fig. 1). The first set was from a historical archive of data contained in the Commonwealth Scientific and Industrial Research Organization (CSIRO) National soil database (NATSOIL) and in databases of soil survey organizations in each of the Australian States and Territories, which were collated during a national soil site data collation (NSSDC) as part of the 'Soil and Landscape Grid of Australia' project (Grundy et al., 2015). It consists of 33 784 sites that were classified by soil surveyors, during numerous projects, according to the Australian Soil Classification (ASC) (Isbell, 2002).

The second set of data was also from the NATSOIL database, but in this case profiles from 3847 sites had no ASC classification assigned to them. Nevertheless in the development of the Australian soil spectroscopic database (Viscarra Rossel and Webster, 2012), we had recorded the visible–near infrared (vis–NIR) spectra of these soil samples and so, we could use discriminant models developed by Viscarra Rossel and Webster (2011) to assign ASC orders to them. The models developed by Viscarra Rossel and Webster (2011) used data from the CSIRO's NATSOIL database, which represented all of the ASC orders. The authors showed that vis–NIR spectra could be used to fairly accurately discriminate among horizons and the orders of the ASC. They describe in detail the spectroscopic measurements and the modelling they performed. Readers are directed to that publication for details.

The third set of data, which provides an even cover of points in central and western Australia are from 1125 sites held in the National Geochemical Survey of Australia (de Caritat et al., 2008). Again, none of the sites had an ASC order assigned, but the soil from all sites had visible–near infrared (vis–NIR) spectra so that the ASC orders could be assigned to them with the models developed by Viscarra Rossel and Webster (2011).

The combined data made up of the NSSDC, NATSOIL, and the vis–NIR estimates represents all of the Australian states and territories and all orders of the ASC (Table 1).

2.2. Digital soil class mapping

We used the Jenny-like DSM framework (McBratney et al., 2003; Jenny, 1941) to model the ASC orders, o, as a function of various environmental predictors:

$$o(\mathbf{u}) = f(s[\mathbf{u}], c[\mathbf{u}], v[\mathbf{u}], r[\mathbf{u}], p[\mathbf{u}]) \quad (1)$$

where the soil–environmental factors represented by the environmental predictors across space ($\mathbf{u} \equiv \{x,y\}$) are soil ($s$), climate ($c$), vegetation ($v$), terrain ($r$), and parent material ($p$). The function used to relate these factors to o was random forest, which we describe below. The proxies for these soil–environmental factors that we used in the modelling are given in Table 2. They were chosen to represent factors that affect the formation and distribution of soil in Australia, and were from several sources, including remote sensing, other soil maps, maps of climatic variables and terrain attributes derived from a digital elevation model (DEM). We used bilinear resampling to harmonise the different resolutions of these data (Table 2) to a common grid with cell size of 1×1 km.

To assess the effect of using the original ASC orders map (Isbell, 2002) in the modelling, we derived two random forest models, one that used all of the predictors shown in Table 2, including the original ASC map, which we refer to as the ASC_{Isbell} map, and the other with all predictors except the ASC_{Isbell} map.

2.3. Data mining

We separated the dataset, $\mathbb{D}$, containing 38 756 observations and their covariates, into a training and a validation set by random sampling. Two-thirds were assigned to the training set, $\mathbb{T} = 25\,837$, and the remaining to the validation set, $\mathbb{V} = 12\,919$.

We used random forest (Breiman, 2001) to classify and map the Australian soil classification orders, o. Random forest is an ensemble of B trees $\{\mathbf{t}_1(\mathbf{p}),\ldots,\mathbf{t}_B(\mathbf{p})\}$, where $\mathbf{p} = \mathbf{p}_1,\ldots,\mathbf{p}_p$ is a p-dimensional vector of covariates (or predictors) that represent the soil–environmental factors (Table 2). The ensemble produces B outputs $\{\hat{o}_1 = \mathbf{t}_1(p),\ldots,\hat{o}_B = \mathbf{t}_B(\mathbf{p})\}$, where $\hat{o}_B$, $b = 1,\ldots,B$, is the classification of the ASC class by the bth tree. Outputs of all trees are aggregated to produce, by majority vote from all trees, the final classification, $\hat{o}$.

Given a set of training data, $\mathbb{T} = \{(\mathbf{p}_1, o_1),\ldots,(\mathbf{p}_n, o_n)\}$, where $\mathbf{p}_i$, $i = 1,\ldots,n$, is a vector of predictors and o_i is the corresponding soil order, training of the random forest proceeds as follows:

1. From $\mathbb{T}$, draw B bootstrap samples. Each bootstrap is the basis for one of the 'trees in the forest'.
2. Then, grow a classification tree for each bootstrap sample with no pruning, to derive the ASC classification, $\hat{o}$.
3. At each node, rather than choosing the best split among all predictors, randomly sample m predictors and choose the best split from among them. The value m is held constant while the forest is grown.
4. Repeat the above steps until B trees are grown.
5. For each tree, predict the data not in the bootstrap sample (*i.e.* the out-of-bag data, which on average, for each data would be approximately 36% of the time) using the tree grown with the bootstrap sample (*i.e.* data that is in-the-bag).
6. Aggregate the out-of-bag predictions and compare the predicted $\hat{o}$ values, with the observed values, o, of each unit in the out-of-bag (oob) sample and calculate the classification error-rate (ER):