Geoderma 232–234 (2014) 10–18

Contents lists available at ScienceDirect

Geoderma

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

Non-parametric imputation of properties for soil profiles with sparse observations

David Clifford [a,*], Melissa J. Dobbie [a], Ross Searle [b]

[a] CSIRO Computational Informatics, P.O. Box 2583, Brisbane 4001, Queensland, Australia
[b] CSIRO Land and Water, P.O. Box 2583, Brisbane 4001, Queensland, Australia

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 18 April 2014
Accepted 23 April 2014
Available online 14 May 2014

Keywords:
Global Soil Map (GSM) project
Missing data
Soil attribute
Non-parametric imputation

ABSTRACT

Soil profile data are a collection of soil property values associated with a series of non-overlapping depth intervals. The CSIRO National Soil (NatSoil) Archive database contains full soil profile data recorded for over 56,000 such depth intervals at approximately 9500 sites around Australia. Another database developed by CSIRO uses spectroscopic estimates of soil attributes for soil sampled as part of Geoscience Australia's Geochemical Survey of Australia. That survey is made up of soils collected at two depth intervals (0–0.1 m and 0.6–0.8 m) from 2244 different sites in Australia. The key question of interest is how complete soil profile data can be used to impute or "fill in" missing soil depth interval data to increase the utility of the two-depth spectroscopic data for a range of environmental analyses. We demonstrate our approach through using the complete NatSoil database to impute missing two-depth interval data to create complete profiles of total phosphorus, total nitrogen and total potassium. A parametric modeling approach to imputation was initially considered but having data at just two depth intervals led us to non-parametric approaches. We simulated from a large quantity of complete profile data with similar first and second order properties as the original data and drew random samples from the simulated data to predict (impute) the two-depth data with quantified certainty for each intervening Global Soil Map depth interval and profile. The complete imputed profiles can be used in future modeling and mapping. We believe our imputation procedure can be extended to other scenarios in soil science where joint imputation of multiple soil properties can be used to fill the gaps arising from incomplete soil profiles.

1. Introduction

A ubiquitous problem with environmental datasets is missing observations (Hopke et al., 2013). The questions then arise as to why data are missing and how an analysis should accommodate missing data. Given missing data occur for a number of reasons, there is no single approach to best accommodate missing values. Understanding how the missing data arise will help with selecting an appropriate approach.

The least attractive option is case deletion, where missing (dependent variable) observations and associated property (independent) data are omitted from an analysis. Not only does this induce data wastage but also because systematic differences (if they exist) between complete and incomplete observations are ignored, this option produces unbiased estimates only if missing data are missing completely at random. Furthermore, standard errors are typically higher, given less information is available to an analysis based on a reduced set of observations. For further details about undertaking statistical analyses in the face of missing data, see Little and Rubin (2002).

Filling-in missing values, known as imputation, is appealing because standard methods and existing software for analyzing complete data can then be used. This reason alone greatly reduces the burden of developing tailored methods and code for analyzing incomplete data. Generally, methods for obtaining complete data can be classified as single or multiple imputation. With single imputation, one value is imputed for each missing value whereas with multiple imputation, missing values are replaced with two or more acceptable values. The latter has the advantage of being able to handle complicated data structures, sophisticated missing-data mechanisms and can more effectively represent imputation uncertainty. On the downside, multiple imputation leads to more computationally intensive data analysis requirements associated with the analysis of a larger dataset.

Soil profile data are a collection of soil property values associated with a series of non-overlapping depth intervals. The collection and measurement of these soil properties is costly and time-consuming (Viscarra Rossel and McBratney, 1998). In the case of soil profile measurements, where there could be one or many missing values for soil depth intervals in a particular profile, imputation is cheaper and more efficient than directly obtaining additional profile measurements. There are several approaches available for imputing values for particular depth intervals including parametric modeling of the attribute as a

[*] Corresponding author. Tel.: +61 7 3833 5532; fax: +61 7 3833 5505.
E-mail addresses: David.Clifford@csiro.au (D. Clifford), Melissa.Dobbie@csiro.au (M.J. Dobbie), Ross.Searle@csiro.au (R. Searle).

http://dx.doi.org/10.1016/j.geoderma.2014.04.026
0016-7061/Crown Copyright © 2014 Published by Elsevier B.V. All rights reserved.

function of depth (Minasny et al., 2006), as well as the use of equal-area quadratic splines (Bishop et al., 1999). Where the profile data is largely missing, both of these approaches are limited and provide little or no information about the certainty associated with imputed values. As such, there is a clear need to explore this issue in more detail.

Our interest in obtaining complete soil depth profiles is motivated by Australia's contribution to the Global Soil Map (GSM) (Minasny and McBratney, 2010). The GSM specification requires estimates of 13 soil properties at 6 depth intervals (in meters); 0–0.05, 0.05–0.15, 0.15–0.3, 0.3–0.6, 0.6–1, and 1–2. The Australian approach for estimating these soil properties is based on predictive soil mapping techniques using observed soil profile information. This digital soil mapping approach (McBratney et al., 2003) relies on relationships between observed soil properties and exhaustively sampled, easily obtainable, digital auxiliary predictor variables or covariates, (e.g. remote sensing data, a digital elevation model and terrain derivatives, geology, and land use), and inference is based on a statistical model that produces quantitative estimates of soil properties and their associated error (Viscarra Rossel, 2011; Viscarra Rossel and Chen, 2011).

We seek to demonstrate how a set of complete soil depth profile data for three soil properties can be used to impute missing values with quantified certainty in another set of largely incomplete soil depth profiles for the same three attributes. We describe a non-parametric simulation approach for undertaking the imputation and illustrate its application to an Australian dataset, but the generality of the approach is emphasized.

2. Methods

2.1. Data description

The CSIRO National Soil Archive (NatSoil) database (Karssies, 2011) is a collation of soil profile data for a range of soil physical and chemical measurements recorded at a range of depth intervals at over 9500 sites around Australia. However, the depth of measurements and the properties recorded at each site are not necessarily consistent or complete.

CSIRO's national visible and near infrared spectroscopic database (Viscarra Rossel and Webster, 2012) comprises many inferred soil attributes based on predictive models. These models have been applied to soil collected at two depth intervals (0–0.1 m and 0.6–0.8 m) from 2244 sites in Australia by Geoscience Australia (GA) as part of their Geochemical Survey of Australia (De Caritat et al., 2007). After checking and cleaning, there were 1116 two-depth spectroscopic profiles of soil property values derived for both the 0–0.1 m and 0.6–0.8 m depth intervals. The challenge is to fill in the gaps for each soil property over the range 0.1–0.6 m. For each database, the soil attributes of interest are total phosphorus (Total P, %), total potassium (Total K, %) and total nitrogen (Total N, %), all recorded as percentage mass.

We demonstrate our technique using these two datasets. However, it is important to highlight one key difference between the datasets is that the original GA soil samples were collected mostly from overbank sediments near the outlets of large drainage basins, and as a result, are fine-grained (De Caritat et al., 2007). On the other hand, the NatSoil soil profiles considered here have no imposed locational spatial structure though they have been collected in regions of Australia that are more intensively sampled, corresponding for the most part, with regions of more intensive agriculture.

While the NatSoil database contains information for a large number of sites around Australia, we reduced the database to those sites that have complete information recorded for soil attributes of interest over the depth range 0 m to 0.80 m (which is the depth range that we need data for in order to form predictions to complete two-depth spectroscopic profiles). These depth intervals were extracted for each property separately, giving 6503, 2631, and 6191 profiles for Total P, Total N and Total K, respectively. Any attempt to impute soil properties at depths deeper than 0.8 m would be based on fewer profiles.

We note that NatSoil profiles with missing values can potentially be included in this kind of analysis, thereby boosting the number of profiles available in each subset. But such an analysis would require additional assumptions about the multivariate distribution of the values for each core. For example, we could assume that observations follow a multivariate Normal distribution (perhaps after log transformation say) and that gaps are a result of values being missing at random, which may or may not be the case. A brief examination of the complete profiles indicates that they do not follow any known multivariate distribution.

Fig. 1 indicates the locations of the NatSoil and two-depth soil profiles that were included in this study.

2.2. Data preparation

The depth interval data in the NatSoil database are harmonized to six depth intervals defined by cutoff values of 0 m, 0.05 m, 0.1 m, 0.15 m, 0.3 m, 0.6 m, and 0.8 m using mass-preserving, equal-area quadratic smoothing splines (Malone et al., 2009). A quadratic spline is a smooth curve through a set of points (e.g. bulk horizon/depth interval data) that is created by fitting a piecewise series of local quadratic polynomials over the depth intervals of the soil profile. These mass-preserving splines are superior to linear and quadratic polynomial regression, as well as exponential decay, in predicting soil depth functions based on bulk horizon data (Bishop et al., 1999) and are now commonly used to obtain a continuously varying depth function (e.g. Berhongaray et al., 2013; Odgers et al., 2012). The six depth intervals defined by our cutoff values are chosen based on the combination of standard depths specified for the GSM (Minasny and McBratney, 2010) and the depths at which the two-depth spectroscopic profiles were recorded. The harmonization was carried out on the natural scale for each variable to preserve the mass of each variable over the observed depth intervals. Fig. 2 illustrates harmonization of a NatSoil profile for Total N.

All analyses and graphing were undertaken using the statistical software package R Version 2.15 (R Development Core Team, 2012).

2.3. Non-parametric data simulation

We consider a non-parametric approach for simulating and imputing soil properties. First we describe how to simulate soil properties at a single depth interval (univariate case). We then describe how to simulate soil properties across multiple depth intervals (multivariate case).

2.3.1. Univariate case

We simulate data by resampling from the observed dataset. One downside to resampling is that only those values that are observed will ever be resampled, a problem we do not face here given the sample sizes of available data from the NatSoil database in this case. A slightly more sophisticated approach that gets around this issue is to simulate uniform random values over the interval [0,1], the range of an empirical cumulative distribution function (CDF) and map them through the inverse empirical CDF of the observed data to simulate data (e.g., see Fig. 3, where the uniform value u drawn from interval [0,1] is mapped to a simulated value of 0.813.). This non-parametric approach produces a rich set of simulated data whose distribution matches the empirical distribution of the observed data.

2.3.2. Multivariate case

Simulating multivariate data is more challenging and at the very least the simulated data should possess the correlation structure we observe between the soil observations across depth intervals. Another attribute of the soil data that needs to be accounted for in the simulation is the high skewness of the data. We transform the data to remove the skewness by mapping the quantiles of the data to corresponding quantiles of the standard normal distribution (Glasbey and Allcroft, 2008). Such a map is invertible and ensures marginal normality of the transformed data at each depth. Marginal normality does not imply