Computers & Geosciences 67 (2014) 62–68

Contents lists available at ScienceDirect

Computers & Geosciences

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

Pragmatic soil survey design using flexible Latin hypercube sampling

David Clifford [a,b,*], James E. Payne [c], M.J. Pringle [c], Ross Searle [a,d], Nathan Butler [a,b]

[a] Commonwealth Scientific and Industrial Research Organisation (CSIRO) Sustainable Agriculture Flagship, GPO Box 2583, Brisbane, QLD 4001, Australia
[b] CSIRO Computational Informatics, GPO Box 2583, Brisbane, QLD 4001, Australia
[c] Department of Science, Information Technology, Innovation and the Arts, GPO Box 2454, Brisbane, QLD 4001, Australia
[d] CSIRO Land & Water, GPO Box 2583, Brisbane, QLD 4001, Australia

ARTICLE INFO

Article history:
Received 16 November 2013
Received in revised form
17 February 2014
Accepted 6 March 2014
Available online 18 March 2014

Keywords:
Soil erosion
Soil survey
Sampling
Digital soil mapping

ABSTRACT

We review and give a practical example of Latin hypercube sampling in soil science using an approach we call flexible Latin hypercube sampling. Recent studies of soil properties in large and remote regions have highlighted problems with the conventional Latin hypercube sampling approach. It is often impractical to travel far from tracks and roads to collect samples, and survey planning should recognise this fact. Another problem is how to handle target sites that, for whatever reason, are impractical to sample – should one just move on to the next target or choose something in the locality that is accessible? Working within a Latin hypercube that spans the covariate space, selecting an alternative site is hard to do optimally. We propose flexible Latin hypercube sampling as a means of avoiding these problems. Flexible Latin hypercube sampling involves simulated annealing for optimally selecting accessible sites from a region. The sampling protocol also produces an ordered list of alternative sites close to the primary target site, should the primary target site prove inaccessible. We highlight the use of this design through a broad-scale sampling exercise in the Burdekin catchment of north Queensland, Australia. We highlight the robustness of our design through a simulation study where up to 50% of target sites may be inaccessible.

1. Introduction

Historically, soil maps have been compiled by qualitative delineation of soil boundaries based on conceptual understanding of soil-formation factors (Jenny, 1941). This implicit understanding is usually developed through sampling of the soil landscape in question. The collection of soil samples is typically resource-intensive and expensive; thus it is important for the sampling programme to be conducted as efficiently as possible, to gain the most information for the least cost. Traditional soil surveys use purposive sampling where data are collected at locations considered to be typical of the soil- or map-unit being quantified (Hewitt et al., 2008). The method is commonly employed in medium- to small-scale surveys (e.g., 1:25,000–1:250,000 scale) and relies heavily on the personal judgement and experience of the surveyor.

In the current working environment, with increasing demand for soil information often accompanied with reductions in field-sampling budgets, there is a strong interest in applying digital soil mapping (DSM) techniques (McBratney et al., 2003) to enhance the efficacy of the soil-mapping process. DSM techniques typically generate statistical relationships between measured soil-profile data and exhaustively sampled, easily obtainable raster surfaces of covariates (e.g., remote sensing data, a digital elevation model and its terrain derivatives, geology, land use). Inferences about soil properties at new locations are based on the model, which produces quantitative estimates of soil properties and their associated error (Viscarra Rossel and Chen, 2011).

Latin hypercube sampling (LHS) is a sampling technique that marries the purposive sampling of traditional soil survey, and the numerical ideas that underpin DSM. LHS was proposed by McKay et al. (1979) as an efficient way to reproduce an empirical distribution function. Helton and Davis (2003) traced the historical development of LHS. In essence, the idea is to divide the empirical distribution function of a variable, X, into n equiprobable, non-overlapping strata, and then draw one random value from each stratum. For k variables, X_1,X_2,…,X_k, the n random values drawn for variable X_1 are combined randomly with the n random values drawn for variable X_2, and so on until n k-tuples are formed, i.e., the Latin hypercube sample (Iman and Helton, 1988). LHS assumes that the k variables are independent, and so extensions that account for correlation have been proposed (Iman and Conover, 1982; Stein, 1987). LHS was readily adopted by the simulation-modelling community as a computationally feasible way to assess the uncertainty of model output, given the empirical distribution functions used as input; indeed, Iman and Helton (1988) showed

[*] Corresponding author at: Commonwealth Scientific and Industrial Research Organisation (CSIRO) Sustainable Agriculture Flagship, GPO Box 2583, Brisbane, QLD 4001, Australia. Tel.: +61 7 3833 5532.
E-mail address: David.Clifford@csiro.au (D. Clifford).

http://dx.doi.org/10.1016/j.cageo.2014.03.005
0098-3004/Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.

that LHS outperformed alternative approaches to uncertainty analysis. Notable soil-related applications of LHS include: simulation of random fields (Pebesma and Heuvelink, 1999) evaluating the probability that cadmium exceeds its contamination threshold (Van Meirvenne and Goovaerts, 2001; Brus et al., 2002); and quantifying the uncertainty of the predictions of pedotransfer functions (Minasny and McBratney, 2002).

The application of LHS most relevant to this study is the design of a soil-sampling scheme in the presence of ancillary information. Minasny and McBratney (2006) blazed a trail with a method they called conditioned Latin hypercube sampling (cLHS). They reasoned that ancillary information should be used to determine soil-sampling locations, provided that it is cheaply obtained, spatially exhaustive, and plausibly related to soil variability. The aim of cLHS is to geographically locate soil samples such that the empirical distribution functions of the ancillary information associated with the samples are replicated, with a constraint that each k-tuple of ancillary information has to occur in the real world. The constraint necessitates conditioning of the Latin hypercube sample. Conditioning is achieved by drawing an initial Latin hypercube sample from the ancillary information, then using simulated annealing to permute the sample in such a way that an objective function is minimised. The objective function of Minasny and McBratney (2006) comprised three criteria: (i) the match of the sample with the empirical distribution functions of the continuous ancillary variables; (ii) the match of the sample with the empirical distribution functions of the categorical ancillary variables; and, (iii) the match of the sample with the correlation matrix of the continuous ancillary variables. The cLHS algorithm has been widely applied (Lin et al., 2009; Kidd et al., 2012; Worsham et al., 2012; Louis et al., 2014; Taghizadeh-Mehrjardi et al., 2014).

Modifications to cLHS have previously been proposed. Minasny and McBratney (2010) have proposed one modification to better sample the edges of the multivariate distribution of the covariates. Roudier et al. (2012) and Mulder et al. (2013) both demonstrated how the cLHS objective function can be modified so that site accessibility is also considered, although it must be pointed out that these modifications do not guarantee accessibility, only increase its probability.

However, unaddressed impracticalities of the approach remain. Cambule et al. (2013) criticised DSM techniques, including cLHS, as being impractical and prohibitively expensive in large regions with access difficulties due to lack of roads or difficult terrain. They also showed that models built on limited data from within accessible regions can be successfully used to predict soil properties in similar but inaccessible regions. Thomas et al. (2012) warned about the need for sensibly chosen ancillary information when using cLHS. Furthermore, they criticised the inflexibility of cLHS, because it does not provide any alternative when the soil surveyor has taken the trouble of travelling to a site, only to find that the prescribed sampling location is inaccessible.

1.1. *Flexible Latin hypercube sampling*

Our goal is to describe extensions to the cLHS method when parts of the survey area are known to be inaccessible prior to sampling. Furthermore, we wish to choose target sites that are more easily accessible than ones that are not and we wish to take prior information into account when selecting new sites to sample from. Finally, and most importantly, we also aim to make cLHS more flexible by highlighting how to choose an alternative site in an objective manner when a particular primary target site is found to be inaccessible when one attempts to visit it. It is important to consider these issues because we may be sampling in large remote areas where travel is restricted due to time and safety constraints.

The goal of cLHS is to optimally sample the covariate space of the region of interest. Ideally, the histograms of the covariate values for the target sites should look the same as histograms of the covariate values for the entire region. We can choose target sites to achieve this but inaccessibility means that the histograms of the covariate values for the sites actually sampled may be quite different to the histograms of covariate values for the target sites.

To explore this issue a little further it may help the reader to consider six different covariate spaces as follows:

(1) the covariate space associated with the region of interest;
(2) the covariate space associated with the subset of the region that is accessible to sampling;
(3) the covariate space spanned by sites previously sampled;
(4) the covariate space spanned by the target sites;
(5) the covariate space spanned by the collection of target sites and previously sampled sites; and
(6) the covariate space spanned by all sampled sites (new and previously sampled sites).

The covariate spaces at positions 1 and 4 in our list are the ones considered in cLHS. When there are no previously sampled sites and all target sites are successfully visited then the covariate spaces at positions 4 and 6 are identical. Depending on the terrain and remoteness of the landscape in question, much of the target region may be inaccessible in all but the most well-funded soil surveys (Cambule et al., 2013). Outside of single-property (Vašát et al., 2010) or small-area surveys (Lacoste et al., 2014), the target sites chosen for sampling are never perfectly sampled (Kidd et al., submitted for publication). Kidd et al. (submitted for publication) reported failing to reach over 40% of target sites in a large cLHS-based study. As such, the spaces at positions 5 and 6 may be quite different from each other, when they should look like the space at position 1.

The algorithm of Minasny and McBratney (2006) does not distinguish between 1 and 2, makes no provision for the inclusion of 3 (thus ignoring 5) and does not build any robustness into the design to try to ensure that 5 and 6 are as close as possible. Our goal is to select target sites from the subset of the region that is accessible to sampling. We will choose these target sites to match space 5 to space 1 as best we can. We also propose a method to help ensure that we match 6 to 1 by objectively ranking alternative sites close to each target in case some prove to be inaccessible on the day of sampling.

2. Methods

2.1. *Study area*

The method presented in this paper was developed as part of a project studying soil erosion vulnerability in the watersheds flowing to the Great Barrier Reef (GBR). The health of the GBR, off the coast of northern Queensland, Australia, is the subject of immense ecological concern. Sediment-laden run-off from agricultural land is considered to be a key factor that influences the quality of water arriving to the GBR (Wooldridge, 2009; Brodie et al., 2013). Catchment-scale modelling of the lands that drain into the GBR has indicated that the Burdekin catchment (with an area of 12.8 million ha) is the largest source of this sediment, exporting about 4 Tg per year or 29% of the total average annual load (Kroon et al., 2010).

The Burdekin is dominated by cattle-grazing of natural vegetation across the majority of the catchment. Past mapping and sampling programmes (Fig. 1) have provided a rich but patchy legacy dataset of site and polygon mapping information for