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.

Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.

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, $X_1, X_2, \ldots, X_k$, 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

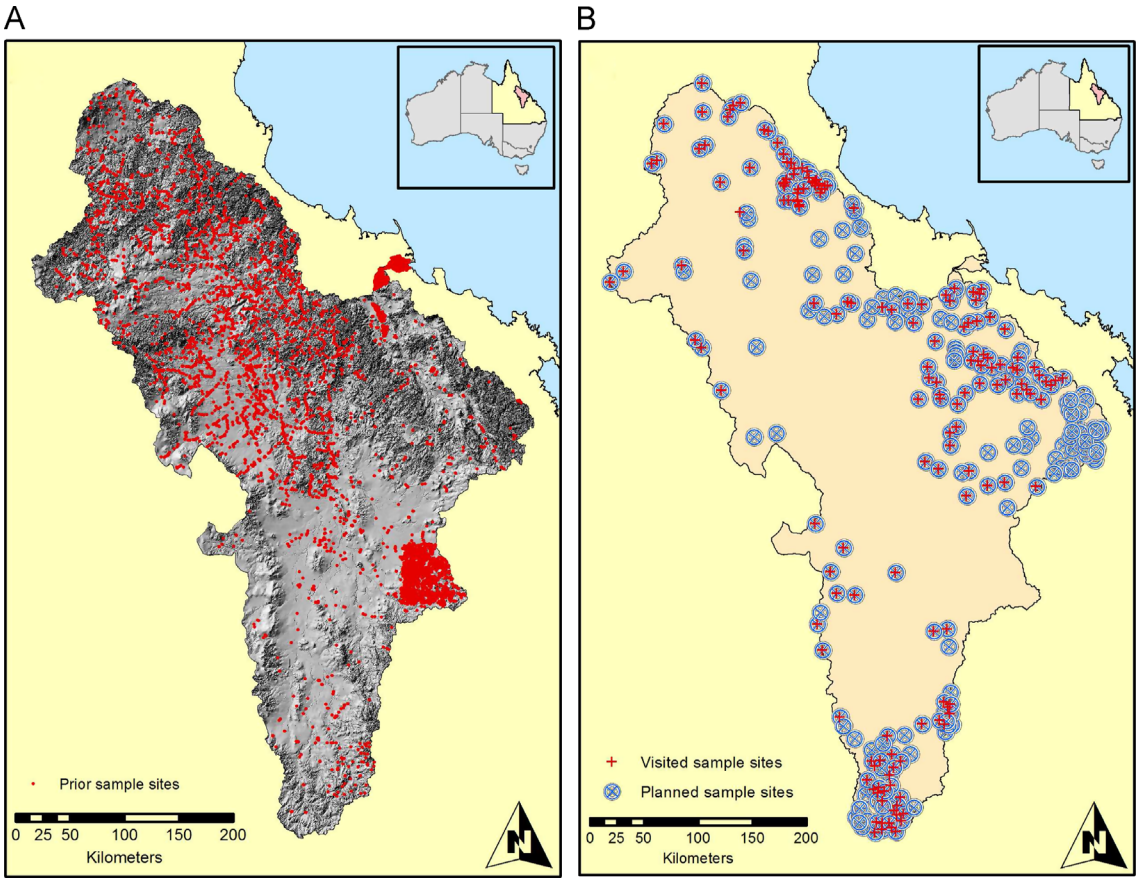

Fig. 1. (A) Locations of 8669 previously sampled sites (red dots) within the Burdekin catchment with hillshade. (B) Locations of $n=300$ planned sampling sites selected by flexible Latin hypercube sampling (blue ⊗) of which 168 have already been visited (red +). Insets in both panels show the location of the catchment within Australia. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the catchment. The ultimate goal of this study is to improve soil information in the grazing areas to improve the management of erosion sources; we aim to achieve this using DSM approaches for efficiency. Central to this concept is the initiation of a new soil survey, designed to efficiently fill the gaps in the catchment space as well as in the covariate space that arise from previous surveys.

Though there are 8669 sites already in existence (see Fig. 1A), they are not evenly distributed around the catchment as they were typically collected to assess land suitability for cropping. Furthermore, previously sampled sites tend to be strongly biased to locations close to roads and tracks. We wished to sample an additional $n=300$ sites prior to mapping soil erodibility. Due to the poor trafficability of the landscape it was preferable for new sampling sites to be within 1.5 km of established roads and tracks.

2.2. Defining and finding optimal sites

We assembled $k=8$ raster surfaces of environmental covariates to be used in the modelling process (Table 1). These surfaces were chosen as they were considered to be representative of landscape processes driving soil erosion; additionally, they were readily available, were at an appropriate scale and covered the entire study area. All rasters were re-sampled to the same spatial extent and dimensions (142 million pixels of size 30 m × 30 m).

We wish to select n sites within the Burdekin catchment for soil sampling. We wish to choose these sites to optimise $k+2$ criteria, where, as above, k is the number of covariates. The first set of k criteria involves choosing our target sites so that the covariate space spanned by the combined group of target sites and previously sampled sites matches the covariate space of the region as best as possible. We achieve this by comparing the distribution of covariates from prior and target sites with the distribution of covariate values for the entire region. With large numbers of sites it is natural to examine histograms to visually compare the distributions. The observed (O_i) and expected (E_i) proportion of sites within each histogram bin (indexed by i) can be compared using the formula $\sum_i O_i(\log_e O_i - \log_e E_i)$. This formula has become known as the Kullback–Leibler (KL) divergence of the expected proportions of sites from the observed proportions of sites (Kullback and Leibler, 1951), with the convention that $0 \log_e 0$ is interpreted as zero as that is the limit of $x \log_e x$ as x decreases to zero. The KL divergence is an example of a metric specifically designed for the comparison of distributions. The KL divergence is defined for both categorical and numerical covariates. In the case of a categorical covariate the bins used correspond to the levels of the covariate; for a numerical covariate we used 25 bins.

Table 1
Environmental covariates used in the study.

	Environmental covariate	References
1–3	P, Th, and K gamma radiometrics, respectively	Minty et al. (2009)
4	Hydrologically enforced DEM – SRTM	Gallant et al. (2011) derived from Farr et al. (2007)
5	Topographic wetness index	Beven and Kirkby (1979)
6	Curvature	Zevenbergen and Thorne (1987)
7	Terrain ruggedness index	Riley et al. (1999)
8	Weathering intensity index	Wilford (2012)

The use of the KL divergence is a minor point of difference between our approach and the original cLHS design. With cLHS each covariate is divided based on its percentiles into n bins of equal size, with the aim of choosing a single site to represent each bin. Then a comparison of observed (O) bin counts with expected ($E=1$) bin counts is made using the quantity $\sum_i |O_i - E_i|$ where i is an index over the bins. This metric penalises all departures from expected values equally. With the KL divergence the penalty of missing a site from the tail of a distribution is larger than the penalty of missing one site close to the mode. With k covariates to consider we have k different KL divergence criteria that we wish to optimise.

The next criterion we wish to optimise is the ease of accessing our target sites. Our selection process will ideally favour sites that are easier to access compared to alternatives that are equal in all other ways. We can achieve such a design by considering the metric $\sum_j r_j$, where r_j is the straight line distance from site j to the closest road. We chose distance from the site to the nearest road as a proxy variable that reflects ease of access. Other proxies that take additional topographic information such as slope and curvature could also be used. The effort required to access a particular site has been modelled based on gradient, landcover and distance (Roudier et al., 2012), however, for broad-scale sampling in remote areas it is not likely that this level of information is readily available at the required resolution.

Finally, as mapping of soil properties across a region is our ultimate goal, we wish to choose sites in a manner that ensures that prior and target sites will span the entire geographic region in as balanced a manner as possible. We chose to optimise this through a dedicated metric of the form $\sum_j d_j^{-1}$ where d_j is the Euclidean distance from site j to its closest neighbouring target or prior site.

Optimising any one of these criteria is relatively easy but it is usually not possible to optimise all $k+2$ criteria simultaneously. One way to optimise multiple criteria, also known as Pareto optimisation, is to combine them in a weighted sum and minimise that linear combination. In such an approach the choice of weights can lead to very different outcomes. Ideally the components of the combined criteria are scale-free. That way, for example, the choice of whether spatial spread is a function of distance in kilometres or in metres will not affect the final result. We also suggest the components of the combined criteria are translated to give the numeric values some meaning in terms of the optimisation.

In designing our study for the Burdekin catchment we combined our optimisation criteria f_i by first rescaling each using its median value m_i and its greatest lower bound L_i. Our single optimisation criterion takes the form

$$F = \sum_{i=1}^{k+2} \frac{f_i - L_i}{m_i - L_i}.$$

We choose to work with medians instead of means as they are less sensitive to outliers. The median values of each criterion are estimated by simple random sampling (SRS) of n sites from the accessible region. The median value of the individual components of F under SRS of n sites will be 1 and so the median value of F is equal to the number of components in its summation, which in this case is $k+2$.

The greatest lower bounds for both ease of access and KL divergence are 0. When a lot of prior information is available the greatest *achievable* lower bound for KL divergence may be a lot higher than 0. Finding a lower bound for spatial spread is not so straightforward and it needs to be estimated. We estimated it by repeated sampling of n sites under SRS. We set L as half the lowest observed value for the spatial spread criteria from such a simulation but a more conservative procedure is advised. As already noted, 0 is also a lower bound for this criterion but such a value may not be achievable.

In an ideal world, all criteria can be optimised simultaneously to their respective lowest values, that is, $f_i = L_i$, and so each of the individual rescaled components of F, and therefore their sum, will be 0. The final value of F once optimisation is complete will lie between 0 and $k+2$ and will give an indication of how much better we have done compared with SRS and also how close our global optimum is to the ideal optimum. We have combined each of our rescaled criteria with equal weights; other weights can also be used and we return to this issue in Section 4.

To optimise F it is impossible to search the entire space of all possible sets of n sites. Instead, we use simulated annealing (Press et al., 2007) to find a set of n sites that approximate the global optimum. Simulated annealing is a computational optimisation routine designed for combinatorial problems such as this and has been used previously to optimise spatial sampling of soil properties (van Groenigen et al., 1999). We use linked parallel simulated annealing running on 12 cores that report current states after blocks of 250 iterations. When no improvements in state were found after several blocks the algorithm was deemed to have converged. We alter our set of sites at each step using one of several strategies – by dropping one site at random and replacing it with a nearby site, by dropping one at random and replacing it with a new site that is not nearby, by dropping sites that perform poorly in terms of spatial spread or ease of access. The proportions of time these strategies are taken are 50%, 40%, 5%, and 5% respectively. At the time of writing, our custom code built for this problem was taking on the order of 12 h to reach convergence, typically requiring 1 million iterations on each of 12 cores. This computation time and effort is low in comparison to the time and effort spent collecting samples in the field but improvements in computation time would be desirable.

Once the set of n target sites have been chosen we wish to create maps to highlight alternative sites close to each target. We created maps to cover a neighbourhood of 40 ha centred at the point on the road that is closest to our target. The mapped value at each site in the neighbourhood should be *related* to F for the combination of the $n-1$ other target sites and this particular site. Instead of mapping F itself for these maps, consider that they only come into use when our soil-sampling team will be in the locality. As such we ignore local changes in the spatial spread component of F by fixing that component to the value associated with the local target site. We use F_l to denote the mapped quantity for the locality.

Experience in the field indicated that maps of F_l are not the most efficient manner to present alternative sites for selection. Field crews reported that the information is better summarised by highlighting only the locations of sites that are as good as the target and ones that are decent alternatives. This classification was created by considering a histogram of the F_l values and noting the rank associated with our target site. By drawing bootstrap samples of the same size we can find the values of F_l that occupy the same rank as the target site. The sites within our locality associated with these values indicate sites that are as good as our target. This process is repeated to fit sites equivalent to the site with the lowest value of F_l that is not as good as the target. Such sites are labelled as decent alternative sites to the target.

3. Results

Our final optimisation criteria value is based on $k=8$ covariates (Table 1) plus criteria for spatial spread and ease of access. By design, when selecting $n=300$ sites to sample in conjunction with data from 8669 previously sampled sites, the median value of our

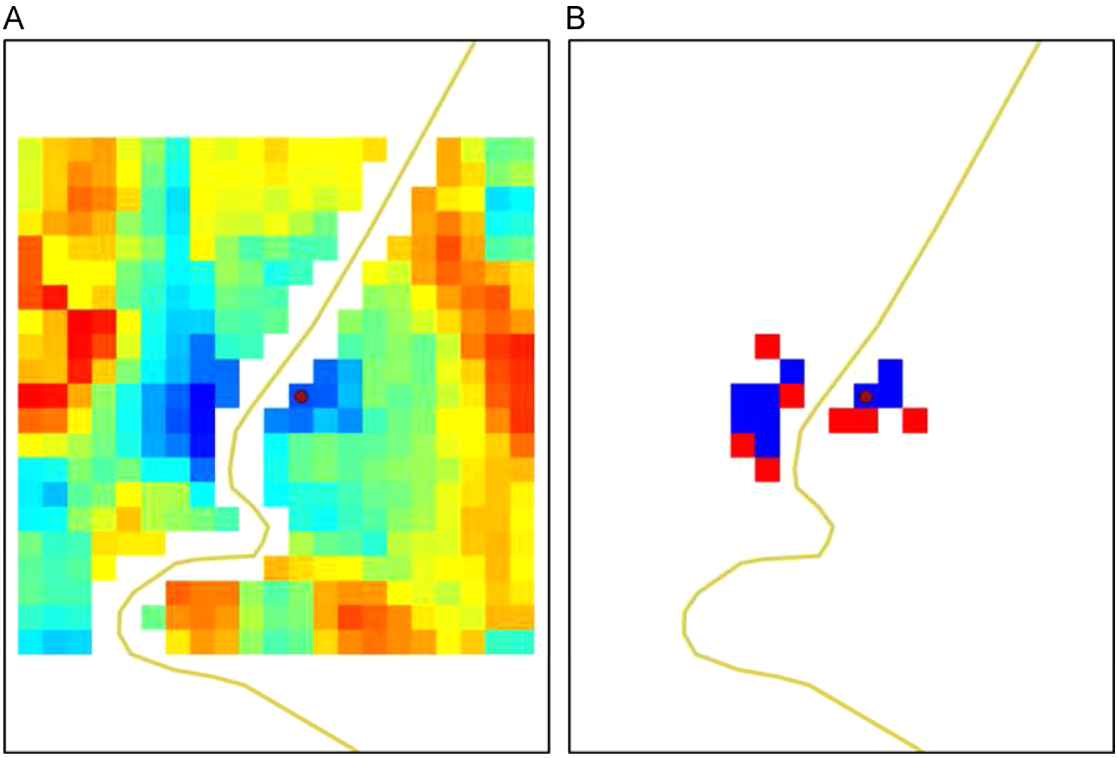

Fig. 2. Maps of local alternatives around our target site indicated by the red circle; access to the sites is via the road (yellow path through map). Panel A is a map of F_1, where blue colour correspond to lower values. Panel B highlights the sites considered to be as good as our target (blue squares) and those that are decent alternatives (red squares). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

criterion under SRS should be $k+2=10$. Ten thousand sets of $n=300$ sites were selected by SRS from the accessible region. After trimming off extremely large outliers due to the form of the spatial spread criteria, the range of values for F for these sites goes from 9.6 to 10.5. After optimisation we were able to lower the value of F to 6.83 – lower than what can be achieved by SRS. The $n=300$ sites selected for sampling by flexible Latin hypercube sampling, shown on the right-hand side of Fig. 1, are generally located in the hills around the catchment boundary where prior surveys did not collect samples. The sampling programme is ongoing and 168 sites that have been visited at the time of writing are also highlighted in Fig. 1.

After simulated annealing was deemed to have completed we created $n=300$ 40-ha maps of local alternative sites around each target, to consider in case a target site is inaccessible. A detailed examination of these sites indicates that 17% of these localities contain a site that would have been a better choice than the target site within that locality. This indicates that we were premature in ceasing our annealing procedure. It also points out that the proposal of new sites within the annealing algorithm may be improved by occasionally doing a detailed local search around the set of target sites at that iteration of the annealing algorithm.

The maps of local alternatives also indicate sites that that are "as good as" the target. The number of such sites differs depending on the target and availability of good alternatives. The most common number of sites found to be as good as the target is 4 but the average number of such sites is 8. Furthermore, the most common number of sites that can be considered a decent alternative to the target is 10 with the average being 16. Fig. 2 shows examples of such maps. In Fig. 2A we plot the criterion value at each alternative site. In Fig. 2B we highlight only sites that are to be considered as alternatives.

Because of the large number of good and decent alternative sites available for consideration, our final design is robust to issues related to access at individual targets. We performed a simulation study to examine more realistic scenarios where the failure to reach target sites was up to 50%. We considered three scenarios. For a given failure rate p, alternative sites were chosen from good alternatives (Scenario 1), decent alternatives (Scenario 2) and by SRS (Scenario 3) from the local area. Fig. 3 indicates the average (solid lines) and 95% confidence intervals (dashed lines) for the values of F for values of p from 0 to 0.5. We see only small changes in the value of F under scenarios 1 and 2 indicating how robust our design is to repeated failures to reach the target site.

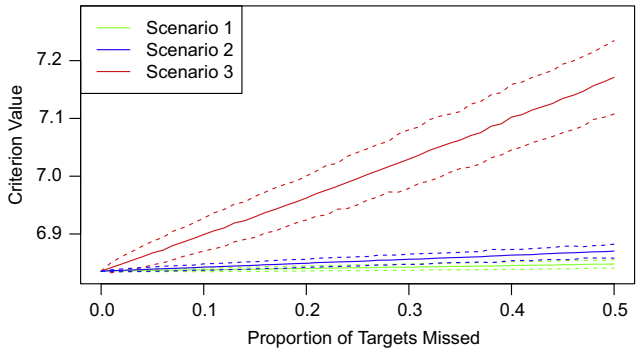

Fig. 3. Examination of robustness of our design. Average criterion value (solid lines) and 95% confidence interval (dashed lines) are plotted as a function of the proportion of target sites that are missed under three scenarios for how to select an alternative site. The alternative sites were chosen from good alternatives (Scenario 1), decent alternatives (Scenario 2) and by SRS (Scenario 3) from the local area (see text for more details).

Fig. 4A highlights for one covariate (wetness index) how the set of previously sampled sites (represented by the histogram) miss part of the covariates space (indicated by the density). Fig. 4B shows a histogram of Wetness Index values for the $n=300$ sites selected by flexible Latin hypercube sampling. These sites fill the gap highlighted by Fig. 4A.

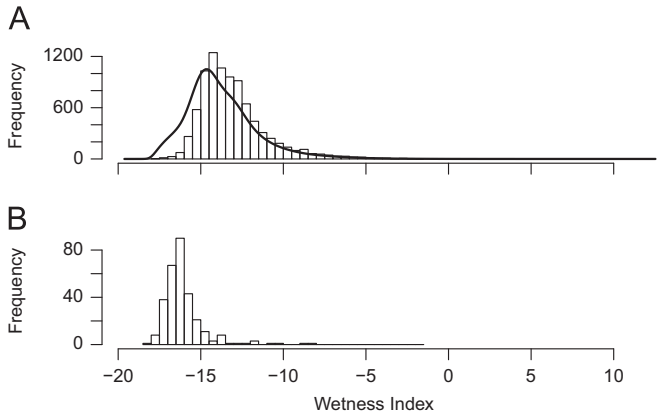

Fig. 4. (A) Histogram of wetness index values of 8669 previously sampled sites together with density of wetness index values for the Burdekin catchment (these distributions are related to covariate spaces 3 and 1 as listed in Section 1.1). The previously sampled sites clearly do not span the entire range of wetness index values. (B) Histogram of wetness index values of 300 target sites (covariate space 4 as listed in Section 1.1). *Y*-axis scales are different in the two panels. The space spanned by the combined 8969 sites is covariate space 5 listed in Section 1.1.

4. Discussion

Any approach to the design of soil surveys using Latin hypercube sampling is based on several assumptions. Firstly, we are assuming that the covariates chosen are driving the soil properties of interest in this region, or they are acting as proxies for what is driving soil properties. Implicit in this assumption is that covariate layers of sufficient quality are available for the entire region of interest at an appropriate spatial scale.

Some readers may be concerned that whole swathes of the region are deemed inaccessible. From a probability-based design and analysis point of view, see for example Cochran (1977), inference outside the accessible region would be biased and invalid as the inclusion probabilities of sites in this region are zero. Fortunately DSM is a model-based approach to analysis (Diggle and Ribeiro, 2007) so this issue does not arise. For DSM to be valid we implicitly assume that any difference in landscape between the region that is deemed accessible and the region that is deemed inaccessible is unrelated to soil function. Such an assumption would be foolish in regions where large portions of the landscape are not suitable for road-building. How to test this assumption within the context of digital soil mapping is an open research question but a good starting point includes approaches based on Mahalanobis distance, such as those used within the applied spectroscopy literature (Whitfield et al., 1987). Such an approach could be used here to discover if inaccessible regions of geographic space occupy unsampled parts of the covariate space.

Our combined optimisation criterion places equal weight on spatial spread, ease of access and matching of the individual covariate spaces. In practice one may wish to consider a weighted combination of these components if one feels strongly about optimising one component more than others. For example, data-mining of prior information may rank the covariates in terms of their importance as predictors of soil properties. Such ranks can be used to specify the weights for the combined optimisation criterion.

Our combined criterion ignores any correlation between covariate layers that may be present. In contrast, the original cLHS approach does attempt to match the sample correlations to those of the global region. In reality, since the number of correlations among k covariates is $1/2k(k-1)$, the effort put into matching the sample correlations to the global ones is a token effort relative to the effort spent matching the Latin hypercube conditions. Preserving the correlation structure will help in modelling soil properties after collection has taken place. If correlation between covariate layers is high we suggest that principal component analysis can be used to convert k correlated covariate layers into k or fewer uncorrelated principal covariates, and the sampling design is carried out in that space. It makes sense in such a setting to have a weighted optimisation criterion where the weights are related to the percentage of variation explained by each principal component (Hengl et al., 2003). The final set of sites will form a Latin hypercube design in the principal component space and will also span the original covariate space. The subsequent analysis does not have to be carried out within the principal component space so there is no loss of interpretability of the covariates in this suggestion; differences in soil properties can be related to environmental drivers themselves.

One criticism of our combined optimisation criterion is that it is overly complex in its reliance on known and estimated values for the median and minimum values of the component criteria. An easier way to achieve spatial spread would be to simply include latitude and longitude as covariates to be matched using the KL divergence. The most important characteristic of F is that it is scale-free, and such a characteristic can be more easily achieved by considering a combined criterion of the form $\sum_{i=1}^{k+2}\log_e f_i$ (McGree et al., 2008). To see that this combined criterion is scale-free consider the task of optimising when all covariates are on different scales. In that case the combined criterion will be $\sum_{i=1}^{k+2}\log_e a_i f_i = \sum_{i=1}^{k+2}\log_e f_i + \sum_{i=1}^{k+2}\log_e a_i = \sum_{i=1}^{k+2}\log_e f_i$ plus a constant associated with the change in units. Ultimately one is optimising the same combined criterion regardless of scale. While we lose the interpretation associated with the value of the combined criteria, that loss is offset by not being required to estimate the medians and greatest lower bounds.

The use of our maps of local alternatives is subject to two additional caveats. Firstly, having arrived at a target location we wish to choose the best alternative in the area and so we ignore any changes to the optimisation criterion to do with the spatial spread. Secondly, these maps have been created under the assumption that all other targets are successfully sampled, and with failure rates as high as 40% such an assumption is not going to be true. This is the reason we investigated the robustness of these designs via simulation. It is not hard to envisage a software system that updates such maps on-the-fly as we visit our target regions, but this was outside the scope of what we could implement in this project. Adaptive sampling surveys have been proposed previously for small-scale projects where sampling of an area takes place in distinct phases (Marchant and Lark, 2006). Sampling in large areas ideally requires that revisits to remote sites are not required.

It is a common practice to release software to accompany the publication of a paper that describes a new computational methodology. Our code is heavily tied to our specific application in the Burdekin catchment and, as such, is unsuited for release. A general implementation of our code is not currently available though we are happy to work with others to achieve its release. Such code should include improvements we had no time to implement, and could include different ways to specify which sites are accessible, different methods for combining our optimisation criteria and allow for the availability of spatial covariate information in different formats.

5. Conclusions

We have presented an approach for pursuing Latin hypercube sampling for large regions that pose significant practical field

challenges while taking prior information into account. Our approach explicitly allows for the failure to reach targets by objectively ranking local alternative sites. A simulation study showed that when failure rates are up to 50% there is little degradation in the final set of sampled sites compared with the original set of target sites.

By way of a practical example of this methodology, we described a remote-area soil survey implemented in the Burdekin catchment, Australia. The methodology received a positive response by surveyors in the field, principally because of the provision of alternative and known-statistically robust sampling sites when target sites were difficult to reach and time was short. Finally, we described modifications to the survey design methodology that we would implement in the light of these results, to further enhance the statistical robustness of the technique.

Acknowledgements

We wish to thank the field crews working in the Burdekin, Flinders, and Gilbert catchments of northern Queensland for their input and feedback of our choice in sampling locations. Rebecca Bartley, Dan Brough, and Mark Thomas have provided valuable comment on early drafts of this report. This research was funded by the Terrestrial Ecosystem Research Network (TERN) and the Reef Water Quality Science Program.

References

Beven, K., Kirkby, M., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24, 43–69.
Brodie, J., Waterhouse, J., Schaffelke, B., Johnson, J., Kroon, F., Thorburn, P., Rolfe, J., Lewis, S., Warne, M., Fabricius, K., McKenzie, L., Devlin, M., 2013. Reef Water Quality Scientific Consensus Statement 2013. Department of the Premier and Cabinet, Queensland Government, Brisbane (⟨http://www.reefplan.qld.gov.au/about/assets/scientific-consensus-statement-2013.pdf⟩)
Brus, D.J., Gruijter, J.J., de Walvoort, D.J.J., de Vries, F., Bronswijk, J.J.B., Römkens, P.F.A. M., de Vries, W, 2002. Mapping the probability of exceeding critical thresholds for cadmium concentrations in soils in the Netherlands. J. Environ. Qual. 31, 1875–1884.
Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353.
Cochran, W.G., 1977. Sampling Techniques, 3rd edition Wiley, New York.
Diggle, P.J., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer
Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The shuttle radar topography mission. Rev. Geophys. 45, RG2004.
Gallant, J., Chan, F., Marinova, D., Andersson, R., 2011. The ground beneath your feet: digital elevation data for today and tomorrow. In: Proceedings of the MODISM2011, 19th International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand. Perth, Australia, pp. 70–76.
Helton, J.C., Davis, F.J., 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab. Eng. Syst. Saf. 81, 23–69.
Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Res. 41, 1403–1422.
Hewitt, A., McKenzie, N., Grundy, M., Slater, B., 2008. Qualitative survey. 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, Canberra, pp. 285–306
Iman, R.L., Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among input variables. Commun. Stat. – Simul. Comput. 11, 311–334.
Iman, R.L., Helton, J.C., 1988. An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal. 8, 71–90.
Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill Book Company, New York, NY, USA
Kidd, D., Webb, M., Grose, C., Moreton, R., Malone, B., McBratney, A., Minasny, B., Viscarra-Rossel, R., Cotching, W., Sparrow, L., et al., 2012. Digital soil assessment: guiding irrigation expansion in Tasmania, Australia. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, 10–13 April 2012, CRC Press, p. 3.
Kidd, D.B., Malone, B.P., McBratney, A.B., Minasny, B., Webb, M., Grose, C.J., Moreton, R.M., 2014. Adaptive sampling for operational digital soil assessment using fuzzy *k*-means covariate stratification. Geoderma (submitted for publication)
Kroon, F., Kuhnert, P.M., Henderson, B., Kinsey-Henderson, A., Turner, R., Huggins, R., Wilkinson, S., Abbott, B., Brodie, J.E., Joo, M., 2010. Baseline pollutant loads to the Great Barrier Reef. CSIRO Water for a Health Country (⟨http://www.csiro.au/Portals/Publications/Research–Reports/GBR-baseline-loads.aspx⟩)
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.
Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311.
Lin, Y.-P., Chu, H.-J., Wang, C.-L., Yu, H.-H., Wang, Y.-C., 2009. Remote sensing data with the conditional Latin hypercube sampling and geostatistical approach to delineate landscape changes induced by large chronological physical disturbances. Sensors 9, 148–174.
Louis, B.P., Saby, N.P.A., Orton, T.G., Lacarce, E., Boulonne, L., Jolivet, C., Ratié, C., Arrouays, D., 2014. Statistical sampling design impact on predictive quality of harmonization functions between soil monitoring networks. Geoderma 213, 133–143.
Marchant, B.P., Lark, R.M., 2006. Adaptive sampling and reconnaissance surveys for geostatistical mapping of the soil. Eur. J. Soil Sci. 57, 831–845.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
McGree, J.M., Eccleston, J.A., Duffull, S.B., 2008. Compound optimal design criteria for nonlinear models. J. Biopharm. Stat. 18, 646–661.
McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.
Minasny, B., Mc Bratney, A.B., 2002. Uncertainty analysis for pedotransfer functions. Eur. J. Soil Sci. 53, 417–429.
Minasny, B., McBratney, A., 2010. Conditioned Latin hypercube sampling for calibrating soil sensor data to soil properties, Proximal Soil Sensing. Springer, pp. 111–119
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388.
Minty, B., Franklin, R., Milligan, P., Richardson, M., Wilford, J., 2009. The radiometric map of Australia. Explor. Geophys. 40, 325–333.
Mulder, V.L., de Bruin, S, Schaepman, M.E., 2013. Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 21, 301–310.
Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics 41, 303–312.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, New York.
Riley, S.J., DeGloria, S.D., Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Int. J. Sci. 5, 23–27.
Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, 10–13 April 2012, CRC Press, p. 227.
Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. Technometrics 29, 143–151.
Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., Malone, B.P., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma 213, 15–28.
Thomas, M., Odgers, N., Ringrose-Voase, A., Grealish, G., Glover, M., Dowling, T., 2012. Soil survey design for management-scale digital soil mapping in a mountainous southern Philippine catchment. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, 10–13 April 2012, CRC Press, p. 233.
van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.
Van Meirvenne, M., Goovaerts, P., 2001. Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. Geoderma 102, 75–100.
Vašát, R., Heuvelink, G.B.M., Borøuvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155, 147–153.
Viscarra Rossel, R.A., Chen, C., 2011. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. Remote Sens. Environ. 115, 1443–1455.
Whitfield, R.G., Gerger, M.E., Sharp, R.L., 1987. Near-infrared spectrum qualification via Mahalanobis distance determination. Appl. Spectrosc. 41, 1204–1213.
Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183–184, 124–142.
Wooldridge, S.A., 2009. Managing local water quality to help combat climate change impacts on the Great Barrier Reef, Australia, Report to the Marine and Tropical Sciences Research Facility. Reef and Rainforest Research Centre Limited, Cairns (⟨http://www.rrrc.org.au/publications/downloads/25i4-AIMS-Wooldridge-S-2009-June-Milestone-Report.pdf⟩)
Worsham, L., Markewitz, D., Nibbelink, N.P., West, L.T., 2012. A comparison of three field sampling methods to estimate soil carbon content. For. Sci. 58, 513–522.
Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Processes Landf. 12, 47–56.