Geoderma Regional 4 (2015) 1–10

Contents lists available at ScienceDirect

Geoderma Regional

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

Operational sampling challenges to digital soil mapping in Tasmania, Australia

Darren Kidd [a,b,*], Brendan Malone [b], Alex McBratney [b], Budiman Minasny [b], Mathew Webb [a,b]

[a] Sustainable Landscapes Branch, Department of Primary Industries Parks Water and Environment Tasmania, 171 Westbury Road, Prospect, Tasmania 7250, Australia
[b] Faculty of Agriculture and Environment, University of Sydney, 1 Central Avenue, Australian Technology Park, Eveleigh, NSW 2015, Australia

ARTICLE INFO

Article history:
Received 6 August 2014
Received in revised form 14 November 2014
Accepted 14 November 2014
Available online 20 November 2014

Keywords:
Digital soil mapping
Soil sampling
Land suitability
Fuzzy k-means
Solonetz
Lixisols
Vertisols
Nitisols
Acrisols
Luvisols
Phaeozems
Gleysols
Fluvisols

ABSTRACT

Digital soil mapping (DSM) was used to generate soil property surfaces at 30 m resolution for Tasmanian Government Land Suitability Modelling in Tasmania, Australia. Soil predictions were required for pH, EC, clay percentage, stone content, drainage, and depth to sodic and impeding layer. Empirical modelling using a suite of environmental covariates and the relevant soil attribute data from field-collected soil cores was used to generate the digital maps. Environmental covariates included: SRTM DEM and derivatives, gamma radiometry, legacy soil maps, surface geology, and multi-spectral satellite imagery.

An integral component of any DSM process is a sound sampling design that represents the full range of environmental variables used. However, in cases where there are operational constraints, the approach needs to remain flexible, efficient, and compatible with project area land use and terrain. In two separate study areas, a combined 700 training and 230 validation sites were sampled over 70,000 ha. A conditioned Latin hypercube (cLHS) sampling design was used for the initial sampling for DSM training sites, with 'contingency sites' created for alternative sampling if access was constrained. The pre-defined ('strict') sample locations proved difficult to implement in the field, with a variety of access issues making sampling slow and arduous. In an attempt to increase sampling progress rates to meet tight project milestones an alternative 'relaxed' sample design based on random sampling of fuzzy k-means covariate clusters (strata) was used for the second study area. A map of clusters provided to soil sampling staff allowed difficult sites to be relocated within the same cluster type, maintaining stratification. The relaxed approach still adequately represented the covariate distribution while providing greater flexibility to site placement. This paper provides background to the Tasmanian DSM project, some discussion of sampling designs for DSM, and the pros and cons of their implementation in the field with due consideration of operational constraints in a Tasmanian case study, highlighting the need for sampling flexibility within 'real-world' conditions.

1. Introduction

Recently, there has been a growing concern over food security in Australia where it is feared that food prices could rise by as much as 50% in the next decade. This is mainly due to a potential scaling back of production in the Murray–Darling Basin as it faces both climate change and a reduction in water allocation for irrigation. Tasmania is seen as a potential and significant part of the solution, with its predicted warming climate allowing a wider variety of food crops to be grown, and a surfeit of water resources. Steps are being made to develop the state as an important new agricultural production area for Australia and the region by development of new irrigation areas, with the aspiration of growing a wider variety of food crops. The basis for the planned development is the efficient and sustainable management, movement, and use of water through new irrigation networks.

The 'Wealth from Water' Project commenced in November 2010 to support irrigated agricultural expansion through land suitability mapping, using digital soil assessment (Carré et al., 2007). It was a partnership between the Tasmanian Department of Primary Industries, Parks, Water and Environment (DPIPWE), the Department of Economic Development, Tourism and the Arts (DEDTA), the Tasmanian Institute of Agriculture (TIA), ACLEP (the Australian Collaborative Land Evaluation Program), and the University of Sydney (through an Australian Research Council Linkage Project). Commencing in the Tasmanian Meander Valley (43,000 ha) and Midlands (Tunbridge, 27,000 ha) irrigation districts, Enterprise Suitability Rules were developed by TIA for 20 enterprises using Tasmanian agricultural research trials, existing literature, and consultation with industry experts. Enterprises included: alkaloid poppies, carrots, hazelnuts, barley, blueberries, pyrethrum, and commercial hemp. The suitability rule-sets required soil property and climate parameters, including pH, EC (electrical conductivity), clay content, depth to sodic layer, depth to impeding layer, stone content, drainage class, frost-risk, chill hours, and growing-degree days (Kidd et al., 2012).

* Corresponding author at: PO Box 46, Kings Meadows, Tasmania 7250, Australia.
E-mail address: Darren.Kidd@dpipwe.tas.gov.au (D. Kidd).

http://dx.doi.org/10.1016/j.geodrs.2014.11.002
2352-0094/

There are now sufficient published examples describing the prediction of soil property surfaces using digital soil mapping (DSM) methodologies based on the *scorpan* approach (McBratney et al., 2003), to make this a scientifically-valid operational approach. These predicted surfaces can provide continuous and quantitative soil property estimates (as opposed to conventionally-derived polygonal soil type surfaces), also having the advantage of statistical validation and associated uncertainty of prediction. Soil property mapping using these methodologies was considered the optimal approach to provide suitability model inputs within available time and resources. An integral component of a DSM process is a sound sampling design that ensures calibration and validation sites are representative of the full distribution of the covariates used for prediction. Ideally sampling should encompass the full range of environmental conditions within a study area. Doing this will limit the subjectivity inherent in traditional sampling approaches such as free-survey (National Committee on Soil and Terrain, 2009). However, for operational endeavours such as the Wealth from Water Project, the sampling approach needs to remain flexible, efficient, and compatible with project area land use and terrain. Large mapping areas will require even greater operational flexibility. Such operational projects often have limited budgets, are time-constrained, and require efficiencies in field effort, often the most expensive component in land resource assessment. The common DSM approach to sampling using a 'strict' sampling design with pre-determined coordinates is often difficult and time-consuming to apply, with numerous access constraints either slowing progress or preventing sampling at desired locations. This paper documents an operational DSM case study, the logistical problems encountered using a popular pre-defined sampling strategy, and the interim solution developed and applied within the tight project time-constraints. The approach used covariate stratification for a randomised sample design which allowed physically impractical sites to be manually re-located within the field to more accessible locations within corresponding strata, while still maintaining the same number of samples from each of the strata types. The thrust of this paper is not to provide an exhaustive review and comparison of the multitude of sampling techniques developed for predictive soil mapping, but to discuss the problems inherent in real-world soil sampling, and document the pragmatic methodological compromise used to improve operational sampling speed and efficiency, while still providing representation of the environmental co-variables used for predictions.

1.1. Soil sampling approaches

Strategies used for soil sampling design generally include; traditional and subjective free-survey for conventional soil landscape, or soil association mapping (National Committee on Soil and Terrain, 2009); geostatistical approaches, that evenly sample the physical geographic space; and techniques developed for digital soil mapping which sample the entire covariate feature space (Minasny and McBratney, 2006; Vašát et al., 2010). Sampling optimisation across the full range of predictor or explanatory variables (covariates) is necessary to maximise environmental correlation (McKenzie and Ryan, 1999). Brus (2010) differentiated between design-based and model-based approaches; design-based sampling mainly uses a statistical approach where a random component is essential in the selection of sampling locations, and the inference is based on the selection probabilities. This is useful if there is a need to know the status or the change in soil properties over an area, e.g. monitoring soil carbon. A model-based approach presumes that the unknown soil attribute value at any location is random; if there is a requirement for mapping or knowing how the soil properties vary in the field the model-based sampling approaches are commonly used.

A sampling strategy can either be undertaken in terms of optimally covering the geographical space, the covariate feature space, or both. There has been some debate as to whether geographic constraints, i.e. spacing or dispersion of the sampling design, or perhaps incorporation of coordinate positions as covariates, is warranted (Minasny and McBratney, 2006). The accuracy of estimating the spatial means of an environmental variable can be increased by dispersing the sample locations uniformly across the study area (Walvoort et al., 2010). However, the need for the spatial dispersion of sample locations could be diminished when using environmental variables for predictions, or when environmental predictors are known and available (Brus et al., 2006), that is, the sampling design is based on the covariate distribution of values.

A popular sampling method used in DSM is the 'conditioned Latin hypercube' (cLHS), a purposive model-based sampling approach that maximally stratifies the full multivariate distribution, where the sample distribution closely replicates the covariate distribution (Minasny and McBratney, 2006). However, such pre-determined, 'strict' sampling methods can be inflexible with little room for alternative site selection in the field. This can be exacerbated when sampling intensively-used agricultural land due to a range of access constraints, such as farmer consent, infrastructure, contamination, travel distance and management phase. Logistical and operational problems have been documented using 'strict' approaches elsewhere; Roudier et al. (2012) incorporated operational constraints into the cLHS design where sampling costs were assimilated as a consideration of distance to roads for ease of access, while Thomas et al. (2012) encountered access difficulties due to extreme terrain, travel distance and vegetation cover while sampling mountainous, heavily vegetated landscapes.

Clifford et al. (2014) also identified operational sampling problems using a pre-defined sampling regime in a large and remote study region in Queensland, Australia, totalling 12.8 million ha. In response, they developed a 'flexible Latin hypercube sampling (LHS)' approach and simulated efficiencies in field effort that potentially increase soil sampling rates with respect to resourced time-constraints. Clifford et al. (2014) aimed to optimally cover the covariate feature space while targeting more easily accessible sites (constrained to buffers around formed roads and tracks), and providing alternative nearby sites (covering a surrounding area of 40 ha) for consideration when initial sampling sites are inaccessible. The flexible LHS approach was developed and documented after completion of the Tasmanian field campaign described in this paper, so was therefore not considered in this project.

Due to the unforseen time taken to carry out an initial cLHS sampling campaign within our case study, a timely and alternative solution was needed to ensure that remaining field sampling was completed by the strict project milestones, and ensure field-work was completed before many areas became too wet to sample due to expected seasonal rain. It was chosen to use 'fuzzy k-means' (FKM) clustering of covariates as sampling stratum, where target sites were equally distributed by number within each stratum, and field staff could move sites within the mapped clusters to maintain stratification and representative covariate distribution.

1.2. k-Means stratification of covariates

k-Means is a popular clustering methodology for multivariate analysis which determines clusters based on multivariate centroids, minimising the mean squared distance between objects and the closest centroid values (Brus et al., 2006; Hartigan, 1985; MacQueen, 1967). Multivariate within-cluster variance is optimised to be as small as possible for each cluster, grouping very similar attribute values for each cluster, and small spatial distances between them for spatially-structured datasets (Burrough et al., 2000). Fuzzy k-means (FKM) is an advanced option of 'hard' k-means where each observation has a degree of belonging to clusters. Burrough et al. (2000) demonstrated the use of FKM for partitioning soil–landscape data, useful for prediction of discrete properties or soil types with boundary overlaps. It has also been used for sampling design, both for geographical clustering, when no environmental variables are used for predictions (Brus et al., 2006), and feature, or covariate stratification (Minasny and McBratney, 2006). However, FKM is not able to accommodate categorical variables,