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

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, therefore un-ordered categorical data, such as soil type or geological mapping, are not able to be used as sampling covariates.

The classical FKM algorithm, as succinctly described by Bezdek et al. (1984), computes a cluster membership function, a value between 0 and 1, which determines the probability that a pixel or raster cell belongs to a particular cluster. The algorithm undergoes a series of iterations, where the value of a fuzzy exponent is set to a value greater than 1 to control the amount of 'fuzziness' of pixel membership to each cluster. The membership function is randomly initialised to a value between 1 and 0, and the cluster centroids computed. The distances (e.g. Euclidean or Mahalanobis) between each pixel value and the cluster centroid measure are determined, and a membership function is generated. A membership value for each pixel belonging to each cluster is obtained by re-running the algorithm such that it is looped until the membership values for all pixels converge and do not change significantly. The membership value for a pixel belonging to a cluster that is closest to 1 implies that pixel most likely belongs to that cluster (Chapron, 2011; Zadeh, 1968). For a 'good' fuzzy k-means classification, the partition coefficient (the ratio of within-cluster variance to between-cluster variance) should be close to 1, and the classification entropy (information entropy) should be close to 0 (Burrough et al., 2000).

2. Aims

The aims of this paper are to;

1. Document a case study of the real-world problems and potential solutions associated with a strict (pre-determined) sampling-design in a Tasmanian operational DSM programme.
2. Test a sampling design that represents the covariate feature space, but is able to increase sampling rates when compared against a 'strict' approach.

3. Materials and methods

3.1. Project areas

Tasmania, as Australia's southern-most, and only island state has a cool-temperate climate, with mean annual rainfall averaging over 1800 mm/yr in the west, to less than 450 mm/yr in the central Midlands (Bureau of Meteorology, Australia). Population is about 500,000 people, with agriculture being one of the most economically important activities, covering a diverse range of soils and landscapes and associated native flora and fauna. The project was undertaken across two separate pilot areas; the Meander and Midlands (Tunbridge) Irrigation Districts, with a total area of approximately 70,000 ha (Fig. 1).

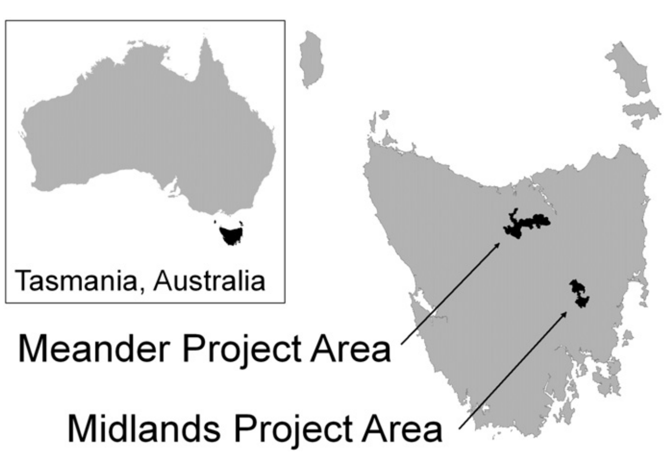

Fig. 1. Project area locations, Tasmania, Australia.

3.1.1. Meander

The Meander project area is primarily used for a wide range of agricultural practices, including grazing (cattle and sheep), dairy, cropping (cereals, vegetables and alkaloid poppies) and perennial horticulture (strawberries, raspberries and hazelnuts). Rainfall averages between 650 mm/yr in the East and over 1000 mm/yr in the West. Mean minimum temperatures range from 0.5 °C in winter to 10.4 °C in summer, and mean maximums of between 22.5 °C and 8.7 °C for summer and winter respectively. The Tasmanian central plateau and the escarpment of the Great Western Tiers mountain range dominate the landscape in the Meander area, where extensive block faulting has disrupted the surface during the lower to middle Tertiary Period. The Meander project boundary follows the Meander Irrigation Scheme and was selected because it has an inherent variety of soils, and a diversity of land uses and landscapes. Tasmanian geological structure largely determines the pattern of soils due to the strong influence of rock type upon soil formation. The Meander area contains soils of the Launceston Tertiary Basin to the East, which comprise a series of alluvial and relict river terraces in association with Smectitic clays in drainage depressions and most recent flood plains (Vertisols; IUSS Working Group, 2007), and sodic (exchangeable sodium % > 6) texture-contrast (sharp change between top-soil and sub-soil textures, with a clay increase of >20%) soil terrace series (Solonetz or Lixisols; IUSS Working Group, 2007), with various distributions of Aeolian cover sands. Red volcanic gradational soils derived from Tertiary Basalt (Nitisols or Acrisols; IUSS Working Group, 2007) dominate the landscape around the Deloraine area, while poorly drained, complex alluvial soils are found in the Meander township area to the South (Gleysols, Fluvisols, and Lixisols; IUSS Working Group, 2007). Outcrops of Jurassic dolerite are scattered through the area which produce diverse soils (Luvisols; IUSS Working Group, 2007) with abundant coarse fragments (Spanswick and Zund, 1999).

3.1.2. Midlands

The Midlands project area covers 27,000 ha of the southern part of the Midlands Irrigation Scheme, an area from Oatlands and North to Tunbridge, and is used for both grazing (predominantly sheep) and cropping (cereals and poppies). Mean minimum temperatures range from 1.5 °C in winter to 10.5 °C in summer, and mean maximums of between 24.5 °C and 11.3 °C for summer and winter respectively. The area is located at the origination of the Launceston Tertiary Basin, an ancient and relict river system that drained central Tasmania to Bass Straight in the North. The soil landscape is comprised of recent and higher level alluvial terraces, bound by Jurassic Dolerite hills to the South. Triassic Sandstone has been capped by the dolerite on foot-slopes to the East, forming dolerite and sandstone fans through alluvial areas. The area receives less than 500 mm/yr. in annual rainfall and is comprised of small areas of primary salinity and widespread sodic soils (Solonetz or Lixisols; IUSS Working Group, 2007), with black cracking clays in drainage depressions and recent flood plains (Vertisols; IUSS Working Group, 2007; Kidd, 2003). The Oatlands area is separated from the Tunbridge area by dolerite hills to the North and is scattered with areas of Triassic Sandstone and Permian mudstone (Luvisols, Lixisols, Phaeozems; IUSS Working Group, 2007; Spanswick and Kidd, 2001).

3.2. Scorpan spatial covariates

Existing legacy soil data was obtained for the Meander and Midland study areas: Quamby Soil Map (Spanswick and Zund, 1999); Interlaken Soil Map (Leamy, 1961); and Oatlands Soil Map (Spanswick and Kidd, 2001). These soil data sources and associated database site density alone were not of the scale or quality to produce reliable suitability surfaces necessary for achieving the outcomes of the project. Nonetheless, these legacy data are still useful as DSM covariate information. In addition the to the legacy soil mapping data, other available *scorpan* covariates were assembled and processed to a common 30 m grid system for

Table 1
Tasmanian spatial covariates.

Scale	Spatial covariates	Scale/resolution	Reference/source
Categorical data			
Regional	Soil map	1:100,000	Spanswick and Zund (1999); Leamy, (1961); Spanswick and Kidd (2001)
Regional	Land capability map	1:100,000	Noble (1993)
Regional	Land use map	1:50,000	DPIPWE (2012) (in prep)
Regional	Vegetation map (TASVEG) v 2.0	1:25,000	DPIPWE (2009)
Regional	Surface geology map	1:25,000	Mineral Resources Tasmania (2008)
Remote sensing			
Local	Rapid eye multispectral	5 m	Cradle Coast Authority (2010)
Local	SPOT bands 1,2 & 3	5 m	SPOT Image (2009)
Local	SPOT NDVI	30 m (processed)	SAGA GIS (system for automated geoscientific analyses, http://www.saga-gis.org), 2009
Local	LandSat principal components	30 m (processed)	SAGA GIS (system for automated geoscientific analyses, http://www.saga-gis.org), 2009
Local	Gamma radiometrics (radioactive nuclides — K, U, Th, total dose)		Mineral Resources Tasmania (2004)
Terrain			
Local	SRTM DEM-S	30 m	1 arc sec digital elevation model, adaptively smoothed, Geosciences Australia (2011)
Local	Slope, aspect, curvatures (plan & profile), topographic wetness index (TWI), multi-resolution valley bottom flatness (MR), multi-resolution ridge top flatness (MRRTF).	30 m	SAGA GIS (System for Automated Geoscientific Analyses, http://www.saga-gis.org)

the study areas using SAGA GIS (System for Automated Geoscientific Analyses, http://www.saga-gis.org), described in Table 1. Terrain derivatives were generated from the Shuttle Radar Topography Mission (SRTM) DEM (Gallant et al., 2011) using the 'Basic Terrain Analysis' SAGA module for slope, plan and profile curvature, valley depth, and aspect. Topographic wetness index was generated using the 'Topographic Wetness Index' module in SAGA, using the 'standard' settings, together with MrRTF (multi-resolution ridge-top flatness) and MrVBF (multi-resolution valley bottom flatness) (Gallant and Dowling, 2003). NDVI (Normalized Difference Vegetation Index) was generated from SPOT imagery using SAGA, where

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{VIS}) / (\mathrm{NIR} + \mathrm{VIS})$$

and NIR is the near infra-red band, and VIS is the visible red band.

3.3. Sampling

At the time of preparing the sampling locations for the Meander area, the motivation was focussed on sampling the covariate feature space rather than equally spacing the sampling locations, so that sites could be used to predict a suite of soil properties. Soil sampling was effectively a 'strict' model-based approach using available covariate data to provide the best chances of predicting multiple properties using a conditioned Latin hypercube (cLHS) sampling design. In an attempt to provide greater sampling flexibility than the predefined cLHS approach used for the initial project sampling, the fuzzy k-means *scorpan* covariate stratification approach was used for validation of soil attribute maps in the Meander area. The desired DSM outcome was to depict the soil property values of each pixel as precisely and efficiently as possible; thus, probability-based validation (Brus, 2010) was not considered necessary or attainable within project time-constraints and resources, and a model-based sampling approach deemed adequate. In the Midlands area, FKM covariate stratified sampling was used for both training, and a separate validation sample set. The approach, therefore, aimed to provide a sampling methodology that adequately represented the covariate feature space, yet remained practical and flexible for rapid operational DSM sampling.

3.4. Meander soil sample design (training)

Two-hundred training sites (data density of 5 samples per 10 km^2) were estimated as appropriate for the required resolution for Meander, a similar density to that determined by Brungard and Boettinger (2010), to be used for inference and mapping of soil properties based on available covariates and observed soil information. The cLHS algorithm (Minasny and McBratney, 2006) generated 200 locations, with 25 additional contingency target sites generated as a separate sample for situations where sampling was not possible due to access, physical sampling constraints or site contamination. Locations were produced using the covariates described in Table 1.

For training sites the cLHS 'strict' sampling design was found to be slow to implement in the Meander area due to operational constraints and the subsequent need to navigate to a completely alternative site. There was a need for a more flexible approach to increase sampling progress rates to meet project milestones, with limited time available to develop or test any novel or adaptive variations to the cLHS approach. It was decided to test and apply fuzzy k-means (FKM) covariate stratification, where pre-determined random sample locations could be manually adjusted in the field when access was constrained.

A fuzzy k-means cluster design was generated for validation of the Meander area soil mapping. This involved stratification of the environmental covariate feature space and a random allocation of field samples from each cluster using JMP software (JMP, Version 9. SAS Institute Inc., Cary, NC, 1989–2012). Covariates included continuous-valued gamma radiometrics, the DEM and several terrain derivatives (Table 1).

In total, an additional sixty locations (30%) were sampled for validation, undertaken as a separate, independent sample set to the calibration sampling to provide an independent or external validation process (Brus, 2010). Six locations were sampled from each of ten clusters, totalling sixty sample sites. The FKM clusters (covariate strata) were generated using the software '*Fuzme*' (Minasny and McBratney, 2002) with an initial fuzzy exponent parameter of 1.30. The Mahalanobis distance option was applied due to unequal variances and associated correlations between each of the covariate data sources (Vrindts et al., 2003). The fuzzy exponent was trialled for values between 1.30 and 1.10 until a good distribution of membership probabilities was obtained, that is, at an exponent value that resulted in a clear probability of a pixel belonging to a single cluster. Optimal cluster numbers of between 5 and 20 were also trialled. For each clustering configuration, FKM cluster output files were generated which assigned a value of membership for each pixel belonging to each cluster. The highest membership value for any individual cluster resulted in that cluster being assigned to that pixel, that is, only one assigned cluster using the 'maximum rule'. Theoretically, since an individual cluster with the highest likelihood of membership was assigned to each pixel, 'hard' clustering (i.e. hard k-means) could have also been used; however this needs to be tested and compared for covariate feature-space representativeness in future sampling exercises, to determine whether spatial membership boundaries would be altered, and any consequential implications.

As the cluster map was provided to field officers to assist in relocation of problematic site locations, it was necessary to generate practical field maps; ten clusters were determined as a good visual stratification with fewer clusters over-simplifying the landscape, and greater than 10 becoming too visually 'noisy'. There was little need to more rigorously determine the optimum cluster number (by determining the partition coefficient and classification entropy) as the clusters were used for a sampling design stratification of covariates, and not modelling purposes. However, the FPI (Fuzzy Performance Index) and MPE (Modified Partition Entropy) values obtained from the *Fuzme* outputs (which should both be close to zero for a good Fuzzy k-mean design (Odeh et al., 1992)) were also considered to ensure that both values were both relatively small in the final clustering design, when compared against other fuzzy exponent/cluster number calculations.

3.5. Midlands sample design (training and validation)

Fuzzy k-means stratified sampling was also used to locate 270 sites for the 27,000 ha Midlands area. Weakly correlated covariates used in a sampling design can lead to suboptimal representation (Brus et al., 2006); therefore the most important covariates were determined from preliminary analysis of the Meander data for the targeted soil properties using a step-wise linear regression (JMP software) and used in the FKM clustering. Site coordinates were determined such that 27 randomly selected locations were sampled from each of ten derived cluster types. A further 8 locations were sampled from each cluster type to provide an additional 80 sites for validation. Validation sampling was undertaken concurrently with calibration sampling to reduce time, travel resources and duplication of field effort.

3.6. Field sampling locations selected with fuzzy k-means

The field map provided covariate clusters (strata) which allowed physically impractical sites to be re-located within clusters to more accessible locations, while still maintaining the same number of samples from each of the cluster types (Fig. 2). Locations perceived to be inaccessible (e.g. vegetation, stock-yards, and dams) (and those that could not be identified by the available spatial data layers for preliminary masking from the covariate layers) were initially identified and shifted within required clusters using a desktop GIS process, to generate a 'relaxed' sampling approach, as opposed to the 'strict' pre-determined location approach. Site locations were established in the field using a regular GPS (with approximately 2 to 5 m precision) linked to a field laptop with GIS and appropriate spatial layers for guidance. A 30 m radial 'tolerance' was allowed for each location, where field staff could sample if sampling at that exact location was not possible. This corresponded to the modelling resolution, as it was not expected to significantly change covariate distribution values, as adjacent pixels should have very similar covariate values. The GIS and GPS guidance was used to visually provide navigation to the pre-determined sampling locations and a visual means to remain within the designated cluster for inaccessible sites. When not a viable location the field officers sampled as closely as possible to the original coordinates, ensuring that they remained within the designated cluster of the original point. For completely inaccessible cluster areas, sites were manually relocated to an alternative spatial cluster of the same type but in another geographical location, to retain stratification. New coordinates and access constraint were recorded.

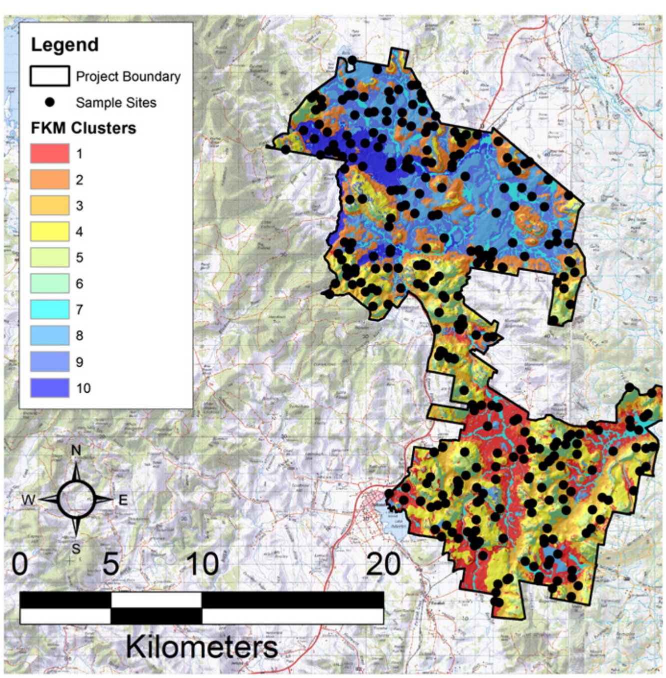

Fig. 2. Midlands FKM clusters and sample locations.

3.7. Field sampling and soil analysis

Samples were taken using a 50 mm diameter percussion soil corer to a depth of 1.5 m and sub-sampled by horizon. Cores and surrounding landscape position were described according to Australian Soil and Land Survey guidelines (National Committee on Soil and Terrain, 2009). Spectral-scanning (MIR) of all training and validation samples was undertaken by CSIRO Land & Water and DPIPWE to predict required soil properties. Fifteen percent of scanned samples were selected and analysed for chemical properties at CSBP Laboratories in Western Australia to provide calibration data for soil property predictions. Full analyses of each sample included pH, EC, exchangeable cations, N, P, K, organic carbon, and particle size distribution.

3.8. Statistical analyses

Sample site covariate distributions for the FKM clusters were compared to the overall covariate distribution using a range of metrics (i.e. the sample slope density, mean, inter-quartile range, median and standard deviation, along with the frequency distribution shape, skewness and kurtosis) to determine whether the variance from the full covariate population was acceptable and comparable to the cLHS approach, as per Brungard and Boettinger (2010). For the Midlands area, a cLHS sample design was generated to compare the sampled covariate distribution against the sampled covariate distribution of the FKM sampling design. These distributions were further compared to the overall covariate distributions for the entire area. The initial FKM design and the overall 'modified' FKM design (the adjusted sample locations due to access constraints for both combined training and validation sites) were also compared to determine whether the actual sample locations remained representative of the covariate feature space. The training and validation sites were tested separately, along with a completely random non-stratified design to determine whether the stratification effort was warranted. Similarly, the Meander area covariate distribution was compared for both the original cLHS design and those based on the actual visited locations to test for any deterioration of the hypercube once the 30 m tolerance or contingency sites had been used.

A potential problem with this approach was that the area of each cluster type was unequal, but an equal number of sites were randomly allocated across each of the cluster types, meaning that some parts of both the covariate feature space and the land surface space would be sampled more intensively than other areas. To test the consequences of this, a validation sample design, based on a random-stratified sample of the FKM covariate stratum was generated and tested using a sample density proportional to cluster area. The spatial area of each cluster type was calculated using GIS and converted to a percentage of the total area; this percentage was then applied to the total sampling site total to determine how many sites were to be located randomly in each cluster type. This would ultimately give an indication whether a more spatially proportional sampling density could be achieved (with respect to cluster area), without compromising the representation of the covariate feature space.

4. Results and observations

4.1. Field constraints

Sampling at the designated coordinates was impractical at many locations in the Meander and the Midlands study areas due to:

1. Physical or consensual access — either due to not being possible to navigate to the exact location due to physical barriers such as tracks, gates, and terrain, within a reasonable time-frame; or actually denied access consent by the land owner.
2. Cropping — the site was currently under a high-valued crop requiring site relocation to avoid damage and potential yield losses.
3. Disturbance — the site was physically disturbed or contaminated by infrastructure construction and earthworks, fertilizer stock-piles etc., therefore not a true indication of surrounding pixels, or identifiable in spatial land-use layers.
4. Infrastructure — impediments to core sampling such as tracks, fences, underground cables, and drainage pipes, which were beyond the resolution of a desktop identification process.
5. Livestock — access not possible or denied to avoid disturbance of livestock at critical times, such as lambing.
6. Stone — too physically stony to sample without damage to equipment.
7. Terrain — too physically rough or steep to allow safe drilling rig access, and compliance with departmental safety policy.
8. Biosecurity and conservation — a high-risk biosecurity or biodiversity area with the potential for spread of weeds, pathogens, or damage to vulnerable or threatened species.

Table 2 summarises the access constraints experienced during fieldwork; the most common being limited access (either physical or denied entry) and inaccessible terrain.

Table 2
Access limitations — reasons for moving sample locations.

Access limitation	Count	Limitation %
Access	56	21
Biosecurity	0	0
Crop	21	8
Disturbed	5	2
Infrastructure	29	11
Livestock	1	0
Stone	33	13
Terrain	48	18
Vegetation	21	8
Wet	48	18
Total not accessed	262	
Total sites	589	
% sites not accessed	44	

4.2. Sample design comparisons (Midlands)

Supplementary data Table 1 shows a comparison in the Midlands area between each soil sampling design (combined training and validation) against the full covariate population and a percentage difference against mean, median, inter-quartile range (IQR), and other statistical outputs. The 'field modified' or 'relaxed' FKM sample design showed the distribution statistics for the actual sample locations after shifting any proposed site locations due to access limitations. The percentage difference between each design and the full covariate distribution for all metrics was generally lowest for the field-modified FKM design, followed by the cLHS. For example, the percentage difference of median values for FKM-final was 1.28, compared to 4.06, 8.23 and 0.85 for the FKM-proposed, random sample and cLHS values, respectively. Mean values generally agreed with this; however, medians are used here for example comparisons as they are less influenced by outlier values. There was a similar pattern for the median values for MrRTF, slope, topographic wetness, total gamma dose, thorium and potassium. The random sample was generally less representative of the tested covariates, as demonstrated by the percentage differences to the other sampling approaches for all metrics. Elevation (DEM), for example, shows a percentage difference of −18.85% for median MrRTF, which was as low as −2.36, 4.67, and −7.06 for FKM-final, FKM-proposed, and cLHS respectively.

Fig. 3 shows the generalised combined training and validation frequency distribution curves for each sample design against the full covariate population, along with quantile box plots. The density curves basically illustrate that the cLHS most closely followed the full covariate distribution; however all sample designs showed a reasonable representation of covariate distribution, following the same general distribution shape.

Supplementary data Table 2 shows the comparison of the separate training and validation sampling statistics against the entire covariate population. Generally, the training FKM design followed a closer representation of the full covariate distribution to the validation sampling with the exception of plan curvature, aspect and valley-depth. Importantly, subjectively moving sites within clusters did not overly reduce the representation of the covariate distribution, for example, median values between the proposed and final training designs were 311.49 and 310.90 for elevation respectively, and an unchanged 0.82 for both proposed and final mid-slope position median values.

Fig. 4 and Supplementary data Table 3 show a frequency distribution comparison of the Meander East cLHS design against the full covariate distribution, and any deterioration of the sample integrity due to using contingency sites when access was constrained. The distributions illustrate that there was negligible deterioration of the cLHS design once contingency sites had been used or sites moved within the allowable 30 m tolerance, i.e., the design was representative of the covariate feature space, following the same general shape. Similarly, insignificant differences between the mean, median, IQR, skewness and kurtosis values suggest that there was negligible deterioration in sample representativeness between the proposed and final cLHS locations, for example, the DEM median values were 165.42 and 163.78 for the proposed and final locations respectively, and no change in median value for radioactive potassium (k percent).

Supplementary data Table 4 shows the frequency distribution statistics where sample density was adjusted proportionally to FKM cluster area size. These were generally close to the covariate distribution of the un-adjusted FKM sampling design, with the exception of some covariables, for example, valley depth having a median value of 60.14, compared to 55.12 and 55.47 for the unweighted sample, and covariate distribution respectively.

The supplied Google Earth '.kmz' files show the Midlands final site locations (categorised by training and validation, 'Sample_sites.kmz') and FKM clusters (strata) (categorised by cluster type, 'FKM_strata.kmz'). Background satellite RGB imagery also highlights the land use and terrain variability throughout each study area.

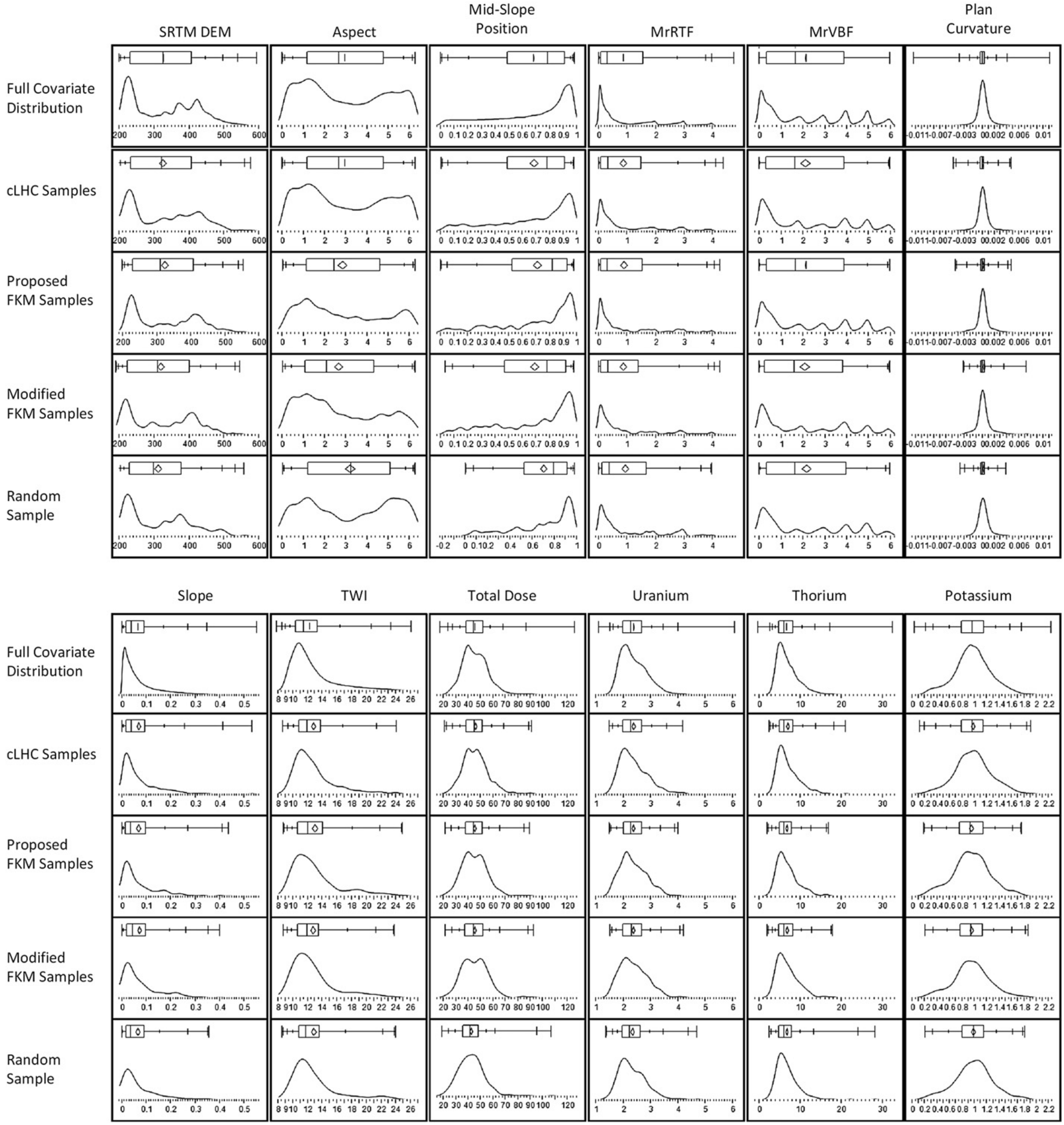

Fig. 3. Comparison of sample designs covariate distribution against the full covariate distribution.

5. Discussion

5.1. Access constraints, and the need for sampling flexibility

Within the Midlands and Meander areas, sites not accessible within a reasonable time frame and departmental occupational health and safety guidelines were encountered at a rate of 44%. This is a substantial (and surprising) proportion, which indicates that 'strict' sampling can be problematic in intensively used areas where impediments to access and physical sampling exist. Within the project areas the most common form of access limitation was due to physical constraints such as locked gates, fences, and tracks that survey staff could not negotiate within reasonable efforts in terms of time, and driving hazard (Table 2). This included land owners who declined to participate in the project.

Departmental field policy specifies that staff health and safety are mandatory. Field staff used a trailer mounted drill rig in excess of 3000 kg which also made access problematic in terms of bogging-potential, stability on steep terrain, and potential paddock damage. The most common constraints to safe towing of the drilling rig were terrain (usually highly undulating, or steep landscapes) to allow safe negotiation, or site wetness causing vehicular traction problems. Other access constraints were due to sites currently under cropping with entry

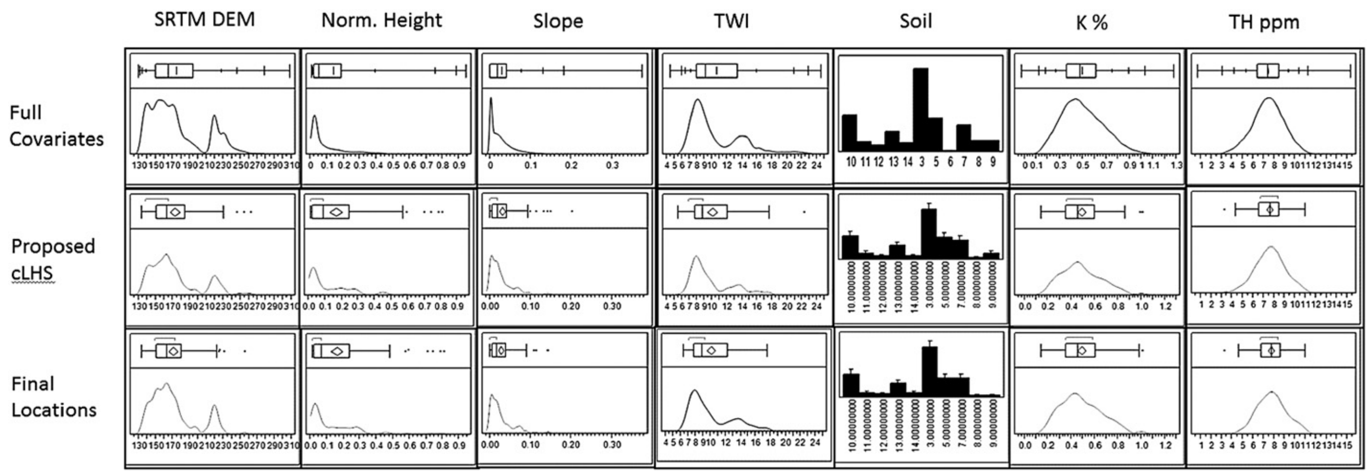

Fig. 4. Comparison of proposed Meander East cLHS covariate distribution against the final sample locations and full covariate distribution.

denied by the farmer to avoid crop damage; site contamination or disturbance, for example, stock camps deemed by the surveyor to bias sampling results due to excessive nutrient inputs; infrastructure, such as communication cables, drainage and water pipes; and vegetation too thick to navigate. Physical sampling was also limited in areas of high stone-content or outcropping. Locations that were too wet or stony to sample were given a stone % estimate and drainage class value (which were to be used as supplementary data for the DSM predictions of these soil variables). Many of the access constraints experienced were not identifiable by spatial GIS layers or were beyond the scale or resolution of the covariates used (for example, drainage lines), and could not be excluded before the sampling design was generated. Although some slope categories could have been removed from the design, these terrain issues were often encountered while traversing to sample locations, which could not be determined directly from the GIS.

A physical map of clusters provided field staff visual guidance to relocate a site as close as possible to the original coordinates. Relocating the site within the same cluster (strata) ensured that the site remained associated to that cluster, effectively maintaining covariate stratification and reducing the risk that sampling distribution would become non-representative of the full covariate feature space. The FKM cluster approach achieved positive feedback from the sampling staff, and resulted in a substantial time reduction for sampling (in terms of sites sampled per day), avoiding the need to navigate to a completely new and alternative site as required with a ‘strict’ approach (the alternative cLHS sites have the potential to be on completely different land titles requiring time and resources to find and contact new farmers, and navigate through a whole new raft of potential sampling constraints).

5.2. Sampling design comparisons

The sample size used was 0.12% of the possible sample locations from the 30 m grid (299,781 cells (covariate population), sample n = 350). From Fig. 3 it can be observed that the cLHS design best accounted for the full covariate distribution with very little deviation from most covariate population's mean, median, interquartile range (IQR), and standard deviation. It also maintained the best curve ‘fit’ in terms of skewness and kurtosis, i.e. it showed little variation to the shape, flatness and distribution of the frequency distribution curves, and therefore had the smallest deviation in the compared sampling design covariate representation.

The FKM sample design, while not as close to the overall distribution density as the cLHS, was generally a good fit, with a slight increase in difference between the median, mean, IQR, and distribution density ‘shape’ (Fig. 2 and Supplementary data Table 1). The mean, median, variance, and curve density of the ‘modified’ FKM design (the final sampling locations after being shifted) was a slightly poorer fit for some covariates, while improved for others. The completely random design (with no stratification), while still a good fit for some covariates, performed poorest overall in comparison to the cLHS and FKM approaches with the greatest variance in terms of IQR, and more-so for skewness and kurtosis (Supplementary data Table 1). This shows that stratification is beneficial for the range of covariates used in this study. For the Meander area, the proposed cLHS showed minor degradation when compared to the final site locations, shifted due to access constraints (Supplementary data Table 3, Fig. 4). The mean, median, IQR, skewness, and general non-parametric density shapes more closely followed the full covariate distribution for the proposed sample design than the final site location covariate distribution.

Ideally, the comparison of these sampling approaches should be undertaken with replicates due to the strong random component in the site selection in all methods. This was done for several replicates before sampling to test the sensitivity of the covariate distribution of the selected sites, and a visual estimate of spatial dispersion. Visual clustering of sites was evident for some replicates in comparison to others; however the covariate distribution remained relatively constant across replicates. These results are not presented here due to the large amounts of data already produced in the supplementary data tables.

A potential weakness with the FKM design is that smaller clusters have an equal number of samples to larger clusters. However larger covariate clusters should intuitively have lower covariate distribution variance per unit area, therefore requiring fewer samples per unit area. Complex terrain comprising smaller area clusters should have greater covariate distribution variability per unit area and will therefore have a higher rate of samples required per unit area. In both project areas the smaller area clusters are located in more topographically complex areas which require a greater sampling density to fully cover the covariate distribution, whereas the larger flood-plain areas have less variability per unit area (see Fig. 2). Supplementary data Table 4 shows a simulation sample design generated such that a proportional number of sites were located by the spatial cluster size. The site covariate distribution statistics had a greater difference to the full covariate distribution statistics for most covariates (with the exception of the DEM, mid-slope position, MrRTF and valley depth), when compared to the proposed (un-weighted) FKM design used for sampling the Midlands area. This suggests that using a sampling density proportional to cluster size would generally not improve sampling the covariate feature space, but could potentially improve proportional sampling of land-surface ‘space’, which would be of more importance to design-based sampling.

5.3. Comparison of training and validation designs

Project time constraints meant that validation sampling could not be delayed until initial soil property predictions had been generated. Consequently, validation sampling was undertaken simultaneously with training sampling, using the covariate FKM clusters as strata, rather than stratification of generated quantitative soil properties or class surfaces. Validation field sampling, co-incident with training sampling, greatly reduced overall time and associated costs for the Midlands area by avoiding duplication of field effort. Validation sites, calculated in number as an additional 30% of training samples and randomly selected by computer from each cluster, were compared against the full covariate distribution (Supplementary data Table 2). Although generally showing a greater variance from the full distribution than the calibration sampling, the sample locations provided reasonable covariate distribution replication. The higher variance is explained by the lower sampling density, with less opportunity to sample the entire covariate range.

Due to the documented access limitations it was not possible to sample all locations (pixels), or identify and mask inaccessible sites of the study area using desk-top imagery analysis. The validation approach used in this study is therefore not a true probability sample as not all locations have an equal chance of being selected, which will reduce the ability to provide statistical estimates of map quality, providing instead a model-based estimate of quality. In effect, this validation approach does not provide a true measure of statistical accuracy, but quantifies the prediction error at each validation point (Laslett 1997) which can be averaged to give an overall model prediction error.

The ability to undertake a probability sampling design is questionable in an operational sense as it is unlikely in any 'real-world' situation that every location will be equally available to be sampled. It is impossible to compute these inclusion probabilities, as inclusion likelihood would not be known until all locations are attempted to be sampled. Subsequently, it was opted for a pragmatic solution to this perceived dilemma by sampling more sites than intentioned for DSM model calibration; using additional sampling resources for model validation. The validation sample is considered independent of the calibration samples in the sense that they were not included in the model fitting. The operational advantage is realised by the fact that both calibration and validation samples can be retrieved together in the same sampling campaign.

5.4. Potential alternative future sampling approaches

The navigation time to locations within the 'strict' design was found to be the main impediment to progress in intensively used areas such as the Midlands and Meander regions of Tasmania. The described 'relaxed' approach increased the number of sites able to be sampled per day, however, the sampling-rate increase is difficult to quantify due to the different drilling conditions between Meander and Tunbridge. Project time-constraints would not allow full testing of variations to the methodology before sampling was undertaken; for future sampling, further investigations into cluster-size sample density and whether optimisation of spatial distribution is warranted, and how this might affect the representation of the sampled covariate distribution, and any advantages to final DSM predictions.

Importantly, this case study and the documented operational problems encountered ultimately highlights the need for a more adaptive approach such as the flexible LHS methodology proposed by Clifford et al. (2014). The approach uses pre-determined-coordinate LHS sampling in a remote Queensland setting, with alternative nearby sites suggested when the first choice site sampling is not possible, while maintaining the hypercube integrity covariate distribution, distance to formed tracks and spatial spread. The 40 ha alternative area criteria used in the flexible LHS simulation would need reducing for Tasmanian conditions due to significantly smaller property holdings, and more intensive infrastructure than the Queensland example. These smaller holdings exacerbate sampling delays in situations where alternative sampling locations are made available outside of current property boundaries; usually a completely different land-holder will need contacting for access permission, taking time to find and make contact during daylight hours. Obtaining contact details for all landholders in a sampling area is impractical due to Tasmanian privacy legislation, and the format of contact availability by telephone companies or land title authorities (with many property owners listed as a registered business name only). The decision to use this approach required consideration of whether it was more beneficial to have an optimal sampling design of fewer sites (due to the slower sampling rates), or a much larger number of sites in less optimal locations; it was decided to use the latter in the interests of obtaining more training data for model formation. The FKM approach used for this DSM sampling example was a pragmatic model-based concession that was able to improve soil sampling rates while still representing the covariate feature space. However for future sampling campaigns, the Clifford et al. flexible LHS sampling will be tested alongside this approach as a measure of covariate distribution, sample spread and field practicality to determine the best possible compromise between statically valid model-based, but operationally feasible soil sampling for DSM.

6. Conclusions

Analyses of the project sampling component demonstrated that the cLHS design for the Midlands area showed the best overall sample based on the distribution of the available covariates, but led to operational sampling problems due to the access constraints. However, the 'relaxed' FKM clustered approach provided an acceptable representation of the available covariates, while the clusters (used as a field map and stratification of the sampling area) provided a guide to where non-accessible sample locations could be moved such that the sample number from each covariate stratum was maintained. The approach was able to improve sampling progress rates when compared to the strict cLHS approach, although not quantifiable in this project due to differing sampling conditions.

An operational digital soil mapping (DSM) sampling process needs to be a compromise between operational practicality, and statistically-sound sampling that should give the best opportunity for predictions based on the full covariate range. Although some subjectivity-bias may be introduced, the 'relaxed' FKM approach can allow a pragmatic and time-saving operational sampling strategy, which was shown to still follow a good representation of the available environmental variables. The relaxed FKM strategy was therefore considered a reasonable, 'real-world' sampling compromise suitable for time-constrained DSM operations in Tasmania, which could be applicable where a model-based validation of map quality is acceptable. However, future operational and 'real-world' testing of more statistically robust methods is still required.

Acknowledgments

The authors acknowledge the following: the Australian Research Council Linkage Project LP110200731 for supporting the project; Chris Grose and Rob Moreton (DPIPWE) for soil profile and access constraint descriptions, feedback, suggestions and advice; Regan Parkinson and Jo Wang (DPIPWE) for field assistance and sample processing; Rhys Stickler, Peter Voller and Ashley Bastock (DPIPWE) for project management; and the farmers of the project areas (who allowed access).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.geodrs.2014.11.002. These data include Google maps of the most important areas described in this article.

References

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the Fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–203.

Brungard, C.W., Boettinger, J.L., 2010. Conditioned Latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping. Progress in Soil Science. Springer Netherlands, pp. 67–75.

Brus, D.J., 2010. Design-based and model-based sampling strategies for soil monitoring. Proceedings; 19th World Congress of Soil Science; Solutions for a Changing World. 1st to 6th August 2010, Brisbane, Australia.

Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2006. Chapter 14 designing spatial coverage samples using the k-means clustering algorithm. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Developments in Soil Science. Elsevier, pp. 183–192.

Burrough, P.A., van Gaans, P.F.M., MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets Syst. 113 (1), 37–52.

Carré, F., McBratney, A.B., Mayr, T., Montanarella, L., 2007. Digital soil assessments: beyond DSM. Geoderma 142 (1–2), 69–79.

Chapron, N., 2011. Classification of soil and vegetation by Fuzzy k-means. Classification and Particle Swarm Optimization. Proceedings, International Conference on Swarm Intelligence, Cergy, France, p. 2011.

Clifford, D., Payne, J.E., Pringle, M.J., Searle, R., Butler, N., 2014. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput. Geosci. 67, 62–68.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12), 1347.

Gallant, J., Dowling, T.I., Read, A., Wilson, N., Tickle, P., 2011. 1 Second SRTM Derived Digital Elevation Models User Guide. Geoscience Australia, Canberra, p. 106.

Hartigan, J.A., 1985. Statistical theory in clustering. J. Classif. 2 (1), 63–76.

IUSS Working Group, W.R.B., 2007. World reference base for soil resources 2006. First update 2007. World Soil Resources Report No. 103. FAO, Rome.

Kidd, D.B., 2003. Land Degradation and Salinity Risk Investigation in the Tunbridge District, Tasmanian Midlands. Department of Primary Industries, Parks, Water and the Environment, Tasmania, Australia.

Kidd, D.B., Webb, M.A., Grose, C.J., Moreton, R.M., Malone, B.P., McBratney, A.B., Misnasny, B., Viscarra-Rossel, R.A., Cotching, W.E., Sparrow, L.A., Smith, R., 2012. Digital soil assessment: guiding irrigation expansion in Tasmania, Australia. In: Minasny, B., McBratney, A.B., et al. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012. Taylor & Francis, Sydney, Australia.

Laslett, G.M., 1997. In Discussion of: D.J. Brus and J.J. de Gruijter, Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. Geoderma 80, 45–49.

Leamy, M.L., 1961. Reconnaissance soil map of Tasmania. Sheet 61 — Interlaken. CSIRO PUBLISHING, Australia.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. California, USA, pp. 281–297.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89 (1–2), 67–94.

Minasny, B., McBratney, A.B., 2002. FuzME Version 3.0., Australian Centre for Precision Agriculture. The University of Sydney, Sydney, Australia.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.

National Committee on Soil and Terrain, 2009. Australian soil and land survey field handbook, Australian Soil and Land Survey Handbooks Series 1, 3rd ed. CSIRO PUBLISHING, Melbourne, Australia.

Odeh, I.O.A., Chittleborough, D.J., McBratney, A.B., 1992. Soil pattern recognition with Fuzzy-c-means: application to classification and soil–landform interrelationships. Soil Sci. Soc. Am. J. 56 (2), 505–516.

Roudier, P., Beaudette, D., Hewitt, A., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Proceedings, 5th Global Workshop on Digital Soil Mapping. Digital Soil Assessments and Beyond. 10th–13th April, 2012. Sydney, Australia.

Spanswick, S., Kidd, D., 2001. Oatlands soil report — a revised addition. Reconnaissance Soil Map Series of Tasmania. Department of Primary Industries, Parks, Water and the Environment, Tasmania, Australia.

Spanswick, S., Zund, P., 1999. Quamby Soil Report — Reconnaissance Soil Map Series of Tasmania, A Revised ed. Department of Primary Industries, Parks, Water and the Environment, Tasmania, Australia.

Thomas, M., Odgers, N.P., Ringrose-Voase, A., Grealish, G.J., Glover, M., Dowling, T.I., 2012. Soil survey design for management-scale digital soil mapping in a mountainous southern Philippine catchment. In: Minasny, B., McBratney, A.B., et al. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012. Taylor & Francis, Sydney, Australia.

Vašát, R., Heuvelink, G.B.M., Borůvka, L., 2010. Sampling design optimization for multivariate soil mapping. Geoderma 155 (3–4), 147–153.

Vrindts, E., Reyniers, M., Darius, P., Frankinet, M., Hanquet, B., Destain, M.F., Baerdemaeker, J.D., 2003. Analysis of spatial soil, crop and yield data in a winter wheat field. Proceedings ASAE Annual International Meeting.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Comput. Geosci. 36 (10), 1261–1267.

Zadeh, L.A., 1968. Fuzzy algorithms. Inf. Control. 12 (2), 94–102.