Geoderma 265 (2016) 152–163

Contents lists available at ScienceDirect

Geoderma

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

Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference

R.A. Viscarra Rossel [a,*], D.J. Brus [b], C. Lobsey [a], Z. Shi [c], G. McLachlan [d]

[a] CSIRO Land and Water, Bruce E. Butler Laboratory, PO Box 1666, Canberra, ACT 2601, Australia
[b] Alterra, Wageningen University and Research Centre, Environmental Science Group, Wageningen, The Netherlands
[c] Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China
[d] Agriculture Flagship, CSIRO, Bruce E. Butler Laboratory, PO Box 1666, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 8 August 2015
Received in revised form 19 October 2015
Accepted 10 November 2015
Available online 3 December 2015

Keywords:
Proximal soil sensing
Soil organic carbon stocks
Visible–near
Infrared spectroscopy
Design-based sampling
Regression estimator
Model-based inference

ABSTRACT

For baselining and to assess changes in soil organic carbon (C) we need efficient soil sampling designs and methods for measuring C stocks. Conventional analytical methods are time-consuming, expensive and impractical, particularly for measuring at depth. Here we demonstrate the use of proximal soil sensors for estimating the total soil organic C stocks and their accuracies in the 0–10 cm, 0–30 cm and 0–100 cm layers, and for mapping the stocks in each of the three depth layers across 2837 ha of grazing land. Sampling locations were selected by probability sampling, which allowed design-based, model-assisted and model-based estimation of the total organic C stock in the study area. We show that spectroscopic and gamma attenuation sensors can produce accurate measures of soil organic C and bulk density at the sampling locations, in this case every 5 cm to a depth of 1 m. Interpolated data from a mobile multisensor platform were used as covariates in Cubist to map soil organic C. The Cubist map was subsequently used as a covariate in the model-assisted and model-based estimation of the total organic C stock. The design-based, model-assisted and model-based estimates of the total organic C stocks in the study area were similar. However, the variances of the model-assisted and model-based estimates were smaller compared to those of the design-based method. The model-based method produced the smallest variances for all three depth layers. Maps helped to assess variability in the C stock of the study area. The contribution of the spectroscopic model prediction error to our uncertainty about the total soil organic C stocks was relatively small. We found that in soil under unimproved pastures, remnant vegetation and forests there is good rationale for measuring soil organic C beyond the commonly recommended depth of 0–30 cm.

1. Introduction

Soil organic carbon (C) helps to maintain soil health and productivity. It provides a primary source of nutrients for plants, helps to aggregate particles and develop soil structure, increases water storage capacity and availability for plants, protects soil from eroding and provides a habitat for soil biota. Capturing and retaining additional C in soil can improve the quality and productivity of the soil to sustain food production and simultaneously also mitigate the emissions of greenhouse gases (GHG).

Thoughtful land use and management practices, such as management-intensive grazing, can help store more soil organic C and offer good potential to improve soil quality, enable profitable food production and reduce net GHG emissions (Machmuller et al., 2015). For baselining and to assess the success of such practices, however, we need to accurately quantify the variability of soil organic C stock in both space and time. Importantly, we should aim to characterize its short range spatial variation, which can be significant, and to monitor over time intervals that enable detection of relatively small changes in C stocks.

Soil sampling protocols and conventional laboratory analyses can be used to directly measure organic C stocks. The protocols typically involve designing a sampling strategy, sampling the 0–30 cm soil layer and measuring the organic C concentration, bulk density and gravel content to derive the organic C stock of the soil in this layer. The methods are time-consuming, expensive, involve much sample handling and preparation and use complex procedures, which can be prone to analytical inaccuracies. The complexity and expense of the conventional approach are greater when there is a need to monitor the organic C stock of deeper soil layers or entire profiles. There is evidence that plants and cultivars with deeper and thicker root systems can input stable forms of organic matter deeper in the soil profile (Jobbágy and Jackson, 2000; Lorenz and Lal, 2005).

Conventional methods for measuring changes in the organic C stocks of soil are therefore impractical. If we are to increase our ability to characterize and monitor changes in soil organic C stocks, we need to

* Corresponding author.
E-mail address: raphael.viscarra-rossel@csiro.au (R.A. Viscarra Rossel).

http://dx.doi.org/10.1016/j.geoderma.2015.11.016

develop rapid, practical, accurate and cheaper methods to measure it (Izaurralde et al., 2013). Proximal soil sensing provides a range of tools that can be used to develop a multi-sensor system to efficiently measure the organic C stock of soil profiles (Viscarra Rossel et al., 2011). For example, electromagnetic induction sensors, gamma radiometers and precise global navigation systems can produce multivariate secondary information to help design sampling strategies and to map soil C (e.g. Simbahan and Dobermann, 2006; Miklos et al., 2010). Soil visible–near infrared (vis–NIR) spectroscopy can be used to measure soil organic C in the laboratory and in situ in the field (Stenberg et al., 2010).

Before we can start measuring with sensors however, we need to know where to sample. Locations can be selected by probability sampling (random sampling with known inclusion probabilities) or by non-probability sampling, giving rise to two widely used philosophies: the design- and the model-based approaches (de Gruijter and ter Braak, 1990; Brus and de Gruijter, 1993; Papritz and Webster, 1995; de Gruijter et al., 2006). In the design-based approach, the source of randomness of an observation is the random selection of the sampling sites. In the model-based approach, randomness originates from a random term in the model of the spatial variation, which is added to the model because our knowledge of the spatial variation is imperfect. Thus, probability sampling is a requirement for the design-based approach, whereas it is not for the model-based.

Choosing the most suitable approach depends, amongst other, on the motivation (Brus and de Gruijter, 1997). For example, the design-based approach might be more suitable if the aim is to obtain estimates of the '*global*' mean or total stock and their accuracies for an area, whose quality is not dependent on the correctness of modelling assumptions. The model-based approach might be preferable if we want to produce a '*local*' map of the soil organic C stock in the area. However, deciding which approach to use is often more complicated because the design-based approach can also be used for estimation of local means, and the model-based approach can be used for global estimation. Further discussion on the merits and disadvantages of each method can be found in de Gruijter and ter Braak (1990), Papritz and Webster (1995), Brus and De Gruijter (1997) and de Gruijter et al. (2006).

The possibility of using a regression model to assist with design-based inference, in a model-assisted approach, was discussed by Särndal et al. (1992) and Brus (2000). The approach uses auxiliary information, captured in a regression model, to improve the accuracy of design-based estimates of means and totals. There are fundamental differences between a model-based and a model-assisted approach. Significantly, the variance of a model-assisted estimate of the mean is a sampling variance, not a model-variance. Unlike the estimates of the model-based variance, the model-assisted estimates of the variance do not rely on the correctness of the model's assumptions. That is, if the assumptions underlying the regression model are violated, the model-assisted approach can still produce an unbiased estimate of the sampling variance (Brus, 2000).

Our aims here are to: (i) demonstrate the use of proximal soil sensors to measure the soil organic C stock of grazing land to a depth of 1 m, (ii) to compare the use of design-based, model-assisted and model-based methods to derive baseline estimates of the mean and total soil organic C stocks and their accuracies in the 0–10 cm, 0–30 cm and 0–100 cm layers, and (iii) to derive maps of soil organic C stocks and their uncertainties for each of the three depth layers.

2. Methods

2.1. Study site

The study area is 2837 ha and is located in the Upper Hunter Valley region, New South Wales, Australia, south of Wollar. It is approximately 300 km northwest of Sydney and 50 km northeast of Mudgee, near the Goulburn River National Park. The region has a temperate climate with an average annual rainfall of approximately 600 mm. Geology consists of shale, sandstone, mudstone conglomerates and coal. Landforms at the site consist of gently sloping colluvium and undulating foothills adjacent to north-flowing tributary creeks that are part of the Goulburn River Catchment. There are steep timbered ridges that surround on the south, west and east. The study area is used mostly for cattle grazing for beef production on rain-fed unimproved pastures, with remnant vegetation and surrounding forests on higher elevations. The soil there belongs to mostly the Dermosol and Kurosol orders in the Australian soil classification (Isbell, 2002), approximately equivalent to Planosols, Phaeozems and Acrisols in the World Reference Base system (IUSS Working Group WRB, 2006).

2.2. Proximal soil multi-sensor survey and data preprocessing

A mobile multi-sensor platform (MMSP) was used to survey the study area. The proximal sensors on the platform were an electromagnetic induction sensor, the EM-38 Mk2 (Geonics, Canada), a gamma radiometer with a 4.2 L NaI crystal detector (Radiation solutions, Canada), and a real-time kinetic global navigation system (RTK-GNS) (Trimble, USA).

The MMSP was driven between 10 to 20 km h^{-1} and the sensor data were recorded at a frequency of 1 Hz on parallel line transects with line spacing between 20 and 60 m. Both the speed and the line spacing depended on the navigability of the terrain. A map of the MMSP tracks is show in Fig. 1a. Using each sensor, the data recorded were: electrical conductivity and magnetic susceptibility recorded from the 0–0.5 m and 0–1 m depths, ($EC_{0.5}$, EC_1, $MS_{0.5}$ and MS_1), respectively; gamma radiometrics total dose, potassium (K), uranium (U) and thorium (Th), recorded from around the top 0.5 m of soil (Cook et al., 1996) and elevation with the RTK-GNS.

We checked the histograms of each sensor's data and checked for outliers using the Mahalanobis distance on their correlations. These and other spurious measurements were removed before proceeding with our analysis.

The gamma U and Th bands possessed significant random noise because of the short integration time that we used for the mobile measurements (i.e. 1 Hz). To improve the signal-to-noise ratio of these data, we aggregated each channel spatially using a moving average and using points within a 20 m radius. Thus, the gamma counts were integrated in space rather than in time (Viscarra Rossel et al., 2007).

We derived variograms for each of the sensor data and interpolated them onto a 5 m grid with ordinary block kriging (Webster and Oliver, 2007). The digital elevation map (DEM), produced by kriging, was used to derive terrain attributes that were thought to help describe the variation in soil organic C across the landscape. To derive the terrain attributes, we used the Geographic Resources Analysis Support System (GRASS) geographic information system (GIS) (GRASS Development Team, 2012). The attributes were slope, aspect, tangential, plan and profile curvatures, flow accumulation and the topographic convergence index (TCI) (GRASS Development Team, 2012).

The maps of the sensor data and the terrain attributes are shown in Fig. 1b–o. Elevation in the study area ranges from 360 m in the north to 485 m on ridges to the south (Fig. 1b). The soil has small electrical conductivity across the site but particularly on the ridges to the west (Fig. 1j). The gamma K counts were generally greater at higher elevations along the ridges (Fig. 1l) and suggest the occurrence of soil derived from parent material that contains K-bearing silicates.

2.3. Soil sampling

We selected sampling locations by probability sampling using a stratified simple random design (de Gruijter et al., 2006). We used the interpolated soil sensor data as the variates in the stratification, but we first reduced dimensionality and eliminated multicollinearity between them, using a principal component analysis (PCA). The PCA was

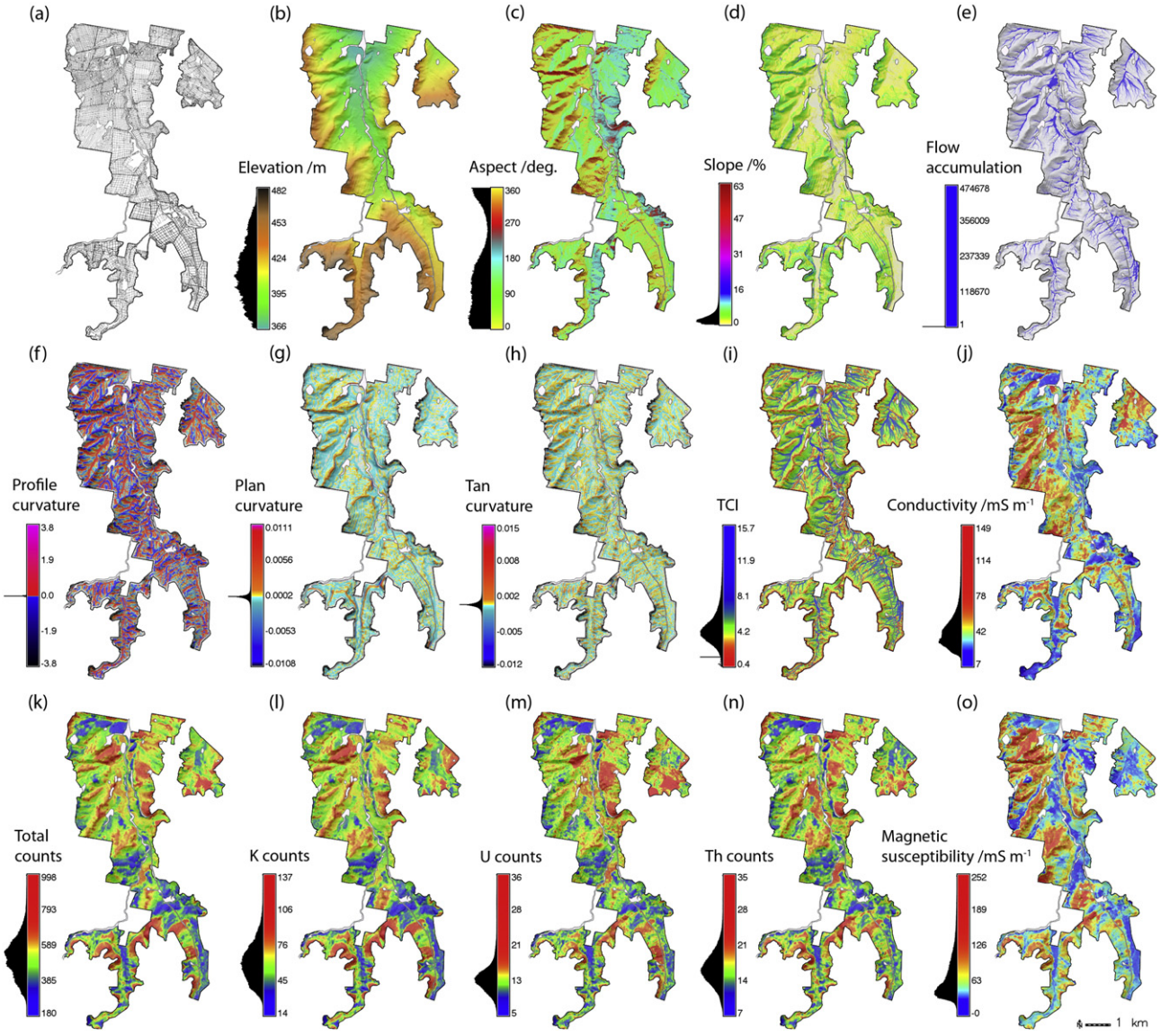

Fig. 1. Maps derived from the mobile multi-sensor platform (MMSP). (a) MMSP tracks over the study area, and (b–o) sensor covariates derived by interpolation of the data recorded by the MMSP.

performed using the iterative NIPALS algorithm (Martens and Næs, 1989). The PCA algorithm produces a set of scores that condense the information content in the samples and a set of eigenvectors which show the variables that load heavily on the particular component. The first principal component accounts for the largest variance, while subsequent components account for decreasingly smaller portions (Table 1).

To derive the stratification across the study area on the 5 m grid, we implemented a *k*-means clustering (MacQueen, 1967) using the PCA scores maps of the first four components (Fig. 2a–d), which accounted for 83% of the variance in the covariates (Table 1).

We could afford to collect soil core samples at 150 sites. After comparing a number of candidate designs that used different number of strata and sampling points per stratum, we decided to use 75 strata with two sampling units within each stratum (Fig. 2e, f), because it produced the best compromise between good sample coverage in feature and geographic spaces. To quantify their performance we used the O1 criterion (Minasny and McBratney, 2006) to assess coverage in feature space, and the mean squared shortest distance (MSSD) to assess spatial coverage (Walvoort et al., 2010).

We used a Geoprobe DT22 soil sampling system to sample the soil cores from the top 1 m of soil, or shallower if bedrock was present. The sampled cores were 50 mm in diameter and were collected in PVC tubes for measurement and storage.

2.4. Proximal multi-sensor measurements on the soil cores

The soil cores were measured at field condition using a vis–NIR spectrometer to infer estimates of soil organic C and soil water, and an active gamma attenuation sensor to measure soil bulk density. We measured simultaneously with each of these sensors at 5 cm intervals down the length of the cores. The first measurement was made at 3 cm from the

Table 1
Principal component analysis of the sensor covariates (Fig. 1b–o).

Component	Eigenvalue	Eigenvalue (%) total	Eigenvalue cumulative (%)
1	1.96	38.4	38.4
2	1.48	21.9	60.3
3	1.13	12.8	73.1
4	0.98	9.5	82.6
5	0.88	7.7	90.3

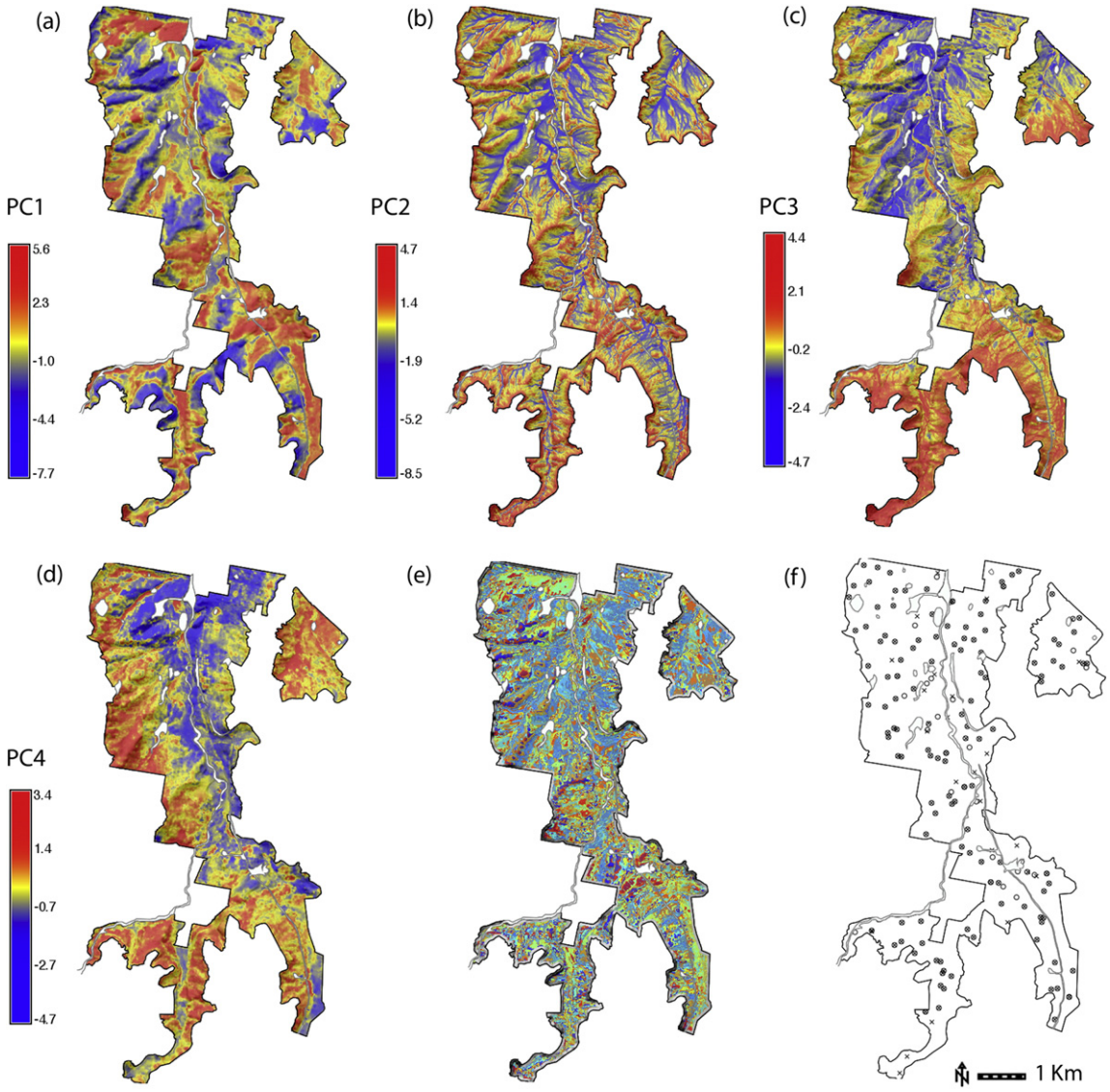

Fig. 2-. (a–d) Maps of the first four principal components of the sensor covariates (shown in Fig. 1b–o), (e) the 75 *k*-means strata and (f) the soil sampling locations.

top, just below the soil surface, thus, in a 1 m core we had a total of 20 measurements from each sensor.

2.4.1. Spectroscopic measurements

The vis–NIR spectrometer was a Labspec (PANalytical, Boulder, Colorado, USA), with a spectral range of 350–2500 nm and spectral resolution of 3 nm at 700 nm and 10 nm at 1400 nm and 2100 nm. We used a modified high-intensity contact probe (also from PANalytical) with a halogen bulb (2901 ± 10 K) for illumination. The contact probe measures a spot of diameter 10 mm, and it is designed to minimize errors associated with stray light. The sensor was calibrated with a Spectralon white reference panel at the start of every core. For each soil measurement 30 spectra were averaged to minimize noise and so to maximize the signal-to-noise ratio. At each measurement depth on the soil core, spectra were recorded with a sampling resolution of 1 nm so that each spectrum comprised reflectances at 2151 wavelengths. To standardize the soil core spectra, we first subtracted the reflectance of the first wavelength (with the minimum reflectance value) to correct for potential baseline shifts between the measurements. Because the spectra are highly collinear, we retained only every tenth wavelength from 350 to 2500 nm, inclusive. This left 216 wavelengths for the analysis. The average spectrum of each depth over all cores is shown in Fig. 3a.

2.4.2. Gamma attenuation measurements

The active gamma sensor (LB444, Berthold technologies, Germany) measured the density of the soil cores. The approach is described in Lobsey and Viscarra Rossel (in press). Briefly, when a gamma-ray beam passes through the soil cores, photons are transmitted following Beer–Lambert's law:

$$I = I_0 \exp[-x(\mu_s \rho_s + \mu_w \theta)] \quad (1)$$

where I and I_0 are the emerging and the incident photon beams, x is the thickness of the soil sample, μ_s and μ_w are the soil and water mass attenuation coefficients ($cm^2\ g^{-1}$), ρ_s is the soil bulk density ($g\ cm^{-3}$) and θ is the soil water content ($cm^3\ cm^{-3}$). If the parameters I_0, μ_s and μ_w are known and we can independently measure the soil water content (θ), then we can determine the bulk density of the soil. In our case, as in Lobsey and Viscarra Rossel (in press), we estimated the water content of the soil core samples using the vis–NIR spectra.

2.4.3. Laboratory analysis and vis–NIR estimates of soil organic C

For the spectroscopic modelling, once the sensor measurements were completed, we selected 50 soil samples from the cores at specific

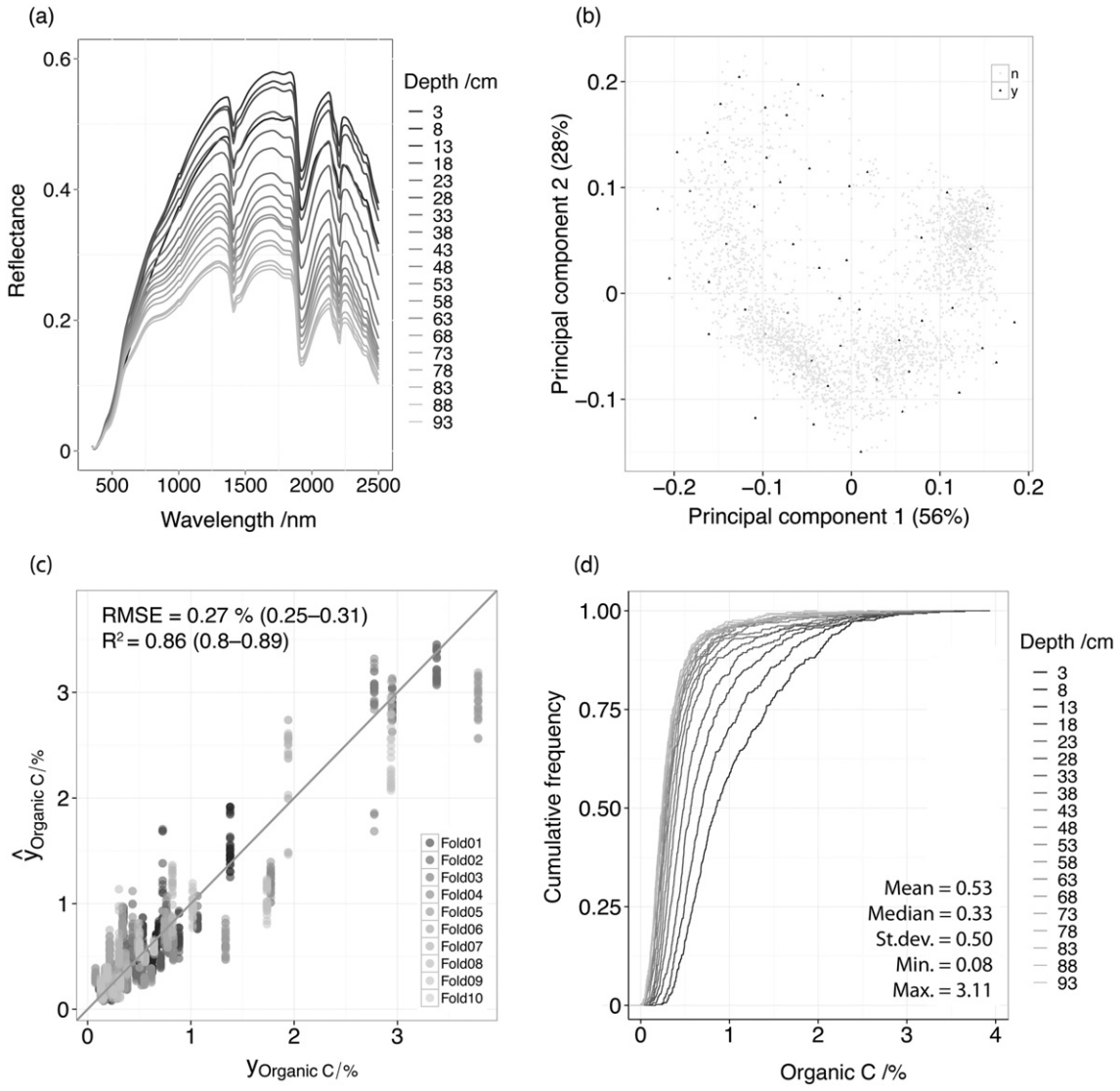

Fig. 3. Visible–near infrared spectroscopy. (a) Average spectrum of each depth over all soil cores, (b) first two principal component scores of the spectra from all of the soil cores (light grey points) showing the 50 samples selected for LECO soil organic C analysis (black points), (c) 10-fold cross validation of the spectroscopic model and (d) spatial cumulative distribution functions of the spectroscopic predictions by depth.

depths and these were analysed for soil organic C content by total combustion using a LECO carbon analyser (Rayment and Lyons, 2011).

To select the 50 samples, we used the spectra from the soil cores as follows. The measured reflectances, R, were first converted to apparent absorbance as $\log_{10}(1/R)$ and then preprocessed with the Savitzky–Golay smoothing (with a second-order polynomial and window size of 10 wavelengths) and first derivatives (Savitzky and Golay, 1964). These spectra were compressed using a PCA, and we used the first two scores, which accounted for 84% of the variance in the spectra, in the Kennard–Stone algorithm (Kennard and Stone, 1967) to select the samples. The algorithm is commonly used in spectroscopy to select representative sets of data for training and validating spectroscopic models.

Fig. 3b shows the PCA scores plot of the spectra from all of the soil cores, and highlights the 50 samples that were selected.

For the LECO analysis, subsamples were taken from the cores, centred on the exact locations where we had measured a spectrum. These subsamples were ground to a particle size $\leq$0.5 mm, any roots present were removed and the subsamples were homogenized by thorough end-over-end mixing. They were tested for the presence of carbonates with a 0.01 M HCl solution and those that contained carbonate, were pre-treated with sulphurous acid H_2SO_3, as a 5–6 wt.% SO_2 solution. Thus the analysis of the 50 subsamples produced data on the total soil organic C content in the sample.

2.4.3.1. Spectroscopic modelling to estimate soil organic C. We modelled the 50 measurements of soil organic C content with their corresponding preprocessed first derivative spectra using the decision trees algorithm Cubist (Quinlan, 1992), which is a form of piecewise linear decision tree, which partitions the response data into subsets within which their characteristics are similar with respect to the predictors. We have already reported the algorithm and its use with soil spectra (Viscarra Rossel and Webster, 2012).

The inaccuracy of the spectroscopic models was assessed by 10-fold cross validations (Fig. 3c) and using the root mean squared error (RMSE), which encompasses bias and imprecision. We also recorded the coefficient of determination R^2 between the observed and predicted values of soil organic C.

The resulting spectroscopic model was used to predict the soil organic C content of the soil cores at the depths where we had measured spectra (Fig. 3d). Thus, we made a total of 2780 estimates of soil organic C. At each depth, the bulk densities and soil organic C content data

were multiplied to derive volumetric estimates of soil organic C in units of g 100 cm^{-3}.

2.5. Estimation of soil organic C stocks at the sampling locations

The first step in estimating the soil organic C stock in the study area was to estimate the stocks in each profile at the sampling locations. As we described earlier, for each soil profile we have measurements of volumetric soil organic C at multiple depths. These measurements in a sampled soil profile i at depth j, denoted hereafter as $y_{i,j}$, are predictions made with the spectroscopic model, not errorless volumetric soil organic C values. The true volumetric soil organic C value of profile i at depth j is the sum of the spectroscopic model prediction and its error: $z_{i,j} = y_{i,j} + \varepsilon_{i,j}$.

For each core, we have model predictions of volumetric soil organic C at regular 5 cm intervals, so they form a 1-dimensional centred systematic sample. Assuming that the spectroscopic model predictions are unbiased, the 1-dimensional systematic sample average

$$\widehat{\bar{z}}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{i,j} \tag{2}$$

is an approximately unbiased estimate of the mean volumetric soil organic C content in soil profile i. In this equation m_i is the number of sampled depths in soil profile i. The aeric soil organic C stock in a sampled soil profile i to a given depth in units of t ha^{-1}, can then be estimated by multiplying this estimated mean volumetric soil organic C content by the thickness of the soil layer of interest, d_i, in units of cm:

$$\hat{t}_i = \widehat{\bar{z}}_i \cdot d_i. \tag{3}$$

In all of the sampled profiles, the thickness d_i is constant for the 0–10 cm and 0–30 cm layers, but varies for the 0–100 cm layer because of the occurrence of bedrock and shallower soil.

There are two sources of error contributing to the inaccuracy about the estimated soil profile mean volumetric organic C content. The first is the 1-dimensional sampling error down the profiles (we do not have a continuous registration of the spectrometer down the profile), and the second is the inaccuracy about the spectroscopic model predictions. The variance of the total error equals the sum of the variances of these two errors:

$$V\left(\widehat{\bar{z}}_i\right) = V_{\mathrm{p}}\left(\widehat{\bar{y}}_i\right) + V_{\mathrm{m}}\left(\widehat{\bar{\epsilon}}_i\right). \tag{4}$$

In this equation, $V_{\mathrm{p}}(\widehat{\bar{y}}_i)$ represents the sampling variance of the estimated mean volumetric organic C content of the soil profile, and $V_{\mathrm{m}}(\widehat{\bar{\epsilon}}_i)$ represents the variance of the average spectroscopic model prediction error at the sampled depths.

The sampling variance of the estimated soil profile mean, $V_{\mathrm{p}}(\widehat{\bar{y}}_i)$, does not need to be explicitly quantified in the design-based, model-assisted or model-based estimation of the soil organic C stock in the study area. In model-based prediction, the variogram is derived from estimates of the soil profile C stocks that already include the 1-D sampling error. In design-based and model-assisted estimation the 1-D sampling error is accounted for in the estimator of the sampling variance of the estimated soil organic C stock in the study area (see next section). In our case for the 0–10 cm and 0–30 cm depth layers, the thickness d_i is constant so that for these two layers the primary units had equal size, implying that within strata the soil organic C profiles were selected with probabilities that were proportional to their size. For the 0–100 cm depth layer, the thickness varied and thus the sizes of the soil organic C profiles were different. We ignored these different sizes of the soil organic C profiles in the 0–100 cm layer and assumed that the effect on the final results would be small.

The variance of the average spectroscopic model prediction error, $V_{\mathrm{m}}(\widehat{\bar{\epsilon}}_i)$, however, must be quantified. Spectroscopic 'measurements' of the soil organic C content at the sampled depths down the profile are not direct measurements, but are predictions from a spectroscopic model, *i.e.* expectations given the vis–NIR spectra in the model (see Section 2.4.3). Therefore, the smoothing of the variation in the predictions of organic C content between locations is not accounted for in the sampling or kriging variances, and must be quantified separately.

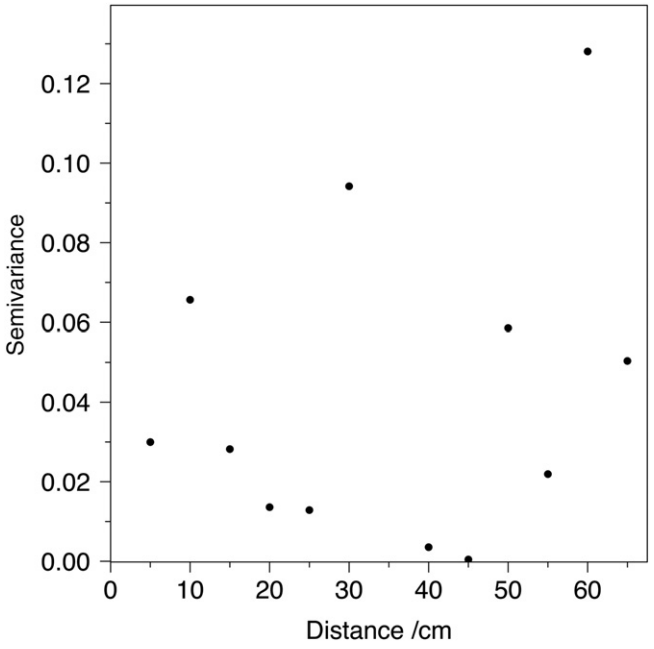

Fig. 4. Pooled 1-dimensional vertical experimental variogram of the spectroscopic model prediction residuals in the soil profiles.

The variance of the average spectroscopic model prediction error, $V_{\mathrm{m}}(\widehat{\bar{\epsilon}}_i)$, at a single new location is approximately equal to the residual variance of the spectroscopic model. In each soil profile we have multiple predictions, one per sampled depth. To estimate the variance of the average prediction error, averaged over all sampling depths, we needed to account for possible autocorrelation in the spectroscopic model prediction errors. Therefore we estimated the pooled 1-dimensional vertical experimental variogram of the spectroscopic model prediction residuals for each sampled soil profile with laboratory measurements of organic C, and computed the weighted average of the semivariances per lag, using the numbers of pairs of points as weights (Fig. 4).

This experimental variogram shows that there was no spatial structure in the spectroscopic model prediction errors and thus no evidence that the prediction errors at the sampled depths in a given soil profile were autocorrelated. Thus, the variance of the average prediction error can be estimated by the residual variance $V_{\mathrm{m}}(\epsilon)$ divided by the number of predictions m_i:

$$V_{\mathrm{m}}\left(\hat{\bar{\epsilon}}_i\right) = \frac{V_{\mathrm{m}}(\epsilon)}{m_i}. \tag{5}$$

By multiplying this variance by d_i^2 we obtain the variance of the model prediction error in the estimated aeric soil organic C stock in soil profile i:

$$V_{\mathrm{m}}(\hat{t}_i) = V_{\mathrm{m}}\left(\widehat{\bar{\epsilon}}_i\right) \cdot d_i^2. \tag{6}$$

2.6. Estimation of soil organic C stocks in the study area and its accuracy

We used design-based, model-assisted and model-based methods to estimate the total soil organic C stocks and their accuracies for each of the 0–10 cm, 0–30 cm and 0–100 cm soil layers in the study area. We describe the methods below, but note that in the model-assisted and model-based approaches we used predictions of the soil organic C

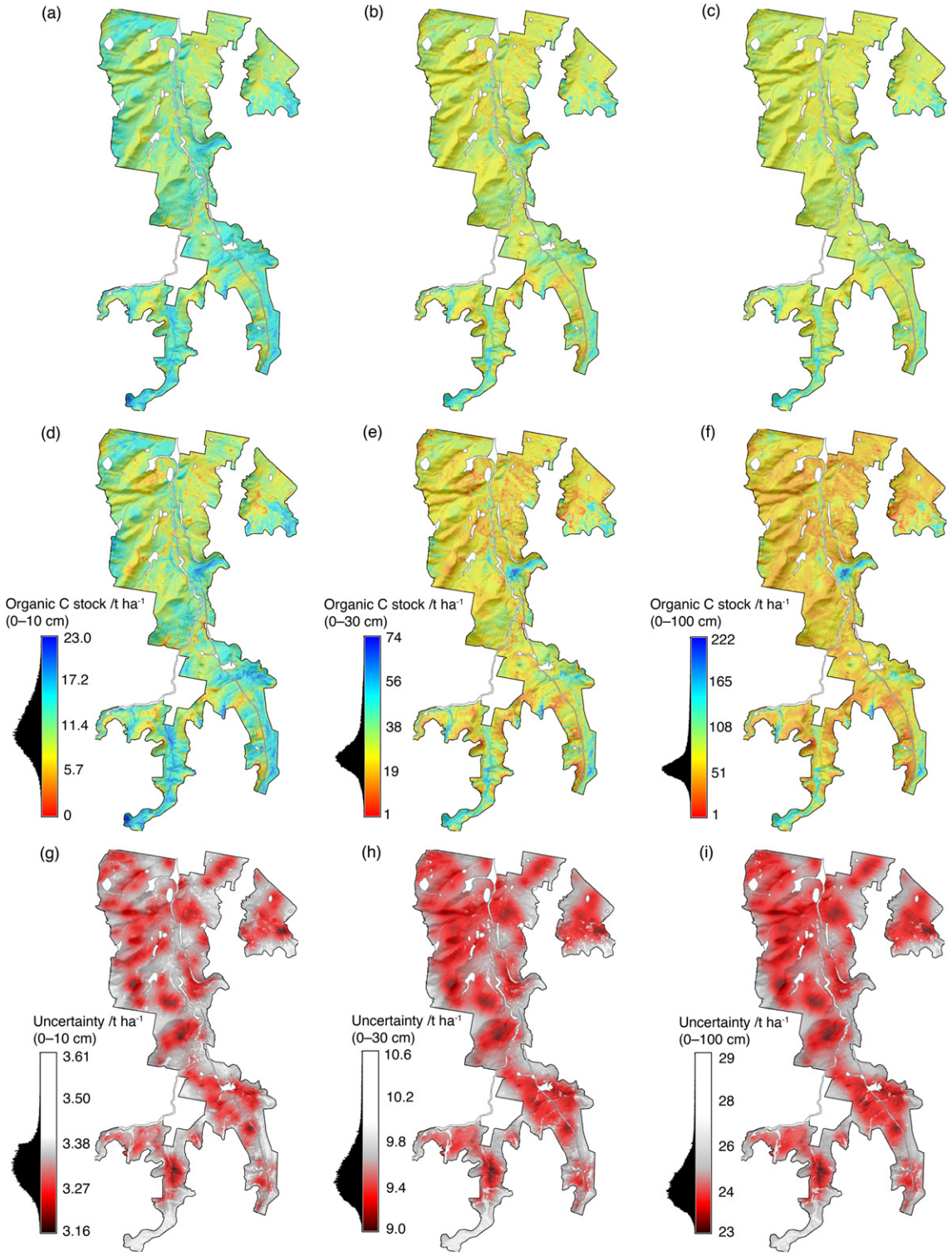

Fig. 5. Maps. (a–c) Soil organic C covariates derived using Cubist, (d–f) kriging with external drift maps of soil organic C stocks and (g–i) their prediction standard deviations, for each of the 0–10 cm, 0–30 cm and 0–100 cm soil layers, respectively. The colour scales for maps in (a–c) are the same as those in (d–f).

stocks at point-locations over the study area as the covariate in the estimation. For this, we set up a Cubist model at the sampling sites for which we had data on the aeric soil organic C stock and also interpolated data from the MMSP (Fig. 1) so that we could use the model to predict it elsewhere. The use of Cubist for spatial data analysis has been reported by many others, for example (Henderson et al., 2005). The soil organic C covariates for each depth layer derived from Cubist are shown in Fig. 5a–c.

2.6.1. Design-based estimation

The design-based estimator of the mean for stratified simple random sampling is:

$$\hat{\bar{t}}_{\pi} = \sum_{h=1}^{H} w_h \hat{\bar{t}}_h \tag{7}$$

where H is the number of strata, w_h is the weight of stratum h which is

equal to the relative area $w_h = A_h / \sum_h A_h$, and $\hat{\bar{t}}_h$ is the estimated mean soil organic C stock of stratum h. Within a stratum a simple random sample is selected, so an unbiased estimate of the stratum mean can be obtained by the sample average:

$$\hat{\bar{t}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\hat{t}_{hi} \tag{8}$$

with n_h the number of sampling points in stratum h. The estimator of Eqs. (7) and (8) is the Horvitz–Thompson estimator for stratified simple random sampling.

The sampling variance of the estimated mean is estimated by

$$\hat{V}_{\mathrm{p}}\left(\hat{\bar{t}}_\pi\right) = \sum_{h=1}^{H} w_h^2 \frac{\hat{S}^2{}_h(\hat{t})}{n_h} \tag{9}$$

where $\hat{S}^2{}_h(\hat{t})$ is the estimated spatial variance of the estimated soil organic C stock per soil profile in stratum h:

$$\hat{S}^2{}_h(\hat{t}) = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(\hat{t}_{hi} - \hat{\bar{t}}_h\right)^2. \tag{10}$$

The sampling variance (Eq. (9)) already accounts for the sampling variance of the estimated soil organic C stocks per soil profile, because the spatial variance (Eq. (10)) is the variance of the *estimated* soil organic C stocks per soil profile, not of the true soil organic C stocks per soil profile.

However, the sampling variance does not account for the variance of the spectroscopic model prediction error (Eq. (5)), so we estimated this variance by:

$$\hat{V}_{\mathrm{m}}\left(\hat{\bar{t}}_\pi\right) = \sum_{h=1}^{H} w_h^2 \frac{\sum_{i=1}^{n_h} V_{\mathrm{m}}(\hat{t}_{hi})}{n_h^2} \tag{11}$$

For the 0–10 cm and 0–30 cm layers, $V_{\mathrm{m}}(\hat{t}_{hi})$ is constant for all soil profiles in a stratum, so that $\sum_{hi=1}^{n_h} V_{\mathrm{m}}(\hat{t}_{hi})/n_h^2 = V_{\mathrm{m}}(\hat{t}_{hi})/n_h$.

To obtain the total variance of the estimated mean of aeric soil organic C stock in the study area we first added the spectroscopic model variance, $\hat{V}_{\mathrm{m}}(\hat{\bar{t}}_\pi)$, to the sampling variance, $\hat{V}_{\mathrm{p}}(\hat{\bar{t}}_\pi)$, as we describe in (Eq. (4)), and multiplied the total variance of the estimated mean by the squared area in ha.

2.6.2. Model-assisted estimation

As mentioned previously, we used the Cubist predictions of soil organic C stock at point locations as a covariate in the regression estimator to increase the precision of the estimated mean and total C stock for each depth layer in the study area (Cochran, 1977; Brus, 2000):

$$\hat{\bar{t}}_{\mathrm{regr}} = \hat{\bar{t}}_\pi + \hat{B}_1\left(\bar{t}_{\mathrm{cub}} - \hat{\bar{t}}_{\mathrm{cub},\pi}\right) \tag{12}$$

where $\hat{\bar{t}}_\pi$ is the Horvitz–Thompson estimator of the mean soil organic C stock in the study area (see previous section), $\hat{B}_1$ is the estimated regression coefficient (slope), $\bar{t}_{\mathrm{cub}}$ is the true mean of the Cubist organic C stock predictions, and $\hat{\bar{t}}_{\mathrm{cub},\pi}$ is the Horvitz–Thompson estimator of the mean of the Cubist predictions.

We estimated the regression coefficient B_1 using (Cochran, 1977):

$$\hat{B}_1 = \frac{\sum_h w_h^2 \frac{1}{n_h(n_h-1)}\sum_i\left(\hat{t}_{hi} - \hat{\bar{t}}_h\right)\left(\hat{t}_{\mathrm{cub},hi} - \hat{\bar{t}}_{\mathrm{cub},h}\right)}{\sum_h w_h^2 \frac{1}{n_h(n_h-1)}\sum_i\left(\hat{t}_{\mathrm{cub},hi} - \hat{\bar{t}}_{\mathrm{cub},h}\right)^2}. \tag{13}$$

This is the *combined* regression estimator because data from all strata are combined to estimate the regression coefficient B_1. To estimate the variance of the combined regression estimator, we first needed to estimate the intercept:

$$\hat{B}_0 = \hat{\bar{t}}_\pi - \hat{B}_1 \times \hat{\bar{t}}_{\mathrm{cub},\pi}, \tag{14}$$

and compute the residuals:

$$\hat{e}_i = \hat{t}_i - \left(\hat{B}_0 + \hat{B}_1 \times \hat{t}_{\mathrm{cub},i}\right), \tag{15}$$

where $\hat{t}_{\mathrm{cub},i}$ is the Cubist prediction of the soil organic C stock for sampling location i. Then, we estimated the variance of $\hat{\bar{t}}_{\mathrm{regr}}$ using:

$$\hat{V}\left(\hat{\bar{t}}_{\mathrm{regr}}\right) = \sum_{h=1}^{H} w_h^2\left(\frac{\hat{S}^2{}_h(\hat{e}_i)}{n_h}\right). \tag{16}$$

As for the design-based approach, to obtain the total variance of the estimated mean aeric soil organic C stock in the study area, we first add the sampling variance (Eq. (16)) to the variance of the spectroscopic model prediction error (Eq. (11)) and multiply the total variance by the squared area in ha.

2.6.3. Model-based estimation

Model-based estimation of the total soil organic C stocks in the study area were obtained by block kriging with an external drift (KED) (Webster and Oliver, 2007), based on the linear mixed model:

$$t(\mathbf{s}) = \beta_0 + \beta_1 t_{\mathrm{cub}}(\mathbf{s}) + \eta(\mathbf{s}) + \delta(\mathbf{s}), \tag{17}$$

where β_0 and β_1 are regression coefficients, $t_{\mathrm{cub}}(\mathbf{s})$ is the field with Cubist predictions of the soil organic C stock, $\eta(\mathbf{s})$ a random field of residuals with zero mean and covariance $C(h)$, in which h is the distance between two locations, and $\delta(\mathbf{s})$ is a random field with the spectroscopic model prediction errors in the soil organic C stocks at the sampling locations, with zero means and variances $V_{\mathrm{m}}(\hat{t}_i)$ (Eq. (6)). We assumed that these spectroscopic model prediction errors $\delta(\mathbf{s})$ were spatially independent (Fig. 4).

Using Eq. (17), we estimated the mean soil organic C stock in the study area using (Diggle and Ribeiro, 2007):

$$\hat{\bar{t}}_{\mathrm{KED}} = \mathbf{d}'\beta + \bar{\mathbf{c}}'\mathbf{C}^{-1}\left(\mathbf{t} - \mathbf{D}\hat{\beta}\right) \tag{18}$$

where $\mathbf{d}$ is the vector with the mean values of the Cubist predictor, $\mathbf{d} = (1, \bar{t}_{\mathrm{cub}})'$, with $\bar{t}_{\mathrm{cub}}$ being the average of the Cubist predictions over all prediction nodes, $\bar{\mathbf{c}}$ is the vector with mean model covariances between the soil organic C stocks at the sampling locations and the study area, $\mathbf{C}$ is the matrix with covariances between the sampling locations, $\mathbf{t}$ is the vector with the soil organic C stocks at the sampling locations, $\hat{t}_i, i = 1 \cdots n$ (Eq. (3)), $\mathbf{D}$ is the matrix containing ones in the first column and the Cubist predictions at the sampling locations in the second, and $\hat{\beta}$ is the vector with the two estimated regression coefficients.

The inaccuracy of $\hat{t}_i$ due to spectroscopy modelling error was accounted for by adding the estimated variance (Eq. (6)) to the diagonal of the matrix $\mathbf{C}$ (de Marsily, 1986; Knotters et al., 1995). The two regression coefficients and variogram parameters were estimated by restricted maximum likelihood (REML) (Lark et al., 2006) and fitted by minimizing the negative loglikelihood using the differential evolution algorithm (Price et al., 2006).

The variance of the model-based estimate of the mean soil organic C stock in the study area was computed using (Corsten, 1989):

$$V\left(\hat{\bar{t}}_{\mathrm{KED}}\right) = \bar{c}_{\mathcal{A}} - \bar{\mathbf{c}}'\mathbf{C}^{-1}\bar{\mathbf{c}} + \mathbf{d}_a'\left(\mathbf{D}'\mathbf{C}^{-1}\mathbf{D}\right)^{-1}\mathbf{d}_a, \tag{19}$$

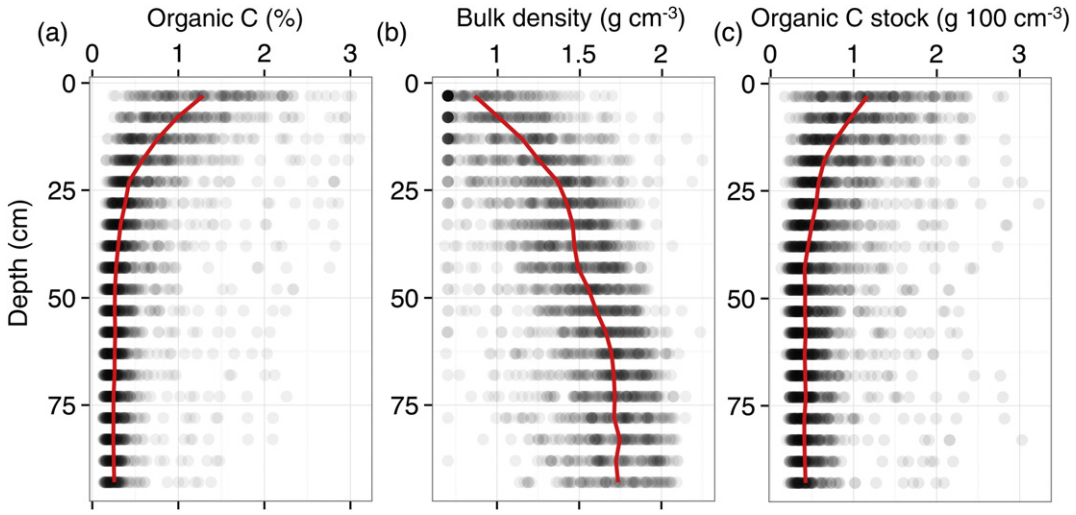

Fig. 6. (a) Spectroscopic model predictions of soil organic C, (b) gamma attenuation measurements of bulk density and (c) volumetric soil organic C at the proximally sensed locations on the 1 m soil cores. The thick (red) lines down the profiles represent the median values of the sensor measurements.

where $\bar{c}_{\mathcal{A}}$ is the mean covariance within the study area, and $\mathbf{d}_a = \mathbf{d} - \mathbf{D}'\mathbf{C}^{-1}\bar{\mathbf{c}}$.

2.6.4. Model-based prediction of soil organic C stock at points–mapping

The linear mixed model in Eq. (17) was also used to predict the soil organic C stocks at the nodes of a 5 m grid discretizing the study area. These predictions at points were obtained by replacing the mean covariances in the vector $\bar{\mathbf{c}}$ by covariances between the sampling points and the prediction node.

We used cross validation to assess performance of the KED predictions, where an observation is removed in turn from the dataset and the remaining observations are used to predict at the site of the removed observation. In the assessment we used the linear correlation coefficient ρ between the observed and predicted values, the mean error (ME) to assess bias and the root mean squared error (RMSE) to assess inaccuracy. We evaluate the performance of the prediction variances by computing the mean ($\bar{\theta}_p$) and median ($\tilde{\theta}_p$) of the standardized squared prediction errors (Voltz and Webster, 1990), $\bar{\theta}_p$ to assess the goodness of the estimate and $\tilde{\theta}_p$ as a more robust estimate of the prediction variance (Lark, 2000). When the fitted KED model is accurate and the uncertainty is well characterized, assuming normally distributed prediction errors, the expected value for $\bar{\theta}_p$ should be 1 and $\tilde{\theta}_p$ should be 0.455.

3. Results

3.1. Soil sampling design

The maps of the first four principal components of the sensor covariates (Fig. 2a–d) have captured the main patterns of the spatial variation in the soil and landscape of the study area that is represented by the sensors (Fig. 1).

The map of the 75 *k*-means strata is shown in Fig. 2e, and the locations of the sampling units are shown in Fig. 2f. In Fig. 2f, with few exceptions, the target locations of the sampling units from our design (crosses) are the same as the actual locations from where the soil samples were taken (circles). The differences between the two sets is caused by inaccessibility of the target locations due to either rough terrain or wet conditions at the time of sampling. For those inaccessible locations, alternate randomly selected sampling locations within each stratum were provided to the surveyors. There were two sites in the southernmost region of the study site that were not sampled because the sites were inaccessible.

3.2. Sensor measurements of soil organic C stocks

The vis–NIR reflectance of the soil cores decreased steadily from the surface to the deepest layer. The spectra of soil that is near the surface is clearly different to spectra from deeper in the soil, they have less pronounced absorptions that are due to minerals and a smoothly increasing shape in the visible range up to around 1000 nm, which is characteristic of soil with more organic matter. Absorptions from iron oxides near 800–1000 nm and those from clay minerals near 1450 nm, 1900 nm, 2160 nm and 2200 nm, are more prominent in the spectra of the deeper layers (Fig. 3a).

Fig. 3c shows a plot of the observed versus predicted values of soil organic C from the 10-fold cross validation of the spectroscopic model that we used to estimate the soil organic C contents of the soil cores at depths where we had only spectra. The RMSE of prediction was 0.27% organic C, ranging from 0.25–0.31% organic C, and the R^2 was 0.86, ranging from 0.8–0.89. The cumulative distribution functions (CDF) of the predictions we made with the spectroscopic model for each depth separately is shown in Fig. 3d. The estimates produced a skewed distribution of soil organic C content, with a mean of 0.53% and median of 0.33% organic C (Fig. 3d). The soil organic C, bulk density and organic C stocks profiles are shown in Fig. 6.

Generally, soil organic C content decreased with depth (Fig. 6a, c), while the bulk density of the soil profiles increased with depth (Fig. 6b). The sensor measurements of bulk density and soil organic C content were skewed. The median bulk density was 1.51 g cm^{-3} and the median volumetric soil organic C content of the profiles was 0.50 g 100 cm^{-3}. In Fig. 6, the thick lines down the profiles represent the median values of the sensor measurements.

Table 2
Statistical summary of the soil organic C stock estimates $\hat{t}_i$ for each depth layer, in t ha^{-1}.

Depth (cm)	Mean	St. dev.	Coeff. var.	Min.	1st qu.	Median	3rd qu.	Max.
0–10	11.45	4.65	0.41	2.49	7.90	11.15	14.57	23.63
0–30	27.31	12.14	0.44	8.25	18.81	24.47	33.51	81.87
0–100	66.43	34.51	0.52	25.64	46.77	56.51	71.01	244.40

Table 3
Design-based, model-assisted and model-based estimates of mean and total soil organic C stocks, and their variances and standard errors.

Estimator	Depth (cm)	$\bar{t}_{cub}$ (t ha^{-1})	$\hat{\bar{t}}$ (t ha^{-1})	$\hat{V}(\hat{\bar{t}})$	$\hat{t}_{\mathcal{A}}$(t)	$\hat{V}(\hat{t}_{\mathcal{A}})$	$\hat{S}(\hat{t}_{\mathcal{A}})$(t)	$\hat{V}_{\mathrm{p}}$	$\hat{V}_{\mathrm{m}}$	$\hat{S}_p$ (t)	$\hat{S}_m$ (t)
Design-based Horvitz–Thompson	0–10		11.44	0.18	31,909.25	1,357,431.06	1165.09	1,323,220.21	34,210.85	1150.31	184.96
	0–30		28.08	1.11	78,324.52	8,625,077.77	2936.85	8,590,866.93	34,210.85	2931.02	184.96
	0–100		62.55	6.22	174,446.26	48,406,856.19	6957.50	48,373,420.60	33,435.59	6955.10	182.85
Model-assisted regression estimator	0–10	11.77	11.38	0.15	31,740.51	1,154,660.82	1074.55	1,120,449.97	34,210.85	1058.51	184.96
	0–30	30.72	28.13	1.01	78,444.98	7,868,329.11	2805.05	7,834,118.26	34,210.85	2798.95	184.96
	0–100	85.60	63.05	5.25	175,850.13	40,791,389.83	6386.81	40,757,954.24	33,435.59	6384.20	182.85
Model-based KED	0–10	11.77	11.43	0.11	31,886.18	846,339.57	919.97				
	0–30	30.72	27.30	0.75	76,146.16	5,864,005.72	2421.57				
	0–100	85.60	61.94	4.69	172,748.75	36,492,213.61	6040.88				

Table 4
Model parameters.

Estimator	Depth (cm)	Regression		Variogram			
		$\hat{B}_0$	$\hat{B}_1$	Model	C_0	C	r (m)
Model assisted combined regression estimator	0–10	1.10	−1.56				
	0–30	0.74	5.44				
	0–100	0.94	−17.30				
Model-based KED	0–10	−3.95	1.31	Spherical	7.96	3.31	817.38
	0–30	−7.42	1.13	Spherical	68.32	28.78	1161.85
	0–100	−44.53	1.25	Spherical	407.82	204.74	1209.10

3.3. Estimates of the soil organic C stocks at the sampling locations

Table 2 summarizes the statistics of the estimated soil organic C stocks at the sampling locations for the 0–10 cm, 0–30 cm and 0–100 cm layers.

The average soil organic C stocks in the 0–10 cm layer is 11.45 t ha^{-1}, it ranged from 2.49 t ha^{-1} to 23.63 t ha^{-1} (Table 2). In the 0–30 cm layer the average stock was 27.31 t ha^{-1} and the range was 8.25–81.87 t ha^{-1}. On average, there is more than twice as much carbon in the 0–100 cm layer compared to the 0–30 cm layer. The average in the 0–100 cm layer is 66.43 t ha^{-1} and the range is 25.64–244.40 t ha^{-1} (Table 2). The frequency distribution of the stocks in the 0–100 cm layer is more skewed than those of the 0–30 cm and 0–10 cm layers, and the data in this layer are also more variable (Table 2).

3.4. Design-based, model-assisted and model-based estimates of soil organic C stocks

The design-based, model-assisted and model-based estimates of the mean and total soil organic C stock and their inaccuracies for each of the three depth layers in the study area are shown in Table 3.

The design-based estimate of the total soil organic C stock, $\hat{t}_{\mathcal{A}}$, in the 0–10 cm layer was 31.909 Gg (1 gigagram (Gg) = 10^9g = 1 kilotonne), in the 0–30 cm it was 78.325 Gg and in the 0–100 cm layer it was 174.446 Gg organic C. For each depth layer, the overall inaccuracy of the total soil organic C stocks, presented as their standard errors, $\hat{S}(\hat{t}_{\mathcal{A}})$, were 1.165 Gg, 2.936 Gg and 6.957 Gg organic C (Table 3).

Using the Cubist predictions of carbon stock at point locations as a covariate in the combined regression estimator of the model-assisted approach produced similar estimates of $\hat{t}_{\mathcal{A}}$ compared to the design-based estimates at each of the three depth layers (Table 3). However, the inaccuracies of these model-assisted estimates were smaller, with $\hat{S}(\hat{t}_{\mathcal{A}})$ values of 1.075 Gg for the 0–10 cm, 2.805 Gg for the 0–30 cm and 6.387 Gg organic C for the 0–100 cm layers (Table 3). The linear regression coefficients used in the combined regression estimator are shown in Table 4.

The estimated slopes, $\hat{B}_1$, of all three layers had negative values, suggesting that the Cubist predictions alone overestimated the soil organic C stocks of the soil profiles. This is particularly evident in the estimate of the 0–100 cm layer where the mean of the Cubist predictions is larger than the design-based estimates of the mean (Table 3). The model-assisted estimator corrected for bias in the Cubist predictions.

For each depth layer, the sampling variance, $\hat{V}_{\mathrm{p}}$, in the design-based and model-assisted estimators accounted for more than 98% of the total inaccuracy in our estimates (Table 3). The contribution to the total inaccuracy of the errors due to the spectroscopic modelling $\hat{V}_{\mathrm{m}}$ was only 2% and almost negligible.

The model-based estimates of the total soil organic C stock, $\hat{t}_{\mathcal{A}}$, in the study area for each of the three depth layers, and their inaccuracies, are shown in Table 3 (bottom block). The model-based estimate of $\hat{t}_{\mathcal{A}}$ in the 0–10 cm layer was 31.886 Gg, in the 0–30 cm it was 76.146 Gg and in the 0–100 cm layer it was 172.749 Gg organic C. The inaccuracies of our estimates, which account for both the sampling error down the profiles and the spectroscopic model errors were 0.920 Gg, 2.422 Gg and 6.041 Gg organic C, for each of the three depth layers, respectively (Table 3). Not accounting for the two sources of error produced smaller inaccuracies with values of 0.770 Gg for the 0–10 cm layer, 2.259 Gg for the 0–30 cm layer and 5.856 Gg for the 0–100 cm layer.

The variance of the model-based estimate of soil organic C stock in the study area is smaller than the variance of the model-assisted regression estimator. Like in the model-assisted method, the model-based

Table 5
Comparison of design-based, model-assisted and model-based estimates of the total organic C stocks, and confidence limits (CL).

Estimator	Depth (cm)	$\hat{t}_{\mathcal{A}}$ (Gg)	Lower 95% CL	Upper 95% CL	Width CL
Design-based Horvitz–Thompson	0–10	31.9092	29.6257	34.1928	4.6
	0–30	78.3245	72.5683	84.0807	11.5
	0–100	174.4463	160.8096	188.0830	27.3
Model-assisted regression estimator	0–10	31.7405	29.6344	33.8466	4.2
	0–30	78.4450	72.9471	83.9429	11.0
	0–100	175.8501	163.3320	188.3683	25.0
Model-based KED	0–10	31.8862	30.0830	33.6893	3.6
	0–30	76.1462	71.3999	80.8924	9.5
	0–100	172.7487	160.9086	184.5889	23.7

Table 6
Quality indices of KED predictions of soil organic C obtained by leave-one-out cross-validation: correlation between the observed and predicted values (ρ), mean error (ME), root mean squared error (RMSE) and the mean ($\bar{\theta}_p$) and median ($\tilde{\theta}_p$) values of the standardized squared prediction errors.

	Depth layer (cm)		
Statistic	0–10	0–30	0–100
ρ	0.715	0.633	0.640
ME (t ha^{-1})	−0.002	0.005	−0.004
RMSE (t ha^{-1})	3.268	9.446	25.240
$\bar{\theta}_p$	0.971	0.969	0.958
$\tilde{\theta}_p$	0.445	0.396	0.290

estimator profited from the relation between the Cubist predictions of the organic C stock and the observed organic C stock per soil profile (to the particular depth), but the former did not exploit the spatial correlation in the models' residuals.

In Table 5 we compare the design-based, model-assisted and model-based estimates of the total organic C stocks in the study area and their confidence intervals.

The widths of the 95% confidence limits of the model-based estimates were narrower than those of the model-assisted and the design-based approaches (Table 5).

3.5. Model-based prediction of soil organic C stock at points—mapping

The REML estimates of the KED model parameters are shown in Table 4. They show that the proportion of the variation that is explained by the regression part of the KED model decreases with depth. There is less spatially correlated variance in the variogram of the 0–10 cm layer compared to the 0–30 cm and 0–100 cm layers (Table 4).

Fig. 5d–f are the maps of soil organic C stocks for the three soil layers and Fig. 5g–iare their prediction standard deviations. In the map of the 0–10 cm layer (Fig. 5d), values range from negligible amounts of soil organic C in the north–northeast, and generally increase towards the south, to around 23 t ha^{-1} in lower and wetter portions of the landscape. Spatial patterns are similar for the deeper layers. In the 0–30 cm layer (Fig. 5e), C stocks range from around 1 t ha^{-1} to 70 t ha^{-1}, and in the 0–100 cm layer (Fig. 5f), from around 1 t ha^{-1} to 220 t ha^{-1}.

The quality indices of the KED predictions of the soil organic C stocks at points for the three layers as obtained by leave-one-out cross-validation are shown in Table 6.

The spatial estimates were relatively unbiased (small ME) so that the primary contributions to their inaccuracies, represented by their RMSE, was from their imprecision. The RMSE and the mean organic C stock estimates (Table 2) increased with the thickness of the soil layer. For all three layers the RMSEs were somewhat smaller than the standard deviations of the soil organic C stocks in the sampled soil profiles (Table 2). The ratio of the standard deviation to the RMSE varied from 1.3–1.4. For each layer, the values of $\bar{\theta}_p$ are all close to 1 (Table 6), suggesting that the estimates of the kriging prediction error variance were unbiased and reliable. However, values of $\tilde{\theta}_p$ were smaller than the expected value of 0.455, which indicates an overestimation of the variance of the prediction errors, particularly in the 0–30 cm and 0–100 cm layers (Table 6).

3.6. Discussion and conclusions

We have shown that data from proximal soil sensors can be used to effectively and efficiently measure soil organic C stocks. The MSSP produced data to inform both the sampling and the estimation of organic C stocks. The soil profile data we used was measured using a multi-sensor system with spectroscopic and gamma attenuation sensors, which produced accurate measures of soil organic C and bulk density at the sampling locations, every 5 cm to a depth of 1 m.

The stratified simple random sampling design allowed design-based, model-assisted and model-based estimation of the total organic C stocks in the study area. If the sampling locations had not been selected by probability sampling, design-based and model-assisted estimation would have been impossible. This flexibility in statistical inference is an advantage of probability sampling.

We chose a sampling design that produced fair coverage in both feature and geographical spaces, so that we could also use the sample for model-based prediction by KED and using the Cubist predictions as the external drift. We do not claim that the design we implemented is optimal, and we welcome research into the design of probability samples that are efficient both in design- and model-based inference.

We did not find published literature on sampling methods that are designed for both design- and model-based estimation. There are efficient designs to construct optimal strata for design-based inference when there is adequate prior information on the target variable, e.g. the cum$\sqrt{f}$ (Cochran, 1977) and the Ospats (de Gruijter et al., 2015) methods. There are sampling designs that are useful for mapping (e.g. Simbahan and Dobermann, 2006), but they do not use probability sampling and thus cannot be used for design-based estimation. Balanced sampling with geographic spreading of sampling locations might be potentially useful for achieving this dual aim (Grafström and Tillé, 2013; Brus, 2015).

The design-based, model-assisted and model-based estimates of total soil organic C stock in our study area were very similar for all three depth layers. However, the variances of the model-assisted and model-based estimates were smaller compared to those of the design-based method. Evidently, the Cubist-derived covariate helped to improve the accuracy of their estimates. At all depths, the model-based method produced the smallest variances. The reason is that the approach is able to use the spatial correlation in the KED model's residuals to improve the accuracy of the estimates.

Baseline estimates with the smallest variances are attractive because they allow future changes in organic C to be more easily detected. However, we note that these estimates of the variance rely on the validity of the KED model assumptions and do not account for inaccuracy in the residual variogram parameters. To explore the sensitivity of the model-based variance to the model parameters, we also fitted double spherical models without nugget. The block-kriging variances were almost the same as the variances that we report in Table 3, obtained with the spherical model with nugget.

An advantage of the design-based and model-assisted approaches is that their estimates of the baseline soil organic C stocks and their variances do not rely on the assumptions of a model, that is, estimation is model-free.

An advantage of the model-based approach is that it can also be used to map the soil organic C stocks. We note that although the model-based approach produced the smallest variance of the predicted total soil organic C stocks, the results cannot be generalized to other sample sizes and types of sampling designs. For smaller sample sizes and consequently lower sampling densities we expect the profit from the spatial autocorrelation of the residuals to become smaller, so that the model-assisted estimates may become as precise.

In our implementation of the design-based, model-assisted and model-based methods, we accounted for the errors due to the regular 5 cm interval sampling down the profiles where the spectroscopic predictions were made as well as the spectroscopic model prediction errors. The variance of the average spectroscopic model prediction error was relatively small. We did not account for the errors of the measurements of bulk density, however, these are likely to be significantly smaller than the error from the spectroscopic modelling (Lobsey and Viscarra Rossel, in press).

We found that, on average, there was more than twice as much organic C in the 0–30 cm layer than in the 0–10 cm, and more than

twice as much in the 0–100 cm layer than in the 0–30 cm layer, suggesting that in soil under unimproved pastures, remnant vegetation and forests there is good rationale for measuring soil organic C beyond the commonly recommended depth of 0–30 cm.

Acknowledgments

We thank Carbon Link Pty. Ltd. for providing us with the opportunity to perform the baseline survey and for the soil sampling. We also thank the Australian Government's Department of Agriculture's Filling the Research Gap Round 2 for funding part of this research through the project (1194194-91), 'An innovative solution for accurate and affordable estimates of soil carbon'. We thank Seija Tuomi for her technical contributions.

References

Brus, D.J., 2015. Balanced sampling: a versatile sampling approach for statistical soil surveys. Geoderma 253–254, 111–121.
Brus, D.J., 2000. Using regression models in design-based estimation of spatial means of soil properties. Eur. J. Soil Sci. 51, 159–172.
Brus, D.J., de Gruijter, J.J., 1993. Design-based versus model-based estimates of spatial means. Theory and application in environmental soil science. Environmetrics 4, 123–152.
Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil with discussion. Geoderma 80 (1–2), 1–59.
Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.
Cook, S.E., Corner, R.J., Groves, P.R., Grealish, G.J., 1996. Use of airborne gamma radiometric data for soil mapping. Soil Res. 34 (1), 183–194.
Corsten, L.C.A., 1989. Interpolation and optimal linear prediction. Statistica Neerlandica 43, 69–84.
de Gruijter, J.J, Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag, Berlin Heidelberg.
de Gruijter, J.J., Minasny, B., McBratney, A.B., 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. J.Surv. Stat.Methodol. 3 (1), 19–42.
de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Math. Geol. 22, 407–415.
de Marsily, G., 1986. Quantitative Hydrogeology: Groundwater Hydrology for Engineers. Academic Press, Orlando, Florida.
Diggle, P., Ribeiro Jr., P.J., 2007. Model-based Geostatistics. Springer Series in Statistics, Springer, New York.
Grafström, A., Tillé, Y., 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24 (2), 120–131. http://dx.doi.org/10.1002/env.2194.
GRASS Development Team, 2012. Geographic resources analysis support system (GRASS) software. Version 6 (4), 1 URL http://grass.osgeo.org.
Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124 (3–4), 383–398.
Isbell, R.F., 2002. The Australian Soil Classification. revised edition. Edition. CSIRO Publishing, Collingwood, Victoria.
IUSS Working Group WRB, 2006. World Reference Base for Soil Resources. World soil Resources Reports No. 103, FAO, Rome.
Izaurralde, R.C., Rice, C.W., Wielopolski, L., Ebinger, M.H., Reeves III, J.B., Thomson, A.M., Harris, R., Francis, B., Mitra, S., Rappaport, A.G., Etchevers, J.D., Sayre, K.D., Govaerts, B., McCarty, G.W., 2013. Evaluation of three field-based methods for quantifying soil carbon. PLoS One 8 (1), e55560 01.
Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10 (2), 423–436.
Kennard, R.W., Stone, L.A., 1967. Computer aided design of experiments. Technometrics 11 (1), 137–148.
Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246.
Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51, 137–157.
Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57 (6), 787–799.
Lobsey, C., Viscarra Rossel, R.A., 2016. Sensing soil bulk density for more accurate carbon accounting. Eur. J. Soil Sci. (in press).
Lorenz, K., Lal, R., 2005. The depth distribution of soil organic carbon in relation to land use and management and the potential of carbon sequestration in subsoil horizons. of Advances in Agronomy vol. 88. Academic Press, pp. 35–66.
Machmuller, M.B., Kramer, M.G., Cyle, T.K., Hill, N., Hancock, D., Thompson, A., 2015. Emerging land use practices rapidly increase soil organic matter. Nat. Commun. 6 (6995).
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley on Mathematical Statistics and Probability vol. 1. University of California, University of California Press, Berkeley, California, pp. 281–297.
Martens, H., Næs, T., 1989. Multivariate Calibration. Wiley.
Miklos, M., Short, M.G., McBratney, A., Minasny, B., 2010. Mapping and comparing the distribution of soil carbon under cropping and grazing management practices in narrabri, north-west New South Wales. Aust. J. Soil Res. 48 (3), 248–257.
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.
Papritz, A., Webster, R., 1995. Estimating temporal change in soil monitoring: I. Statistical theory. Eur. J. Soil Sci. 46, 1–12.
Price, K.V., Storn, R.M., Lampinen, J.A., 2006. Differential evolution — a practical approach to global optimization. Natural Computing. Springer-Verlag.
Quinlan, J., 1992. Learning with continuous classes. In: Adams, A., Sterling, L. (Eds.), Proceedings AI'92, 5th Australian Conference on Artificial Intelligence. World Scientific, Singapore, pp. 343–348.
Rayment, G., Lyons, D., 2011. Soil Chemical Methods — Australasia. CSIRO Publishing, Collingwood, Victoria.
Särndal, C., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36 (8), 1627–1639.
Simbahan, G.C., Dobermann, A., 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma 133 (3–4), 345–362.
Stenberg, B., Viscarra Rossel, R., Mouazen, A., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Adv. Agron. 107, 163–215.
Viscarra Rossel, R., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time. Adv. Agron. 113, 237–282.
Viscarra Rossel, R.A., Taylor, H.J., McBratney, A.B., 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. Eur. J. Soil Sci. 58 (1), 343–353.
Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. Eur. J. Soil Sci. 63 (6), 848–860.
Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil propoerties from sample information. J. Soil Sci. 41, 473–490.
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.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. 2nd Edition. John Wiley & Sons, Ltd.