Science of the Total Environment 631–632 (2018) 377–389

Contents lists available at ScienceDirect

Science of the Total Environment

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

Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions

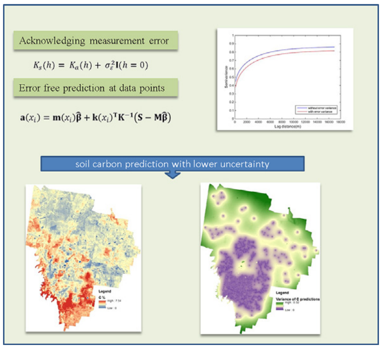
P.D.S.N. Somarathna *, Budiman Minasny, Brendan P. Malone, Uta Stockmann, Alex B. McBratney

Sydney Institute of Agriculture, The University of Sydney, New South Wales, Australia

HIGHLIGHTS

- Measurement errors can be filtered through incorporating in the covariance structure of the spatial model.
- Acknowledging measurement errors in spatial modeling yields a lower uncertainty in spatial predictions.
- MCMC techniques can be used to define the posterior density of measurement error variance.
- Performance of REML-EBLUP approach is comparable to MCMC techniques in terms of bias correction of the spatial model.

GRAPHICAL ABSTRACT


Acknowledging measurement error
$K_s(h) = K_a(h) + \sigma_\varepsilon^2 \mathbf{I}(h = 0)$
Error free prediction at data points
$\mathbf{a}(x_i) = \mathbf{m}(x_i)\hat{\boldsymbol{\beta}} + \mathbf{k}(x_i)^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{S} - \mathbf{M}\hat{\boldsymbol{\beta}})$
Semivariance
Lag distance(m)
without error variance
with error variance
soil carbon prediction with lower uncertainty
Legend
C %
High : 7.54
Low : 0
Legend
Variance of C predictions
High : 0.32
Low : 0

ARTICLE INFO

Article history:
Received 11 September 2017
Received in revised form 11 February 2018
Accepted 25 February 2018
Available online xxxx

Editor: Ouyang Wei

Keywords:
Infrared spectroscopy
Measurement error
Variogram
REML-EBLUP
MCMC
Filtered kriging

ABSTRACT

Spatial modelling of environmental data commonly only considers spatial variability as the single source of uncertainty. In reality however, the measurement errors should also be accounted for. In recent years, infrared spectroscopy has been shown to offer low cost, yet invaluable information needed for digital soil mapping at meaningful spatial scales for land management. However, spectrally inferred soil carbon data are known to be less accurate compared to laboratory analysed measurements. This study establishes a methodology to filter out the measurement error variability by incorporating the measurement error variance in the spatial covariance structure of the model. The study was carried out in the Lower Hunter Valley, New South Wales, Australia where a combination of laboratory measured, and vis-NIR and MIR inferred topsoil and subsoil soil carbon data are available. We investigated the applicability of residual maximum likelihood (REML) and Markov Chain Monte Carlo (MCMC) simulation methods to generate parameters of the Matérn covariance function directly from the data in the presence of measurement error. The results revealed that the measurement error can be effectively filtered-out through the proposed technique. When the measurement error was filtered from the data, the prediction variance almost halved, which ultimately yielded a greater certainty in spatial predictions of soil carbon. Further, the MCMC technique was successfully used to define the posterior distribution of measurement error. This is an important outcome, as the MCMC technique can be used to estimate the measurement error if it is not explicitly quantified. Although this study dealt with soil carbon data, this method is amenable for filtering the measurement error of any kind of continuous spatial environmental data.

* Corresponding author.
E-mail address: sanjeewani.pallegedaradewage@sydney.edu.au (P.D.S.N. Somarathna).

https://doi.org/10.1016/j.scitotenv.2018.02.302
0048-9697/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Soil carbon is recognized as a variable central to soil fertility and agricultural productivity. It is also well known for its capacity to serve as a store for atmospheric carbon. Transferring atmospheric CO_2 into long-lived pools and securely storing so that it is not immediately remitted is known as carbon sequestration (Lal, 2004; Yigini and Panagos, 2016). Small increases in soil carbon stocks per unit land area are anticipated to result in significant changes in climate and land use management (Falloon and Betts, 2010). Understanding soil carbon processes for implementing "best practice" for balancing carbon budgets is pivotal for carbon sequestration programs (Dawson and Smith, 2007). These programs need extensive sampling for auditing soil carbon stocks. Similarly, the assessment of soil health would also require conducting extensive measurement of soil carbon.

With the growing need for detailed soil carbon data, existing soil carbon maps and inventories are becoming inadequate, especially for large scale projects (Stevens et al., 2013). Standard techniques of soil carbon measurements such as dry combustion and oxidation analyses can be tedious, time consuming and expensive (Nocita et al., 2014). Conversely, infrared spectroscopy has been demonstrated to be a near comparable measurement technique that has the added advantage of being relatively low cost (Janik et al., 2007; Reeves III, 2010; Rossel and Webster, 2012; Stevens et al., 2013; Viscarra Rossel et al., 2006). The low cost associated with this technique means that mapping studies can afford higher sampling densities, thus enabling a detailed understanding soil carbon spatial variation across landscapes.

The use of infrared spectroscopy for soil analysis has been thriving over the past decade (Bellon-Maurel and McBratney, 2011). These studies have mostly focused on predicting basic soil composition, particularly soil organic carbon (SOC) and texture (Stenberg et al., 2010). Bellon-Maurel and McBratney (2011) provide a detailed review of the studies on the use of NIR and MIR spectroscopic studies for soil carbon inference. The review showed that these soil spectral inference studies are largely dedicated to predicting soil carbon content for point locations. However, it is proposed that these soil spectral inference studies could be further expanded into a spatial context for *optimally* predicting soil carbon content at unsampled locations, and ultimately for soil mapping purposes.

Infrared spectroscopic soil carbon measurement is an indirect mode of measurement. The carbon concentrations are inferred using calibration models based on the characteristics of the absorption spectrum of scanned soil samples. One drawback of using these soil carbon data is the comparatively larger measurement error associated with calibration models compared to the data acquired through standard dry combustion techniques (Bellon-Maurel et al., 2010).

When predicting the soil carbon content spatially, we are interested in the actual value rather than the value distorted by the measurement error. More often than not, measurement error is disregarded. For example, a recent study by Rial et al. (2017) mapped topsoil organic carbon content using Visible-Near Infrared (VNIR) spectroscopic measurements without accommodating within the methodology a procedure for handling the measurement errors in the data.

To achieve an optimal prediction in a spatial modelling exercise, the measurement errors should be filtered out (Cressie, 1991). One way of accounting for the measurement error is to include measurement error variance (σ_ε^2) in the variogram or covariance structure of the spatial model. This is also known as kriging with uncertain data, where the error variance is added to the diagonal of the spatial covariance matrix (Delhomme, 1978; Knotters et al., 1995; Laslett and McBratney, 1990). This filters the measurement error variance from the nugget component of the experimental variogram, ultimately leading to lower uncertainty of spatial predictions.

The accuracy of the spatial predictions can also be influenced by the techniques of model parameter estimation. Conventional techniques using method-of-moments can be biased (Lark et al., 2006), and thus the Residual Maximum Likelihood Method (REML) and Bayesian inference from Markov Chain Monte Carlo (MCMC) analysis are the established techniques for unbiased parameter estimation (Poggio et al., 2016). Lark et al. (2006) used REML for estimating parameters of the covariance function directly from the data, and then the estimated parameters were used for the spatial prediction in what is termed as an empirical best linear unbiased predictor (EBLUP). MCMC simulation can also be applied for estimating the variogram and trend model parameters directly from data. Minasny et al. (2011) advocated the use of MCMC simulation for parameter inference in model-based soil geostatistics including the spatial prediction of soil carbon. The basic advantage of MCMC over REML is that MCMC estimates the underlying uncertainty of the parameters, whereas REML relies on a single realisation of the variogram parameters. However, MCMC estimations are computationally expensive compared to the REML approach due to the slow convergence rates of the former (Mossel and Vigoda, 2006; Poggio et al., 2016).

In this study, we explored the applicability of REML-EBLUP and MCMC simulation for measurement error parameter inference for soil carbon spatial modelling. A combination of laboratory measured (dry combustion), near infrared red (NIR) and mid infrared (MIR) spectra estimated soil carbon data and associated σ_ε^2 were used for predicting soil carbon content across the Hunter Valley region, NSW, Australia. Subsequently, we compared the prediction capability of each model, i.e. incorporating σ_ε^2, and without σ_ε^2.

2. Theoretical context

The stochastic spatial process of soil carbon can be expressed by a linear mixed model.

$$\mathbf{S} = \mathbf{M}\beta + \mathbf{W}\mathbf{u} + \mathbf{e} \tag{1}$$

S is the vector of n observations, **M** is the $n \times p$ design matrix that associates with each value of p fixed effects, and β is the vector of p fixed effect coefficients. **u** is the vector of q random effects, realisations of variable **u**, which is associated with the n observations by an $n \times q$ design matrix **W**. It is assumed that **u** is the spatially dependent random variable, while **e** independent random errors and **u** and **e** are independent to each other. Hence, assuming **u** and **e** are jointly Gaussian,

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 \xi \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I} \end{bmatrix} \right) \tag{2}$$

where σ^2 is the variance of the independent error, ξ is the variance ratio between **u** and σ^2 and **G** is the correlation matrix of **u**. **e** represents both measurement errors and the short scale variations of the spatial process which is geo-statistically known as the nugget effect. Assuming **u** is drawn from second order stationary random process, **G** can be characterised by a suitable covariance function since it only depends on the relative locations of the observations (Lark et al., 2006).

The Matérn covariance function has been effectively used in soil science (Minasny and McBratney, 2005) to model the covariance structure of the random effects. The Matérn covariance function (**K**) is given as,

$$K_{ij} = c_0 \delta_{ij} + c_1 \left[\frac{1}{2^{\upsilon-1} \Gamma(\upsilon)} \left(\frac{h}{r} \right)^{\upsilon} K_\upsilon \left(\frac{h}{r} \right) \right] \tag{3}$$

where K_{ij} is the covariance between observation i and j, h represents the separation distance between i and j, δ_{ij}denotes the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ when $I \neq j$), $c_0 + c_1$ signifies the sill variance, K_v is the modified Bessel function of the second kind of order υ. Γ is the gamma function, r denotes the distance or 'range' parameter and υ is the spatial 'smoothness'. The latter parameter allows greater flexibility in modelling the local spatial covariance. The parameters of the covariance function along with σ^2 and ξ can be estimated using REML. This counters the dependence of the estimates on the fixed effects β which are the "nuisance" parameters in this spatial problem (Lark et al., 2006).

2.1. Incorporating measurement error in the spatial model structure

The spectroscopic soil carbon estimates $\mathbf{S}(x_i)$ are usually associated with measurement errors and different to the true or actual values $\mathbf{a}(x_i)$. The data model for spatial random process can also be written as,

$$\mathbf{S}(x_i) = \mathbf{a}(x_i) + \mathbf{e}(x_i) \tag{4}$$

where, the process model $\mathbf{a}(x_i) = \mathbf{M\beta} + \mathbf{Wu}$.

Therefore, measurement error variability can be acknowledged through the inclusion of the underlying spatial correlation process $\mathbf{u}$.

$$K_s(h) = K_a(h) + \sigma_\varepsilon^2 \mathbf{I}(h = 0) \tag{5.1}$$

where h is the separation distance between the observation points. I is a binary function where I = 1 when $h = 0$ and otherwise I = 0. Then, Eq. (5.1) can be elaborated to include σ_ε^2 in the covariance structure of the spatial model.

$$K_s(h) = \begin{bmatrix} K_a(h) + \sigma_\varepsilon^2 & K_a(h) \\ K_a(h) & K_a(h) + \sigma_\varepsilon^2 \end{bmatrix} \tag{5.2}$$

Eq. (5.1) enables the predictor to "filter out" the measurement error from the data. Then the predictions at sampled locations will be measurement error free estimates of the data (Eq. (6)).

$$\mathbf{a}(x_i) = \mathbf{m}(x_i)\hat{\boldsymbol{\beta}} + \mathbf{k}(x_i)^\mathbf{T}\mathbf{K}^{-\mathbf{1}}\left(\mathbf{S}{-}\mathbf{M}\boldsymbol{\beta}\right) \tag{6}$$

Once this covariance is determined, it can be used to predict the values at un-sampled locations x_0 . The ordinary kriging predictor can be written as (Cressie and Wikle, 2011).

$$\mathbf{a}(x_0) = \mathbf{m}(x_0)\hat{\boldsymbol{\beta}} + \mathbf{k}(x_0)^\mathbf{T}\mathbf{K}^{-\mathbf{1}}\left(\mathbf{S}{-}\mathbf{M}\hat{\boldsymbol{\beta}}\right) \tag{7}$$

$\mathbf{a}(x_0)$ is the vector of predicted soil carbon at N un-sampled locations $\mathbf{m}$ is the N x p design matrix with p fixed effects and $\mathbf{k}$ is the covariance matrix between x_i and x_0.

And the prediction variance given by:

$$\tau^2 = \boldsymbol{K}_a(x_0, x_0){-}\boldsymbol{k}_a(x_0)^T\mathbf{K}^{-\mathbf{1}}\boldsymbol{k}_a(x_0) \tag{8}$$

Hence, when predicting at sampled locations with known σ_ε^2, the proposed method yields smoother, error free predictions with lower prediction variance. When predicting at un-sampled locations, predictions will be identical whether measurement error is acknowledged or not. However, prediction variance will be higher when the measurement error is ignored. If all data points share a common σ_ε^2, the prediction variance will be lowered by an amount of σ_ε^2 when the errors are accounted for. Since our soil carbon data are sourced from three different sources, the respective prediction variance will not be lowered by exactly σ_ε^2. This clearly shows how the uncertainty of predictions reduces when the measurement error is accounted for. A more detailed theoretical explanation about filtering the measurement errors of spatial data using aforementioned technique can be found in Cressie (1991) and Cressie and Wikle (2011).

Filtered kriging or FK (Cressie, 1991; Schabenberger and Gotway, 2017; Waller and Gotway, 2004) involves a similar technique of filtering out the measurement error to achieve noise free predictions. Filtered kriging requires a known σ_ε^2 which is common across the spatial locations. Christensen (2011) proposed a heterogeneous filtered kriging (HFK) approach to address the heterogeneity of measurement errors in FK. The main difference between the LMM and FK is, LMM includes σ_ε^2 in the covariance structure while in FK, σ_ε^2 is included in the variogram of the kriging system. Although these parameter estimation techniques are different, both methods ultimately yield similar results. Table 1 compares these two methods.

3. Methods

Fig. 1 shows a flow diagram of the methodology. Each step will be discussed in the following sections.

3.1. Study area

The study area is situated in the Lower Hunter Valley, NSW, Australia. The area is known specifically as the Hunter Wine Country Private Irrigation District (HWCPID). The district has an area of approximately 220 km2. The area experiences a temperate climate, with warm humid summers and relatively cool winters. It receives a uniformly distributed rainfall with an average annual amount of 740 mm. The

Table 1
Comparison of LMM and HFK.

	LMM	Kriging
Model	$\mathbf{S}(x_i) = \mathbf{M\beta} + \mathbf{Wu} + \mathbf{e}(x_i) =$ **S** is the vector of *n* observations, **M** is the *n x p* design matrix that associates with each value of *p* fixed effects, and $\boldsymbol{\beta}$ is vector of *p* fixed effect coefficients. **u** is the vector of q random effects which is associated with the n observations by *n x q* design matrix **W**. $\mathbf{u} \sim N(0,\boldsymbol{G})$, $\mathbf{e}{\sim}N(0,\sigma^2)$ and $cov(\ \mathbf{u}, \mathbf{e}){=}0$ Where G is the covariance matrix of u and σ^2 is the variance of **e**	$S(x_i) = a(x_i) + e(x_i)$ $S(x_i)$ is the observed spatial process, $S(a_i)$ is unobservable spatial process. $e(a_i)$ is the error term as in LMM
Prediction model	Maximised joint distribution of s and u gives $\mathbf{C}\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{Ms} \\ \mathbf{W}^\mathbf{T}\ \mathbf{s}\end{bmatrix}$ Where $C = \begin{bmatrix}\mathbf{M}^\mathbf{T}\mathbf{M} & \mathbf{M}^\mathbf{T}\mathbf{W} \\ \mathbf{W}^\mathbf{T}\mathbf{M} & \mathbf{W}^\mathbf{T}\mathbf{W}{+}\xi^{-\mathbf{1}}\mathbf{G}^{-\mathbf{1}}\end{bmatrix}$ ξ^{-1} is the ratio of variance of **u** to σ^2	$\begin{bmatrix}\lambda \\ \mu\end{bmatrix} = \begin{bmatrix}\hat{\partial}_S & 1 \\ 1' & 0\end{bmatrix}\begin{bmatrix}\dot{\gamma}_{s,a((x_0)} \\ 1\end{bmatrix}$ Where 1 is an *n*- vector of ones, the (I, j) element of the *n x n* matrix $\hat{\partial}_S$ is $(\hat{\gamma}_s(x_i{-}x_j), \dot{\gamma}_{s,a((x_0)}$ is the cross semivariogram of s with $a(x_0)$ error- free unobservable value and μ is the Lagrange multiplier
Structure of the Random effect corrected for the σ_ε^2	Covariance function $\boldsymbol{K}_s = \boldsymbol{K}_a + \sigma_\varepsilon^2\,\mathbf{I}, h = 0$	Variogram $\dot{\gamma}_{s,a((x_0)} = \hat{\gamma}_s - \frac{1}{n}\sum_{i=1}^n \sigma_\varepsilon^2(s_l), h{\neq}0$
Prediction error variance	$\tau^2 = K_a(x_0,x_0) - \boldsymbol{k}_a(x_0)^T\boldsymbol{K}^{-1}\boldsymbol{k}_a(x_0)$ Where $\boldsymbol{k_a}(x_0) =\ cov\ (a(x_0), a(x_i))$ and	$\tau^2 = \lambda^T\dot{\gamma}_{s,a((x_0)} + \mu$
Parameter estimation	ML, REML or MCMC on the data, maximising a loglikelihood function (Cressie and Wikle, 2011)	Method of moments calculation for variogram and fitting via nonlinear least squares (Christensen, 2011)

HWCPID has an undulating topography with hills ascending to the South-West (Fig. 2a). The underlying geology is comprised of predominantly Early Permian siltstones, marl and some minor sandstone and Late Permian siltstones, Middle Permian conglomerates, sandstones and siltstones in minor amounts (Hawley et al., 1995; Malone et al., 2016). The HWCPID is mainly occupied by viticultural enterprises, followed by dry land grazing systems.

3.2. Data

The soil carbon data consists of 'data' collected between the years of 2001 and 2015 during the annual soil surveys carried out by the students of the University of Sydney soil sciences group, and the data from Malone et al. (2011) and Odgers et al. (2011). The term 'soil carbon' in this research is analogous to the total carbon content in the soil. Soil samples were collected from the topsoil (0–10 cm) and subsoil (40–50 cm). As the soil carbon data were derived from several years of surveys, the method of soil carbon measurement also varied. Methods included laboratory analysis using dry combustion method, and spectrally inferred from NIR and MIR diffuse reflectance measurements. Dry combustion of the soil samples was done using an ElementarVario Max CNS macro elemental analyser (Elementar Analysesysteme GmbH, Hanau, Germany) where the carbon content is determined by the loss on ignition at 400 °C (Zobeck et al., 2013). The standard deviation of the soil carbon measurement of the ElementarVario Max CNS analyser is 0.001–0.004 g 100 g^{-1} based on standard soil samples.

The estimation of soil carbon content using infrared spectroscopy is done based on the absorption spectrum which is produced after scanning the soil sample with an analytical spectral instrument. The absorption spectrum has a characteristic shape produced based on the constituents of the soil. The spectrum is then used to infer the soil carbon content via calibration models. NIR spectroscopic measurements were made using an Agrispec portable spectrophotometer with a contact probe attachment (Analytical Spectral Devices, Boulder, Colorado). Bruker TENSOR 37 Fourier Transform (FT) mid-infrared (MIR) spectrometer was used to measure the MIR spectral reflectance of soil samples. The collected NIR/MIR spectra were prepossessed to remove the noise, followed by normalising before using them for the calibrations. Calibration models were derived using a regression tree method called Cubist (Minasny and McBratney, 2006; Quinlan, 1992), where spectral data is linked with the soil carbon content measured via the dry combustion method. The calibration data came from a library of 316 soil profile samples from the wheatbelt of southern NSW and northern Victoria (Geeves et al., 1995). See Minasny et al. (2008) for the MIR spectra calibration model.

Some sampling points of the study area consisted of more than one type of measurement for soil carbon. Accordingly, some data points contained all three types of measurement: laboratory analysis using dry combustion (C_{ea}), NIR-inferred and MIR-inferred measurements. Some sampling points had only two types of measurements: either C_{ea} and MIR or C_{ea} and NIR, while the rest of the sampling points consisted of only one measurement type.

Altogether, there were 1679 soil samples for the top 0–10 cm soil layer. The observed data consisted of 681 C_{ea} measured values, 266 NIR inferred values and 732 MIR inferred values. There was a total of 1129 samples for the 40–50 cm subsurface layer. There were 43 C_{ea} measured values, 767 NIR inferred values and 319 MIR inferred values. The total carbon content was measured in g 100 g^{-1} of soil (Fig. 2b). A detailed analysis of data emphasizing spatial and temporal distribution of measured carbon is presented in Appendix A.

3.2.1. Data pre-processing

Measured soil carbon concentration data had a skewed distribution for all methods, resulting in the need to transform it via a square root transformation to approximate a normal distribution. Then a linear relationship was developed between the values of standard dry combustion technique (CNS) and NIR, and MIR. These relationships were used for bias correction at sampling points where there were no CNS measurements. Measurement errors for NIR and MIR data were calculated as.

$$\boldsymbol{\varepsilon}_{(\mathrm{NIR/MIR})(x_i)} = s_{Cea(x_i)} - s_{NIR/MIR(x_i)} \tag{9}$$

where $s_{Cea(x_i)}$ is the measured value at a location x and, while $s_{NIR/MIR(x_i)}$ denotes the NIR/MIR measured value at the same location. Then the measurement error variance for NIR/MIR data is given by

$$\sigma^2_{\varepsilon(\mathrm{NIR/MIR})} = \frac{\sum\left\{\left(\boldsymbol{\varepsilon}_{(\mathrm{NIR/MIR})} - \bar{\boldsymbol{\varepsilon}}_{(\mathrm{NIR/MIR})}\right)^2\right\}}{\mathrm{n}} \tag{10}$$

where $\bar{\varepsilon}_{(\mathrm{NIR/MIR})}$ denotes the mean error and n is the number of observations for NIR/MIR data.

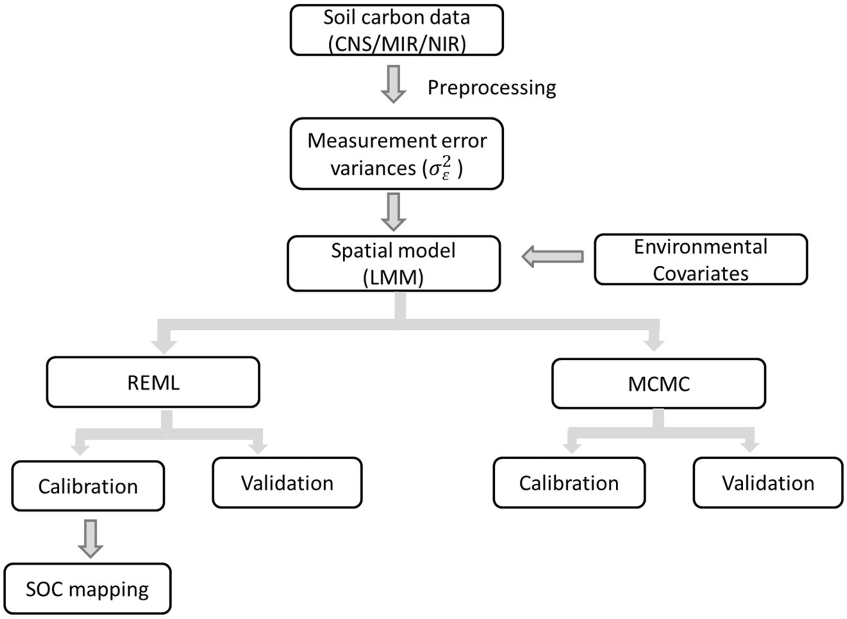


Fig. 1. Methodology flow diagram.

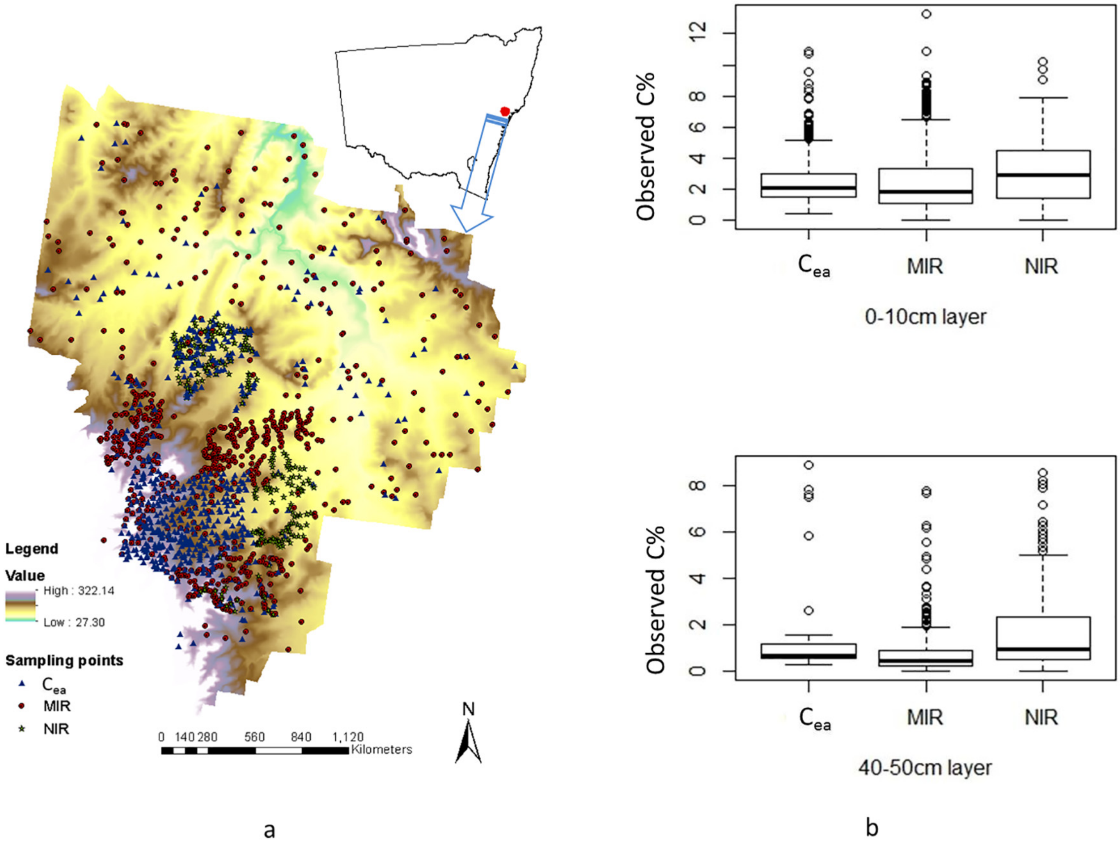


Fig. 2. (a). The spatial distribution of observation points for the 0–10 cm layer. (b). The composition of observation data (C_{ea}) measured using CNS Vario Max and NIR/MIR spectral inference for the two soil layers.

The calculated σ_ε^2 was 0.21 and 0.07 for NIR and MIR respectively for the 0–10 cm layer, while for the 40–50 cm layer the error variance was 0.22 and 0.17 respectively for NIR and MIR instruments.

70% of the data from each layer was randomly selected for model training, with the remaining 30% allocated for validation. To ensure there was no co-located data for a sampling location, we selected the most accurate type of measurement that was available. The ranking for accuracy was 1) dry combustion measurement, 2) MIR spectrally inferred measurement, and 3) NIR spectrally inferred measurement.

3.3. Spatial model of soil carbon

Stepwise regression and correlation coefficients were used to select the most important environmental covariates for 'scorpan' modelling (McBratney et al., 2003) of soil carbon content. Covariate selection was done using a covariate pool which consisted of 15 covariates: Easting, northing, aspect direction, Landsat 5 ETM bands 1,2,3,4,5, and 7, catchment area, Digital Elevation Model (DEM), land cover, light isolation, normalized difference vegetation index (NDVI), plan curvature, profile curvature and slope direction. The selection was done through the analysis of correlation coefficients and the stepwise regression method. Landsat band 5 (NIR band), DEM, and NDVI were statistically significant for soil carbon predictions over the study area. A linear mixed model (LMM) was fitted to the 0–10 cm and 40–50 cm data for estimating the soil carbon content across the study area. The LMM model is given as

$$\mathbf{S}(x_i) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \textit{Landsat band } 5 + \boldsymbol{\beta}_2 \textit{Filled DEM} + \boldsymbol{\beta}_3\ \textit{NDVI} + \mathbf{u} + \mathbf{e}(x_i) \tag{11}$$

where $\mathbf{S}(x_i)$ is the observed value of soil carbon and $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$, are parameters of the fixed effects.

Model calibration was done using REML and Bayesian MCMC models, and the parameters were estimated for both scenarios: inclusion and exclusion of measurement error variance. Then we selected the best performing calibration technique as confirmed by the validation dataset, to implement the spatial prediction of soil carbon at unsampled locations i.e. digital soil mapping.

3.3.1. Estimation of model parameters via REML

In geostatistics, the linear spatial model parameters are usually inferred using the standard ordinary least squares method. Then the variogram structure of the model residuals is estimated separately using the method-of-moments procedure. This method underestimates overall variability, and also the spatial structure estimates maybe similarly inaccurate. REML can be used as a solution to this problem since it provides unbiased and robust estimates of the parameters directly from the data (Minasny and McBratney, 2007).

In the REML approach, an optimization algorithm (Nelder-Mead Simplex method) was used to find the parameters that maximise the following log-likelihood function:

$$L(\boldsymbol{\theta}|\mathbf{y}) = \frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}| - \log|\mathbf{W}| - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{Q}\mathbf{y} \tag{12}$$

where $[\boldsymbol{\theta} = \boldsymbol{\beta} \mid \boldsymbol{\phi}]$ denotes the vector of parameters to be estimated for the linear spatial model,$\boldsymbol{\beta}$ is the linear spatial trend and $\boldsymbol{\phi} = [c0, c1, r, v, \sigma_\varepsilon^2]$ defines parameters of the covariance function. $\mathbf{M}$ is the design matrix of the trend function and $\mathbf{W} = \mathbf{M}^T\mathbf{K}^{-1}\mathbf{M}$, $\mathbf{Q} = \mathbf{I} - \mathbf{M}\mathbf{W}^{-1}\mathbf{M}^T\mathbf{K}^{-1}$ and $\mathbf{y} = \mathbf{TS}$, denote a stationary data increments transformation of $\mathbf{S}$ with transformation matrix $\mathbf{T} = \mathbf{I} - \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$.

Estimation of the REML variogram parameters using the profile likelihood method can be summarised as follows:

- Choose a set of values for v and r
- Maximise the log-likelihood by estimating c_0 *and* c_1 for each combination of v and r using an optimization algorithm.
- Plot the likelihood L, against v and r *for* finding the respective v and r values that have the largest L (log-likelihood) value.

A detailed explanation of REML-EBLUP for Matérn covariance function can be found in Lark and Cullis (2004) and Minasny and McBratney (2007).

Accordingly, REML was used to generate the spatial model parameters for the two soil layers separately. Models for each soil layer were trained with added measurement error variance and without measurement error variance. As explained earlier, the measurement error variance was included in the diagonal of the covariance matrix for the prior approach. REML was applied with initial guesses for c_0 and $c_{1,}$ and then each model was optimised via a profile-likelihood method using a combination of v and r values. The parameter combination that furnished the maximum likelihood (Eq. (12)) was then selected.

3.3.2. Bayesian inference using MCMC simulation

Unlike classical statistics, Bayesian inference treats both parameters (of the statistical model) and the sample data as random, and draws conclusions about the population based on those samples. The Bayes theorem states that the posterior probability, $p(\boldsymbol{\theta}|\overline{S})$ of a hypothesis ($\boldsymbol{\theta}$) is proportional to the product of likelihood $L(\boldsymbol{\theta}|\overline{S})$ and the prior probability, $p(\boldsymbol{\theta})$ of the hypothesis given the new observations $\overline{S}$ or $p(\boldsymbol{\theta}|\overline{S})\alpha\, p(\boldsymbol{\theta})$ $L(\boldsymbol{\theta}|\overline{S})$. In this context, $\boldsymbol{\theta}$ is the spatial model (Vrugt, 2016). Thus, it captures the probability distribution of posterior parameters from the likelihood and probability of the distribution of prior parameters too.

However, for practical geostatistical problems which have a high dimensionality, it is near-impossible to obtain the posterior distribution through analytical means or by analytical approximation. Only recently a new analytical approximation via Integrated Nested Laplace Approximation (INLA) has been proposed (Huang et al., 2017; Poggio et al., 2016). Commonly, iterative approximation methods such as Markov Chain Monte Carlo (MCMC) are used to approximate the target distribution (Minasny et al., 2011). The differential evaluation adaptive metropolis (DREAM), an algorithm proposed by Vrugt et al. (2008) is a multi-chain MCMC simulation algorithm that automatically tunes the scale and orientation of the proposed distribution en route to the target distribution. In addition, DREAM has an efficient sampling strategy on high-dimensional and multi-modal posterior distributions (Vrugt, 2016). DREAM implementation can be summarised as follows:

- Define the parameter ranges, initial sampling distribution and likelihood function to compare model prediction with the observation data.
- List the upper and lower bounds of the parameters and use Latin Hypercube sampling over d dimensional hypercube to initialize the initial points of the N number of Markov chains.
- Optimise the log-likelihood function:

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{S}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}(\mathbf{S}-\mathbf{M}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{S}-\mathbf{M}\boldsymbol{\beta}) \tag{13}$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta} \mid \boldsymbol{\phi}]$ denotes the vector of parameters to be estimated for the linear spatial model as in Eq. (12).

More detailed theoretical and technical explanation about DREAM can be found in Vrugt (2016). Minasny et al. (2011) also provide a detailed discussion about its application in soil geostatistics.

Accordingly, four calibration models were derived using DREAM for the two soil layers for both scenarios; with added σ_ε^2 and without σ_ε^2. We also explored the possibility of defining the posterior distribution of σ_ε^2 of NIR and MIR predictions directly from the data. If the posterior distribution of σ_ε^2 can be successfully defined, MCMC can be used to determine σ_ε^2, when σ_ε^2 is not explicitly quantified.

3.4. Model validation

The calibrated models for each soil layer and for each scenario were validated using a subset of the data. With these data, models were compared using root mean squared error (RMSE) and Lins' concordance correlation coefficient (CCC) (Lin, 1989).

3.4.1. Uncertainty assessment

The predictions from a spatial model are done with a certain degree of uncertainty. Uncertainty can be simply defined as the variability of model predictions. Parameter uncertainty and structural uncertainty are the two major sources of uncertainty of a spatial model. Parameter uncertainty is caused by the uncertainty of model parameters and the structural uncertainty caused by approximate or incomplete treatment of the spatial relationship of the process being modelled (McKay, 1995).

Prediction variance is considered as a measure of uncertainty of model predictions caused by the uncertainty of input parameters of the model. We calculated the variance of the model predictions to see how the models behave when the uncertainty of data is included or excluded. Prediction variance is given by,

$$\sigma_x^2 = \frac{\sum_{i=1}^{n}\left\{\hat{S}(x) - \hat{\mu}(x)\right\}^2}{n} \tag{14}$$

where $\hat{S}(x)$ is the predicted value and $\hat{\mu}(x)$ is the mean of predicted values.

The standardised squared deviation SSD(x) measures a prediction model's goodness of fit. It is an indication of the quality of estimate of the prediction variance. A value closer to 1 for mean SSD(x) indicates a good estimate (Voltz and Webster, 1990) and a median value closer to 0.455 (Lark, 2000) symbolises kriging with a correct variogram:

$$SSD(x) = \frac{\left\{S(x) - \hat{S}(x)\right\}^2}{\sigma_x^2} \tag{15}$$

where $S(x)$ is the measured value, $\hat{S}(x)$ denotes the predicted value with variance σ_x^2.

In the REML approach, prediction variance at each prediction point was calculated along with the predicted value. Similarly, for MCMC approach, prediction variance was calculated at each prediction point given the realisations of the last 1000 MCMC simulations, and the predicted value was taken as the average of all simulated predicted values.

3.5. Mapping the carbon content

Finally the best performing REML–inferred models were selected to predict the carbon content to a 25 × 25 m grid over study area. The maps were produced using trained models for with (a) and without (b) the inclusion of σ_ε^2 scenarios for both 0–10 cm and 40–50 cm soil layers. The spatial predictions for these two scenarios were subsequently compared by subtracting (b) from (a). Also the prediction uncertainties of two scenarios were compared in a similar manner.

4. Results & discussion

This study focuses on filtering measurement errors or observed uncertainty of data via expressing the estimated uncertainty in the spatial model. The study compares the accuracy of a spatial model when the measurement error variance is included and when it is excluded. In any research field, measured data is affected by some degree of uncertainty. For example ecological data are almost always observed

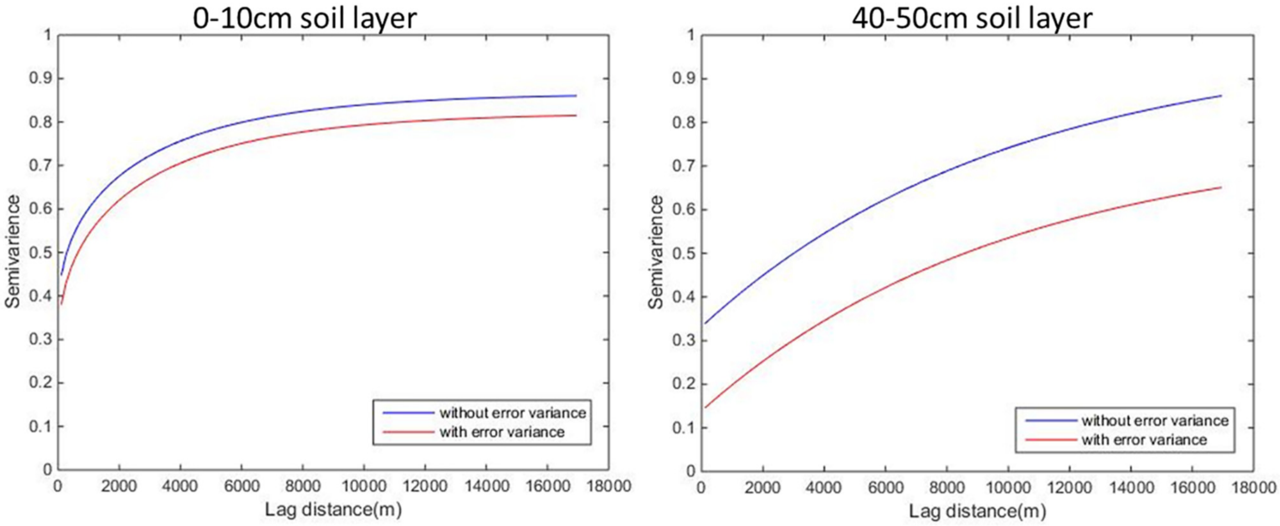


Fig. 3. REML estimated variograms for the scenarios with and without inclusion of σ_ε^2 for the two soil layers.

incompletely with large and unknown amounts of measurement error or data uncertainty (Cressie et al., 2009).

4.1. REML inferred variogram parameters

In this study, the REML approach was used to find the optimum values that maximise the log-likelihood (Eq. (12)). We compared the aforementioned scenarios; with σ_ε^2 and without σ_ε^2 using derived variogram structures. For the top 0–10 cm layer, among the thirty six combinations of v and r values, 0.05 was the optimum v while the range was around 2500 to 5000 m (Fig. 3). In the modelling exercise without the inclusion of σ_ε^2, the optimum range is slightly reduced to 1500 m while the smoothness parameter (v) increased to 0.1. Very low v values and short ranges suggest that the soil carbon process at the study site is a highly variable spatial process. In other words, a presence of rapid variations of soil carbon content at small lags can be observed. With the inclusion of σ_ε^2 the value of v has doubled suggesting the addition of σ_ε^2 increased the smoothness of the spatial process of soil carbon.

The variogram parameters of the 40–50 cm layer had a similar optimum ($v = 0.2$) and a similar range of 2000–4000 m for both scenarios. The smaller v value also indicated that a rapid variation in soil carbon content occurs at small lags. However, the variation is comparatively low compared to that of the top 0–10 cm layer. Variograms of the two scenarios for both soil layers show a similar structure apart from the different sill values. The total variance is reduced with the added σ_ε^2 for both layers, and the nugget (c_0) value decreased with the inclusion of σ_ε^2 for both soil layers. This is mainly due to the exclusion of σ_ε^2 in

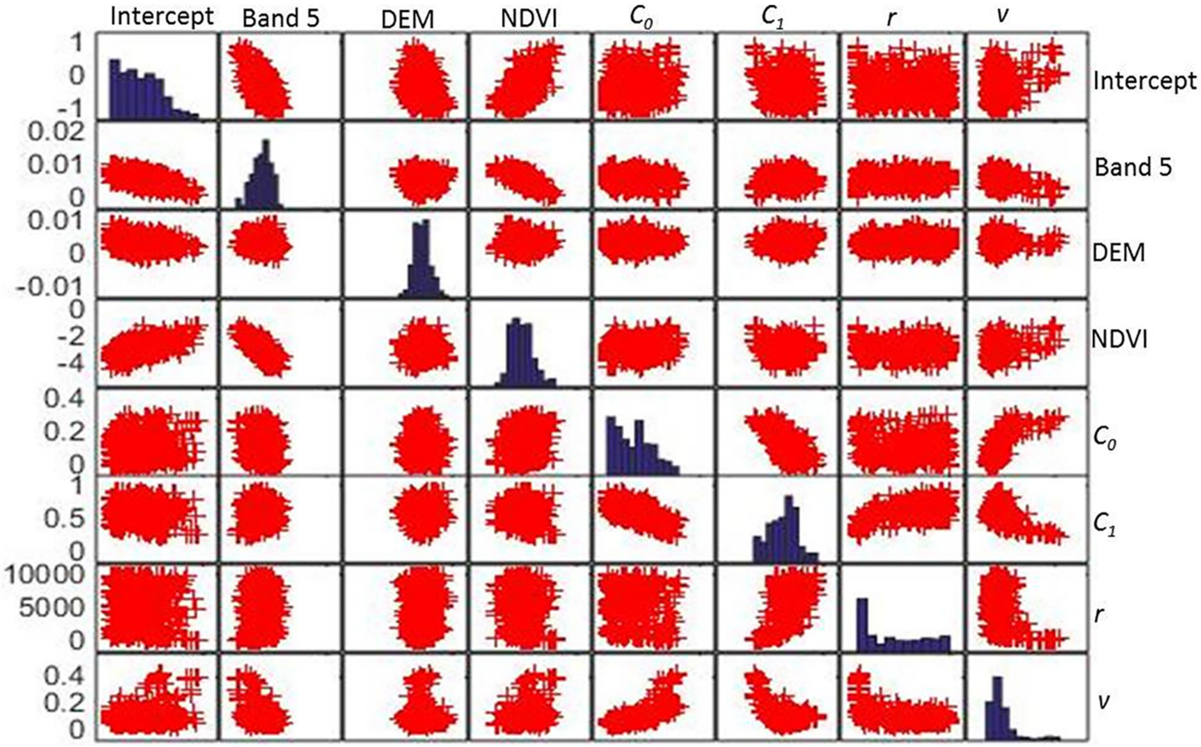


Fig. 4. Marginal distributions (diagonal) and binary scatter plots (off-diagonal) of posterior parameters for 0–10 cm soil carbon content.

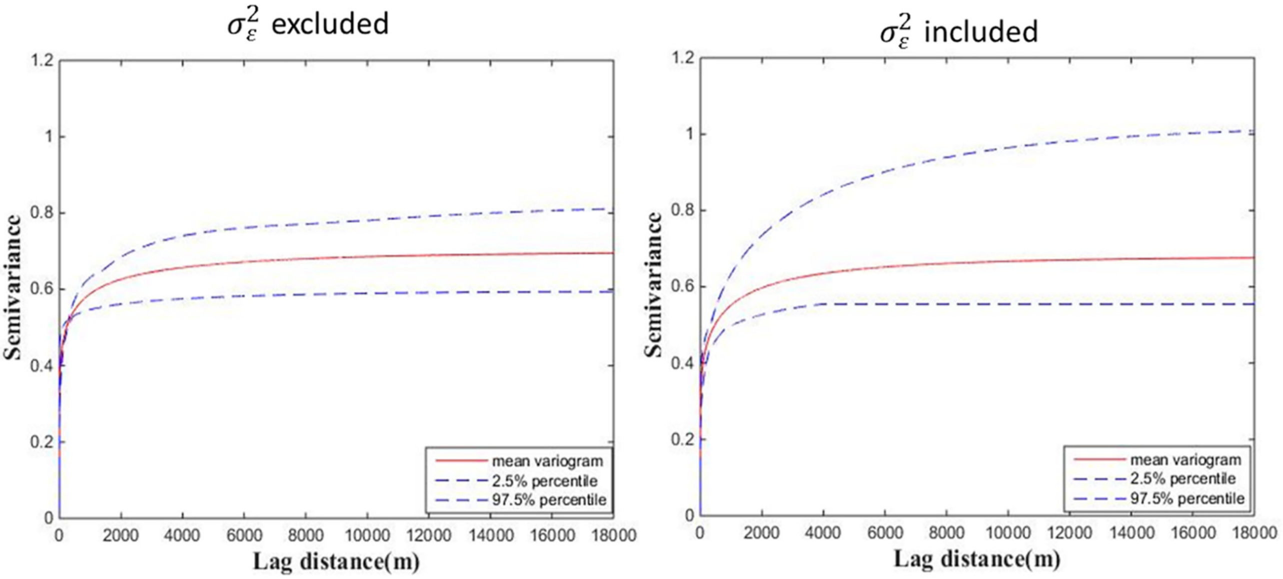


Fig. 5. MCMC estimated variograms without and with inclusion of σ_ε^2 for 0–10 cm soil layer with uncertainty levels.

modelling the variogram structure. Clark (2010) also found that the apparent estimation of variance can be significantly reduced by acknowledging the measurement error.

4.2. MCMC parameter estimation

The marginal distributions of Matérn variogram parameters along with the linear trend model parameters, including the two-dimensional scatter plots of posterior samples are illustrated in Fig. 4. The probability distribution functions (pdfs) of the linear model parameters appear to be well defined, and approximate a Gaussian distribution. There is a significant correlation between the intercept and the parameters of covariates; Band 5 and NDVI.

The near normal marginal distribution of c_0 and c_1 of variogram parameters appear to be well defined. However, the pdf of r extended over the entire prior range, implying that the pdf of r is poorly defined. Also, there were significant scatter correlations between variogram parameters. c_0 and c_1 were positively correlated, and the smoothness parameter (v) was positively correlated with the nugget (c_0) while r and c_1 were also positively correlated. These correlations between variogram parameters add significant uncertainty to variogram parameters.

The modelling with σ_ε^2 displayed similar marginal distributions of model parameters while the range was poorly defined for the topsoil layer. Correlation between trend models and variogram parameters for both soil layers displayed a similar pattern for both scenarios of testing; with and without σ_ε^2.

Variogram structures from both scenarios were well defined by the MCMC simulations. Similar to the REML approach, modelling with σ_ε^2 reduced the total variation of the variogram and the nugget value (Fig. 5). Unlike REML, the uncertanity of the variogram models

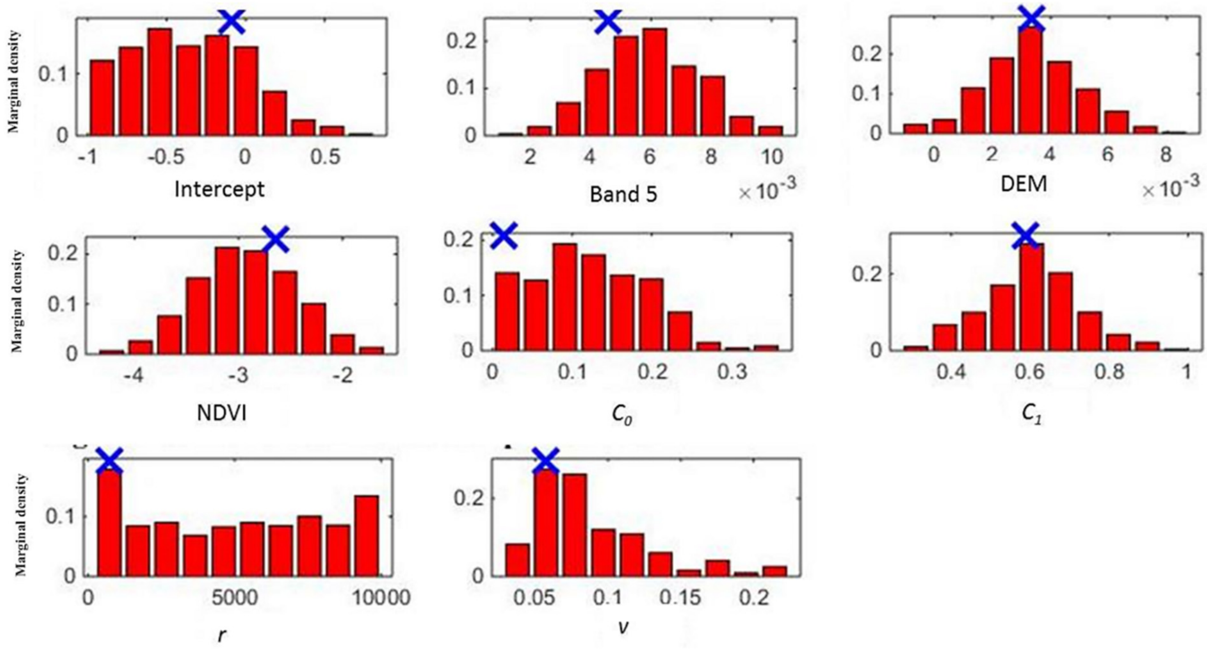


Fig. 6. Marginal distributions of posterior parameters for 40–50 cm soil carbon content without added σ_ε^2. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

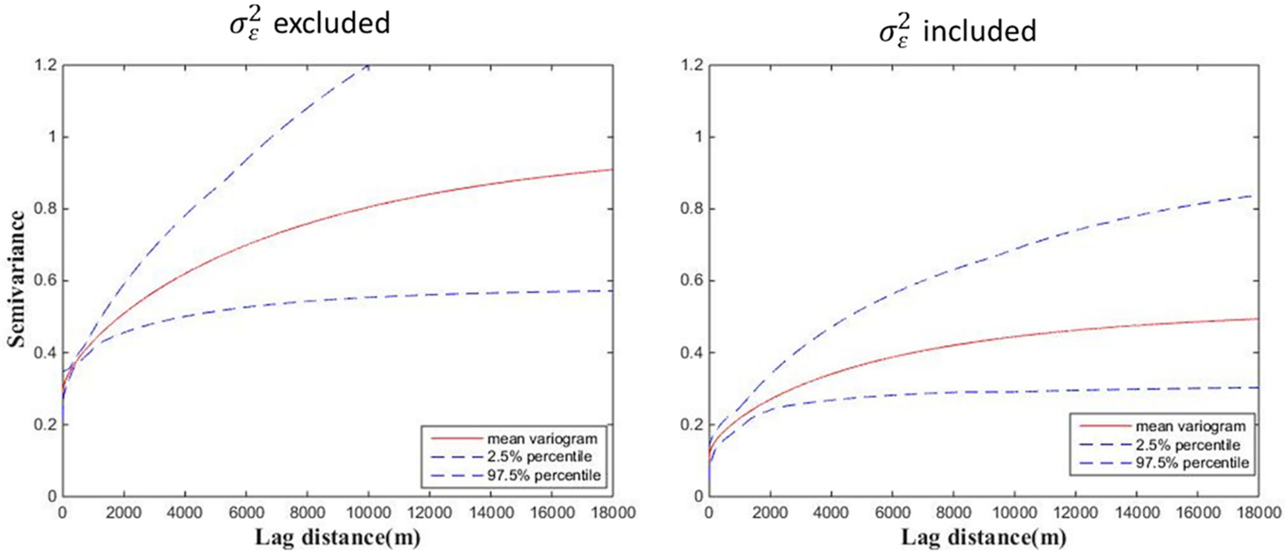


Fig. 7. MCMC estimated variograms without and with inclusion of σ_ε^2 for 40–50 cm soil layer with uncertainty levels.

could be estimated through the MCMC simulation which are expressed by 2.5 and 97.5 percentile values of the simulated distributions in Tables 2a and 2b. When the mean parameter values from MCMC were compared with the REML values, the total variation was reduced.

Fig. 6 presents the histograms of the marginal posterior distributions of the linear and variogram models' parameters for the 40–50 cm layer. The maximum a-posteriori probability (MAP) of each model parameter is depicted by the (blue) cross in each histogram. These values are the mode of the posterior distribution of each parameter.

As per the top layer, the linear model parameters appeared rather well defined and exhibit approximate Gaussian distribution. However, the posterior variogram parameters extend through the entire range of the prior parameters, which is indicative of the variogram parameters being not well defined. This relatively large parameter uncertainty can lead to unrealistically large prediction uncertainties. The presence of outliers along with the correlations between variogram parameters can cause this type of behaviour, and is also seen in the exponentially increasing 97.5 percentile of the variogram (Fig. 7a). Noticeably, the sill value is reduced by approximately 50%,

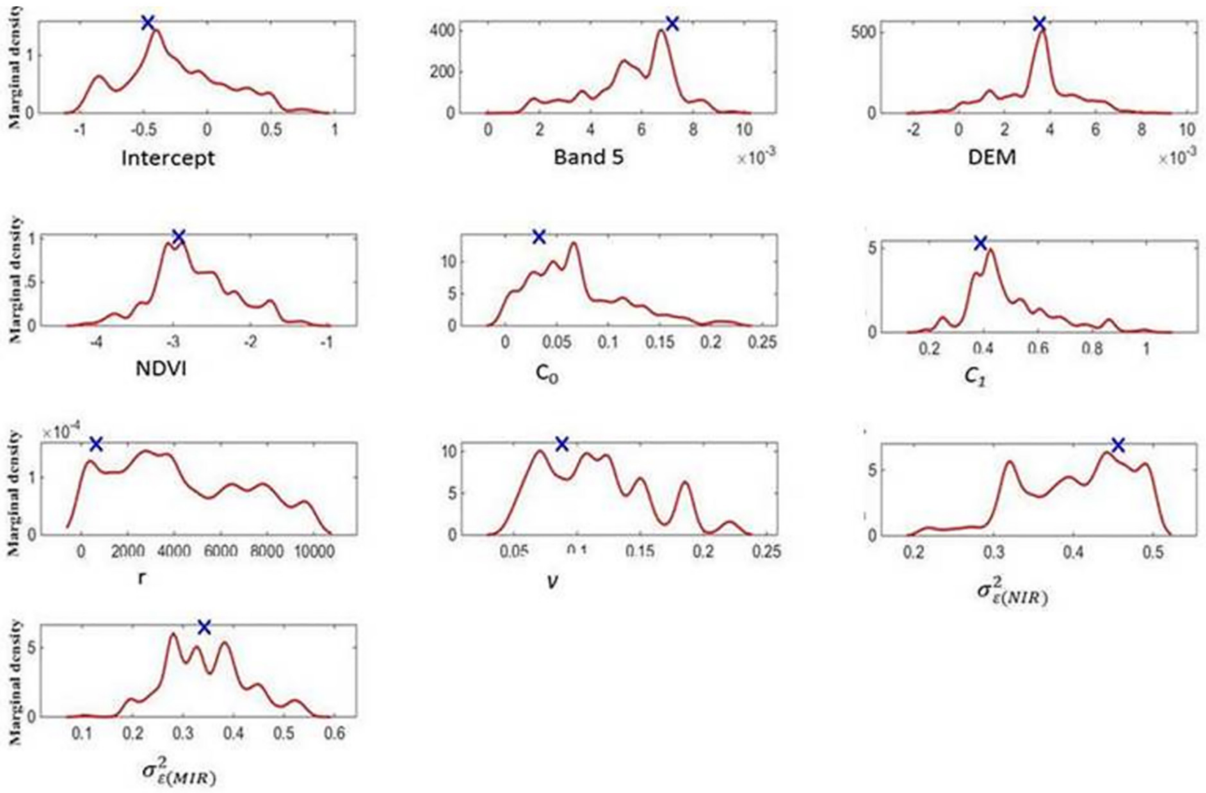


Fig. 8. Empirical probability density functions of posterior parameters for 0–10 cm soil carbon content. (Blue) the cross indicates the MAP value of each parameter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2a
Posterior parameters of the linear spatial model and the variogram models for 0–10 cm layer. β represents coefficients of the linear trend. c_0, C_1, and r are the variogram parameters; nugget, sill, and range respectively. v is the smoothness parameter of the covariance function.

Method	β_0	β_1	β_2	β_3	β_4	C_0	C_1	r	v
REML	-0.488	0.006	0.003	0.004	-3.159	0.340	0.554	5000.0	0.200
REML $+\sigma_\varepsilon^2$	-0.482	0.006	0.004	0.003	-2.960	0.267	0.526	5000.0	0.200
MCMC- 2.5 percentile	-0.985	0.002	0.001	-0.014	-3.896	0.015	0.319	659.4	0.021
MCMC mean	-0.427	0.006	0.004	0.010	-2.916	0.143	0.556	5030.2	0.091
MCMC- 97.5percentile	0.362	0.008	0.006	0.027	-1.609	0.280	0.721	9833.9	0.174
MCMC$+\sigma_\varepsilon^2$ -2.5 percentile	-0.921	0.005	0.002	-0.031	-3.583	0.007	0.251	572.9	0.054
MCMC$+\sigma_\varepsilon^2$- mean	-0.302	0.006	0.004	0.008	-2.813	0.104	0.623	5041.5	0.095
MCMC$+\sigma_\varepsilon^2$-97.5 percentile	-0.157	0.010	0.006	0.027	-2.444	0.342	0.979	9177.4	0.248

Table 2b
Posterior parameters of the linear spatial model and the variogram models for 40–50 cm layer. β represents coefficients of the linear trend. C_0, C_1, and r are the variogram parameters; nugget, sill, and range respectively. v is the smoothness parameter of the covariance function.

Method	β_0	β_1	β_2	β_4	C_0	C_1	r	v
REML	0.563	-0.001	0.003	-0.420	0.329	0.370	5000.0	0.500
REML $+\sigma_\varepsilon^2$	0.599	-0.002	0.003	-0.319	0.137	0.358	5000.0	0.500
MCMC 2.5-percentile	-0.778	-0.003	0.000	0.985	0.216	0.296	1964.4	0.169
MCMC- mean	0.359	-0.001	0.003	-0.315	0.289	0.694	6616.8	0.380
MCMC- 97.5percentile	0.910	0.002	0.060	0.224	0.389	1.443	9778.3	0.690
MCMC$+\sigma_\varepsilon^2$ 2.5-percentile	-0.559	-0.004	0.000	1.478	0.031	0.159	1294.7	0.150
MCMC$+\sigma_\varepsilon^2$ mean	0.344	-0.001	0.004	0.516	0.111	0.475	5786.7	0.425
MCMC$+\sigma_\varepsilon^2$ 97.5 percentile	-0.962	0.002	0.007	0.379	0.163	1.419	9822.8	0.929

and the uncertainty levels are considerably reduced with the inclusion of σ_ε^2 (Fig. 7b).

4.3. Estimating error variance through MCMC simulation

The possibility of defining the pdf of error variance parameters is also examined. The empirical probability density functions of the posterior parameters including the error variance parameters $\sigma^2_{\varepsilon(MIR)}$ and $\sigma^2_{\varepsilon(NIR)}$ for topsoil carbon content are given in Fig. 8. The marginal distribution of $\sigma^2_{\varepsilon(MIR)}$was relatively well defined here, with MAP values of 0.35 and 0.47 for the top layer and 0.14 and 0.23 for the bottom layer for $\sigma^2_{\varepsilon(MIR)}$and $\sigma^2_{\varepsilon(NIR)}$ respectively. The respective measured values were 0.07, 0.21 and 0.17, 0.22. These results indicate that there is a possibility of estimating the σ_ε^2 through the MCMC simulation.

Tables 2a and 2b illustrate the posterior parameters of the linear trend model and the variogram parameters. The MCMC derived posterior parameters display a significant uncertainty. It is difficult to compare the optimised parameters of each model since they are derived from different likelihood functions. However, the optimised linear model parameters of REML and the mean values of the MCMC simulations were quite similar between the scenarios for the top soil layer (Table 2a). The variogram parameters differed between the scenarios and between the models. There is a considerable difference in the c_0 and v of the optimised variogram parameters while the range values of all models were quite similar. With the inclusion of σ_ε^2, the c_o value was reduced due to filtering out σ_ε^2 from the nugget.

The optimised linear model parameters from REML and the means of the MCMC simulations were quite similar for all models tested for the 40–50 cm layer (Table 2b). The c_1 variogram parameter values of all models were quite similar and the other parameters c_0, r, v were significantly different between models. The spatial range parameter of the MCMC approach was almost half of the REML derived estimate.

4.4. Comparing model performance

4.4.1. Prediction accuracy

Tables 3a and 3b summarises the validation results of all models for the 0–10 cm and 40–50 cm layers. The MCMC results represent the averages of 1000 simulations from the last 1000 MCMC parameter realisations. The RMSE and CCC values for all models were quite similar indicating that the differences between the prediction accuracy of the models were insignificant. Although the values slightly differ between the two spectral inferencing techniques, the RMSE and CCC values stayed the same, confirming the predictions from both scenarios were almost identical. The same statistics were calculated separately for samples which were measured using the lab-based measurements and spectrally inferred soil carbon values. The accuracy of all models was comparatively higher for the measured soil carbon (C_{ea}) than the NIR/MIR inferred carbon content.

When we consider the 40–50 cm validation results, overall the conclusions are similar to the top 0–10 cm layer. However, the CCC values for all modelling scenarios of the bottom layer were greater than the respective values of the upper layer. Usually the accuracy of carbon predictions decreases with increasing depth. However, this study produced contradictory results. This can be due to the environment in which the study is situated. In the study site, the sub-soil variation is driven by the presence of marl (loose, earthy deposits consisting chiefly of an intimate mixture of clay and calcium carbonate) rather than the

Table 3a
Comparison of validation results of REML-BLUP, MCMC methods with and without σ_ε^2 for 0–10 cm layer. C_{ea} is the measured value using dry combustion.

Test Statistics	REML-EBLUP	REML-EBLUP $+\ \sigma_\varepsilon^2$	MCMC	MCMC $+\sigma_\varepsilon^2$
RMSE-all (C_{ea},NIR,MIR)	0.20	0.20	0.19	0.19
RMSE- C_{ea}	0.05	0.05	0.11	0.11
RMSE- NIR and MIR	0.15	0.15	0.25	0.25
CCC-all (C_{ea},NIR,MIR)	0.45	0.47	0.45	0.45
CCC-all - C_{ea}	0.44	0.55	0.46	0.47
CCC-all - NIR/MIR	0.45	0.47	0.45	0.45
Mean prediction variance	0.51	0.25	0.51	0.49
Mean SSD-all (C_{ea},NIR,MIR)	0.39	0.80	0.39	0.40
Median SSD-all (C_{ea},NIR,MIR)	0.10	0.21	0.11	0.11
Mean SSD- C_{ea}	0.22	0.45	0.23	0.23
Median SSD- C_{ea}	0.08	0.16	0.08	0.08
Mean SSD- NIR and MIR	0.50	1.08	0.53	0.52
Median SSD- NIR and MIR	0.15	0.32	0.17	0.16

Table 3b
Comparison of validation results of REML-BLUP, MCMC methods with and without σ_ε^2 for 40–50 cm layer. C_{ea} is the measured value using dry combustion.

Test Statistics	REML-EBLUP	REML-EBLUP $+ \sigma_\varepsilon^2$	MCMC	MCMC $+\sigma_\varepsilon^2$
RMSE-all (C_{ea},NIR,MIR)	0.19	0.19	0.19	0.19
RMSE- C_{ea}	0.19	0.09	0.12	0.11
RMSE- NIR and MIR	0.19	0.19	0.20	0.19
CCC -all (C_{ea},NIR,MIR)	0.63	0.64	0.63	0.60
CCC - C_{ea}	0.87	0.90	0.88	0.85
CCC - NIR/MIR	0.60	0.62	0.61	0.58
Mean prediction variance	0.38	0.19	0.38	0.19
Mean SSD-all (C_{ea},NIR,MIR)	0.51	1.10	0.49	1.29
Median SSD-all (C_{ea},NIR,MIR)	0.17	0.35	0.16	0.42
Mean SSD- C_{ea}	0.32	0.62	0.30	0.92
Median SSD- C_{ea}	0.11	0.20	0.12	0.31
Mean SSD- NIR and MIR	0.51	1.04	0.50	1.30
Median SSD- NIR and MIR	0.17	0.36	0.17	0.44

other environmental factors used in this study. Thus, the carbon content in the subsurface layer appears to be mainly determined by a spatial random process, rather than being a reflection of the environmental covariates. This indicates that the measured soil carbon content is an indirect measure or proxy for the presence of sub-soil marl. Rather usefully, with the models used in this study, once the covariance structure is accurately estimated, the accuracy of the predictions increases. Carbon content in the top soil layer of the study site is closely related to environmental factors which make it difficult to fully capture the soil carbon variability through the deterministic model. There are always unaccounted relationships in the linear trend, and hence the prediction accuracies remain low.

4.4.2. Prediction uncertainty

For both REML and MCMC modelling approaches, the median and mean SSD values greatly improved (almost doubled) when σ_ε^2 was included in the spatial model. This is mainly due to the much lower prediction variance associated with the aforementioned scenario (Tables 3a, 3b). The prediction variance values show that when the measurement error is acknowledged, the prediction uncertainty nearly halved.

We also compared the reliability of the uncertainty of prediction estimates by calculating the percentage of predictions that occupy the prediction range within the defined 95% confidence limits for all tested models. For the REML approach, 96.6% of data were within the 95% confidence range while for the MCMC approach 96.8% was within the CI range. Thus, all approaches display more or less a similar outcome.

4.5. Comparing predictions at un-sampled locations

Since the REML approach produced comparable results to the more technically and computationally expensive MCMC techniques, we selected the former approach to predict the soil carbon content across the study area.

Fig. 9 illustrates the effect of filtering σ_ε^2 from the data. We drew 0.2% contours of the predicted carbon content from both methods to examine the smoothing effect when measurement error is acknowledged. When σ_ε^2 is ignored, the predictions are made using a more continuous variogram leading to intense contouring around the data points. By contrast, the σ_ε^2 filtered map provides more realistic and smoother predictions of soil carbon content.

Fig. 10 depicts the comparison of the spatial prediction of soil carbon over the study area for both modelling scenarios. Fig. 10a represents the soil carbon content when modelling with σ_ε^2. Soil carbon

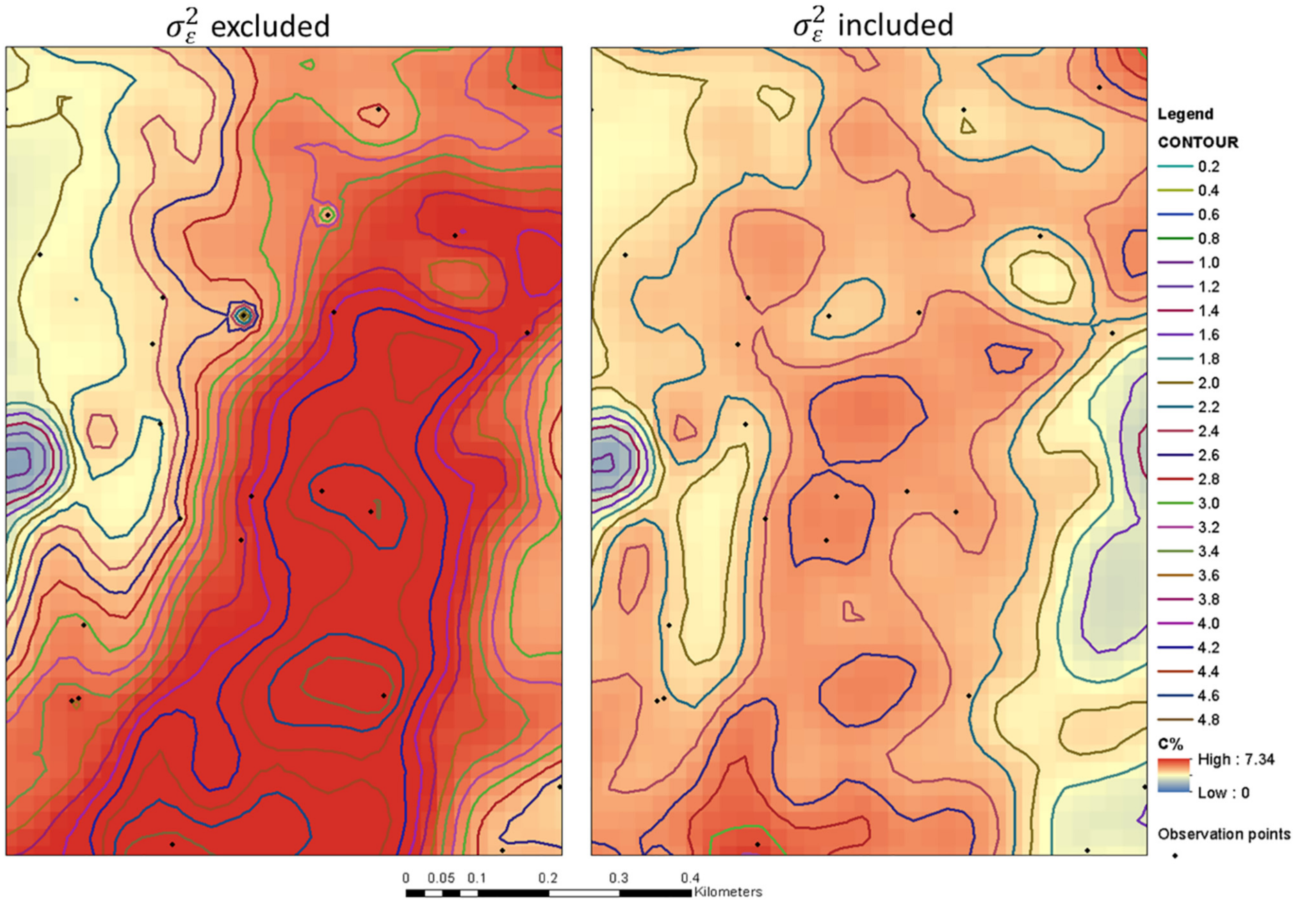


Fig. 9. 0.2% contouring of spatially predicted soil carbon for 0–10 cm layer. Figure shows smoother predictions around data points when the measurement error is acknowledged.

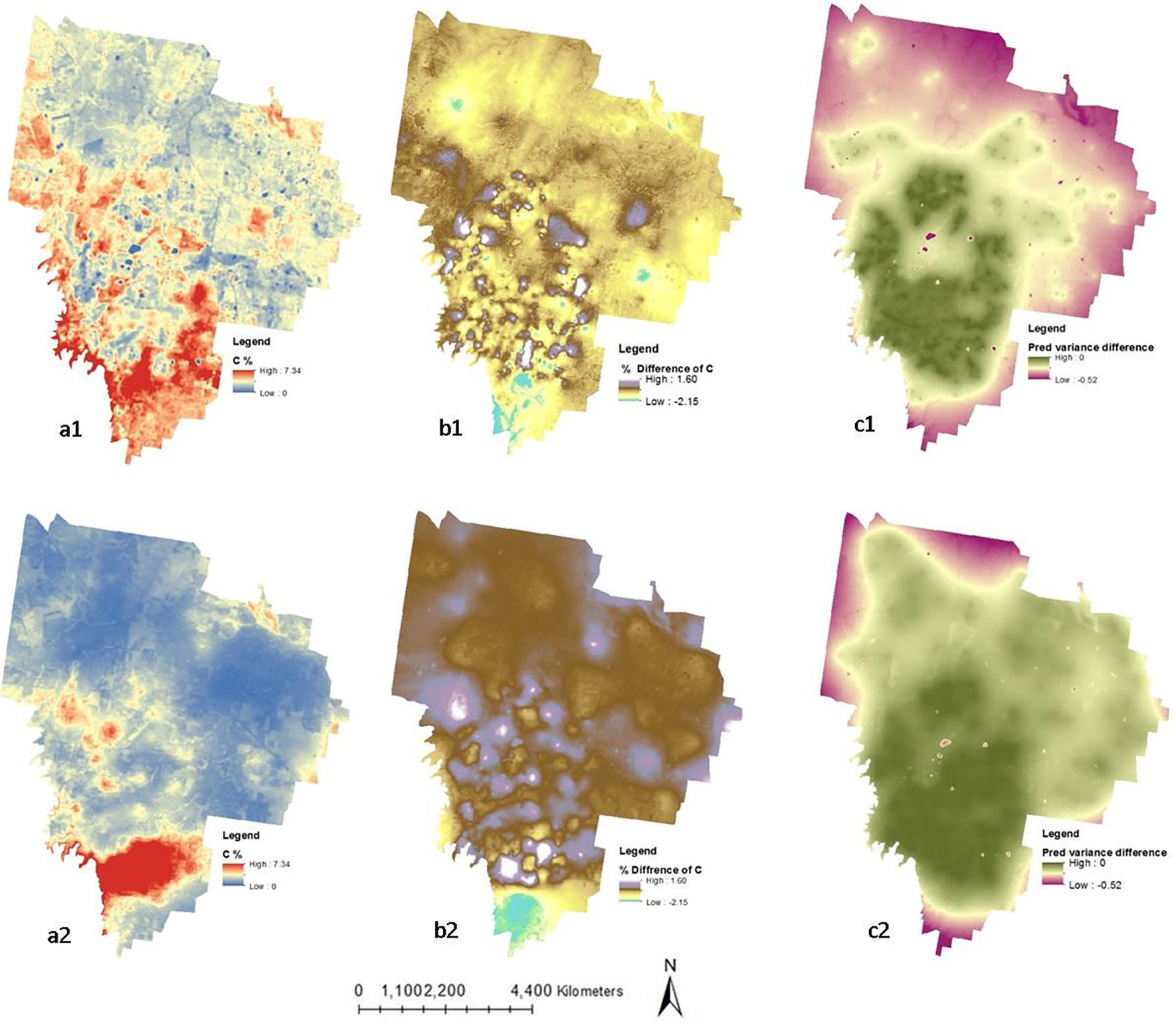


Fig. 10. REML-BLUP predictions and prediction variances of soil carbon for (1)0–10 cm and for (2) 40–50 cm soil layers. (a) predictions for modelling with σ_ε^2, (b) the difference of soil carbon predictions with and without incorporating σ_ε^2. (c) difference between prediction variances between the two scenarios

content of the topsoil layer is closely related to the environmental covariates NDVI and the elevation of the study area. High carbon contents are mostly seen in highly elevated forested areas. High carbon content towards the south of the study area is caused by the presence of marl. Fig. 10b and c respectively are the differences between predictions and prediction uncertainties of both modelling scenarios. Negative values indicate higher corresponding cell values when σ_ε^2 is ignored. If we used a common σ_ε^2 for data points, theoretically, the difference between predictions of both scenarios should be zero. However, in this study we used different σ_ε^2 for each data type leading to a difference between predictions. The colours towards the top of the colour legend indicate positive values closer to zero, whereas the colours towards the bottom of the colour legend indicate negative values closer to zero. Accordingly, soil carbon content of 0-10 cm soil layer is mostly over predicted when σ_ε^2 ignored. For the 40–50 cm soil depth, a substantial part of the study the corresponding map shows that the soil carbon content is slightly under predicted when σ_ε^2 is ignored. The influence of environmental covariates on soil carbon content is minimal for the 40–50 cm soil depth and the prediction differences of this soil layer is more closer to zero than the top soil layer. Perhaps, the effect of environmental covariates has also contributed to non- zero prediction differences between the two testing scenarios other than the difference between σ_ε^2 among data points.

Fig. 10c shows the difference between prediction variances for the two scenarios; with and without inclusion of σ_ε^2 for both soil layers. The difference is minimal in the vicinity of the observation points, and it gradually increases as distance increases from the points for both soil layers. Negative values indicate the exclusion of σ_ε^2 has resulted in higher prediction uncertainty, but in contrast, the prediction variance have clearly reduced by the inclusion of σ_ε^2. This indicates that the filtering of measurement error from data yields a higher prediction certainty.

5. Conclusions

- The use of rapidly acquired spectroscopic measurements of soil attributes as input data in spatial modelling is rapidly growing. It is important to use this data appropriately. In short, measurement error should be accounted for in the digital soil mapping framework.
- Filtering measurement error from the data leads to filtering the σ_ε^2 from the total variation (sill) of the variogram. The inclusion of error variance lowers the overall uncertainty of spatial predictions.

Acknowledging the measurement error is an effective way to improve confidence in prediction.

- MCMC can be used for estimating the distribution function of error variance parameters. This is an important finding that can be used in spatial modelling of soil to compute σ_ε^2 estimated directly from data in the absence of laboratory accuracy comparisons.
- In terms of the accuracy of predictions and goodness of the calibration models of REML-EBLUP $+\sigma_\varepsilon^2$ provides a comparable accuracy to MCMC.
- The LMM approach utilised in this study can be used to account for other types of measurement error, including data estimated from pedotransfer functions. Further work needs to include the effect of spatial position errors in the LMM.
- Although this study is focused on filtering the measurement errors of soil carbon data, the technique is amenable for other data (for example soil, air, and water data) sources with known uncertainty. If the uncertainty of data is not explicitly quantified, the MCMC techniques can be used to quantify the measurement errors.

Acknowledgement

The authors are grateful to the Australian Department of Agriculture, Round 2- Filling the Research Gap Program for supporting this research (Grant number - 1194105-66). A special thank is extended to Professor Gerard Heuvelink, Wageninen university, the Netherlands for giving valuable insights to this manuscript. We also extend our acknowledgement to the editor and the two anonymous reviewers whose constructive comments helped to improve the content of the manuscript.

Appendix A

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2018.02.302.

References

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils - critical review and research perspectives. Soil Biol. Biochem. 43 (7), 1398–1410.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends Anal. Chem. 29 (9), 1073–1081.

Christensen, W.F., 2011. Filtered kriging for spatial data with heterogeneous measurement error variances. Biometrics 67 (3), 947–957.

Clark, I., 2010. Statistics or geostatistics? Sampling error or nugget effect? J. South. Afr. Inst. Min. Metall. 110 (6), 307–312.

Cressie, N.A.C., 1991. Statistics for Spatial Data. Wiley, New York.

Cressie, N.A.C., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. Wiley, Hoboken, N.J.

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V., Wikle, C.K., 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol. Appl. 19 (3), 553–570.

Dawson, J.J.C., Smith, P., 2007. Carbon losses from soil and its consequences for land-use management. Sci. Total Environ. 382 (2), 165–190.

Delhomme, J., 1978. Kriging in the hydrosciences. Adv. Water Resour. 1 (5), 251–266.

Falloon, P., Betts, R., 2010. Climate impacts on European agriculture and water management in the context of adaptation and mitigation—the importance of an integrated approach. Sci. Total Environ. 408 (23), 5667–5687.

Geeves, G.W., New South Wales. Department of, C., Land, M, Soils, C.D.o, 1995. The Physical, Chemical and Morphological Properties of Soils in the Wheat-belt of southern N.S.W. and Northern Victoria. 9780643053984;0643053980. CSIRO Division of Soils, Glen Osmond, S. Aust.

Hawley, S., Glen, R., Baker, C., 1995. Newcastle Coalfield Regional Geology 1: 100 000. Geological Survey of New South Wales. Sydney, Australia.

Huang, J., Malone, B.P., Minasny, B., McBratney, A.B., Triantafilis, J., 2017. Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. Sci. Total Environ. 609 (Suppl. C), 621–632.

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Aust. J. Soil Res. 45 (2), 73–81.

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 (3–4), 227–246.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304 (5677), 1623–1627.

Lark, R.M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. Eur. J. Soil Sci. 51 (4), 717–728.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. Eur. J. Soil Sci. 55 (4), 799–813.

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.

Laslett, G., McBratney, A., 1990. Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes—a case study for soil pH. Eur. J. Soil Sci. 41 (3), 451–471.

Lin, L.I., 1989. A concordance correlation-coefficiecnt to evaluate reproducibility. Biometrics 45 (1), 255–268.

Malone, B.P., de Gruijter, J.J., McBratney, A.B., Minasny, B., Brus, D.J., 2011. Using additional criteria for measuring the quality of predictions and their uncertainties in a digital soil mapping framework. Soil Sci. Soc. Am. J. 75 (3), 1032–1043.

Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. Geoderma 262, 243–253.

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

McKay, M.D., 1995. Evaluating Prediction Uncertainty. US Nuclear Regulatory Commission.

Minasny, B., McBratney, A.B., 2005. The Matérn function as a general model for soil variograms. Geoderma 128 (3–4), 192–207.

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.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matern covariance function. Geoderma 140 (4), 324–336.

Minasny, B., McBratney, A.B., Salvador-Blanes, S., 2008. Quantitative models for pedogenesis - a review. Geoderma 144 (1–2), 140–157.

Minasny, B., Vrugt, J.A., McBratney, A.B., 2011. Confronting uncertainty in model-based geostatistics using Markov chain Monte Carlo simulation. Geoderma 163 (3–4), 150–162.

Mossel, E., Vigoda, E., 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. Ann. Appl. Probab. 16 (4), 2215–2234.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347.

Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma 163 (1–2), 38–44.

Poggio, L., Gimona, A., Spezia, L., Brewer, M.J., 2016. Bayesian spatial modelling of soil properties and their uncertainty: the example of soil organic matter in Scotland using R-INLA. Geoderma 277, 69–82.

Quinlan, J.R., 1992. Learning with continuous classes. Proc. of the Fifth Australian Joint Conference on Artificial IntelligenceWorld Scientific, Singapore, pp. 343–348.

Reeves III, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma 158 (1–2), 3–14.

Rial, M., Martínez Cortizas, A., Rodríguez-Lado, L., 2017. Understanding the spatial distribution of factors controlling topsoil organic carbon content in European soils. Sci. Total Environ. 609 (Suppl. C), 1411–1422.

Rossel, R.A.V., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63 (6), 848–860.

Schabenberger, O., Gotway, C.A., 2017. Statistical Methods for Spatial Data Analysis. CRC Press.

Stenberg, B., Rossel, R.A.V., Mouazen, A.M., Wetterlind, J., 2010. Visible and near Infrared Spectroscopy in Soil Science. In: Sparks, D.L. (Ed.), Advances in Agronomy. Advances in Agronomy Vol 107, pp. 163–215.

Stevens, A., Nocita, M., Toth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS One 8 (6).

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131 (1–2), 59–75.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic-splines and classification for predicting soil properties from sample information. J. Soil Sci. 41 (3), 473–490.

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44.

Waller, L.A., Gotway, C.A., 2004. Applied Spatial Statistics for Public Health Data, 368. John Wiley & Sons.

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Sci. Total Environ. 557–558 (Suppl, C), 838–850.

Zobeck, T.M., Baddock, M., Van Pelt, R.S., Tatarko, J., Acosta-Martinez, V., 2013. Soil property effects on wind erosion of organic soils. Aeolian Res. 10, 43–51.