Geoderma 403 (2021) 115232

Contents lists available at ScienceDirect

Geoderma

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

Using remote sensors to predict soil properties: Radiometry and peat depth in Dartmoor, UK

B.P. Marchant

British Geological Survey, Keyworth NG12 5GG, UK

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Peat
Radiometric surveys
Geostatistics
Splines
Optimized Sampling

ABSTRACT

Remote sensors provide high resolution data over large spatial extents that can potentially be used to map soil properties such as the concentration of organic carbon or its moisture content. The sensors rarely measure the property of interest directly but instead measure a related property. There is a need to make ground measurements of the property of interest to calibrate a model or relationship between the soil property and the sensor data.

We develop a framework for optimizing the locations and number of ground measurements of a soil property for surveys incorporating sensor data. The data are used to estimate a linear mixed model of the property where the fixed effects are a flexible spline-based function of the sensor measurements.

The framework is used to map peat depth across a portion of Dartmoor National Park using radiometric potassium data measurements from an airborne survey. The most accurate maps result from using a geostatistical predictor to combine the relationship with the sensor data and the spatial correlation amongst the peat depth measurements. The optimal sampling designs suggest that ground measurements should be focussed where peat depths are largest and most uncertain. When measurements are made at 25 optimally selected sites, predictions that do not utilise the sensor data have 20% larger root mean square errors than those that do. For 200 ground measurements this benefit is 14%. The maps produced using the sensor data and 25 ground measurements have smaller root mean square errors than those based only upon 200 ground measurements.

1. Introduction

Soil forms over long time-scales and can be considered to be a non-renewable resource. Many natural and anthropogenic processes threaten soil health and quality. At the European scale, Stolte et al. (2015) identified and reviewed 11 of these threats. These were soil erosion by water, soil erosion by wind, decline of organic matter in peat, decline of organic matter in minerals soils, soil compaction, soil sealing, soil contamination, soil salinization, desertification, flooding and landslides and decline in soil biodiversity. There is an urgent need to measure soil health, quality and function to understand where these threats apply and to quantify their potential impacts. However, soil measurements are generally costly and time consuming and often require samples of soil to be collected and taken to a laboratory for preparation and analyses. Also, many such samples are required for broad scale prediction of the variation of soil properties such as the concentrations of soil nutrients and contaminants or the depth of the soil. Remote sensing (Ravi Shanker, 2017) offers an alternative approach to monitoring soils that in many cases utilises existing data consisting of many more measurements and which cover a wider area than could be achieved by other means.

Generally, remote sensing approaches do not directly measure the soil property that is of interest. For example, Minasny et al. (2019) review the use of remote sensing data to map the extent and quality of peatlands. Visible and infrared sensors provide an indication of the land cover and vegetation at a location and specific spectral signatures associated with peatlands can be identified (Dissanska et al., 2009). Radar sensors measure the energy backscattered from the surface and can provide an indication of soil moisture (Poggio and Gimona, 2014). Gamma radiometric sensors measure emanations of a set of radioactive isotopes from the soil and underlying rocks. These emanations occur from the natural decay of radioactive elements such as potassium, thorium and uranium. The amplitude of the radiometric signal emanating from rocks, which is dependent on the geological setting, is attenuated by overlying layers of peat and variation in the measured amplitudes can be used to infer variation in the depth of the peat (Beamish, 2014). It is therefore important to appreciate, quantify and take steps to minimise the uncertainties associated with inferring soil

E-mail address: benmarch@bgs.ac.uk.

https://doi.org/10.1016/j.geoderma.2021.115232
Received 25 February 2021; Received in revised form 18 May 2021; Accepted 20 May 2021
Available online 8 June 2021

properties from these indirect measurements.

Remote sensing data might be used in one of two ways to map soil properties. A generally applicable relationship between the sensor data and the property of interest could be determined and then applied across either a subset of or the entire spatial-extent of the remote sensing data. This could lead to regional, national or even global maps of soil properties. For instance, Airo et al. (2014) placed a threshold on radiometric potassium measurements to distinguish between shallow (less than 0.6 m) and deep (greater than 0.6 m) peat across Finland. Alternatively, geostatistical approaches (Webster and Oliver, 2007) could be used to integrate measurements of the property of interest with the remote sensing data and to interpolate predictions of the property of interest where it has not been measured. Such approaches utilise any spatial autocorrelation amongst measurements of the property of interest which implies that measurements made a short distance apart are relatively likely to be similar. Keaney et al. (2013) used cokriging (Webster and Oliver, 2007) to integrate field measurements of peat depth with airborne gamma radiometric data and map peat depth across part of the Republic of Ireland. The inclusion of the radiometric data within their geostatistical model led to reduced uncertainty in the predictions of peat depth relative to that arising from interpolation of the measured peat depths. Both approaches require some measurements of the soil property of interest to calibrate a generally applicable model and in the second approach to include within the geostatistical predictor. The number and locations of the measurements must be chosen carefully to ensure that the full potential of using remote sensing data for mapping soil properties is realised.

In this paper we develop a framework for integrating remote sensing data with ground measurements of the soil property and for selecting the number and locations of ground measurements that are required. We use linear mixed models (Lark et al., 2006) to represent the spatial variation of the soil property of interest and the relationship with relevant covariates such as radiometric data, satellite imagery or elevation. This approach has similarities with the classical geostatistics techniques known as universal and regression kriging (Webster and Oliver, 2007). However, our approach differes in that we use the maximum likelihood estimator (Lark et al., 2006) to estimate model parameters meaning that the likelihood can be used as a criterion by which the appropriateness of different models can be compared.

Within the linear mixed models we use spline basis-functions (Wood, 2017) to relax the standard assumption of linear relationships between the property of interest and the covariates. Once a linear mixed model has been calibrated it can be used to simulate realistic sets of measurements of the property of interest. We use such realisations to compare the effectiveness of different survey designs for the necessary ground measurements of the property of interest. We consider three cases. In the first, the ground measurements are used to estimate the relationship between the property of interest and the covariates and then this relationship is used to predict the property of interest across the study region without utilising the spatial autocorrelation of the measurements of the property. This is a linear model rather than a linear mixed model. The second case utilises the spatial autocorrelation to interpolate measurements of the property of interest but does not use the remote sensing data. Finally the third case utilises both the underlying relationship with the remote sensing data and the spatial autocorrelation of the property of interest to predict this property where it has not been measured.

In each case we use spatial simulated annealing (SSA) to optimize the configuration of measurement locations to minimise a relevant objective function (van Groenigen et al., 1999). The objective functions reflect the average expected errors in the predictions for the different cases. The SSA approach has been widely used to optimize geostatistical surveys. When it is used to optimize surveys for interpolation of a single variable using a known model of spatial autocorrelation it leads to measurement locations being spread evenly over the study region (van Groenigen et al., 1999). When it is used to optimize predictions using a linear model in isolation then the measurements are restricted to the locations of the largest and smallest values of the covariate and when a linear relationship and a geostatistical model are combined in a linear mixed model the resultant scheme both disperses the measurement locations and ensures that the extremes of the covariate are sampled (Brus and Heuvelink, 2007). If the measurements in the design are to be used to estimate parameters of the geostatistical model then this leads to a proportion of closely located pairs of measurements in the design (Marchant and Lark, 2007). We explore the optimal sample designs that are required for our extended version of the standard linear mixed model and compare the effectiveness of the three prediction cases in terms of the errors that result from specified numbers of measurements.

We consider these approaches in the context of mapping peat depths in the Dartmoor National Park, England using radiometric data. The Dartmoor National Park was covered by the Tellus South West airborne radiometric survey (http://www.tellusgb.ac.uk/) which was flown throughout 2013 (Fig. 1). With reference to this and other airborne radiometric surveys, Beamish (2014) used a theoretical model to demonstrate how the bedrock radiometric signal is attenuated by overlying peat. This attenuation was used to map peat zones in case studies from across the UK (Beamish, 2013, 2014). However, Beamish (2014) concluded that the radiometric signal could not be used to generally map variations in peat depth since (i) for 80% saturated peat, 90% of attenuation occurs in the top 60 cm and it is therefore difficult to discriminate between peats that are deeper than 60 cm, (ii) the degree of attenuation varies according to soil moisture levels, porosity and density, and (iii) the radiogenic parent signal varies according to properties of the bedrock. Beamish (2015) studied the attenuation of the Tellus South West radiometric signals across different bedrock units. Beamish (2015) was able to accurately map peat soils based upon radiometric attenuation within subregions of this dataset. Intra-peat variations in attenuation were observed and interpreted as variations in moisture content, with the lowest amplitude zones corresponding to 100% saturation.

Ground measurements of peat depth are generally made by inserting a thin (typically 1.5 cm diameter) metal probe into the peat until resistance from the underlying soil or bedrock is felt or by using a proximal ground-penetrating radar (Gatis et al., 2019). The peat depth

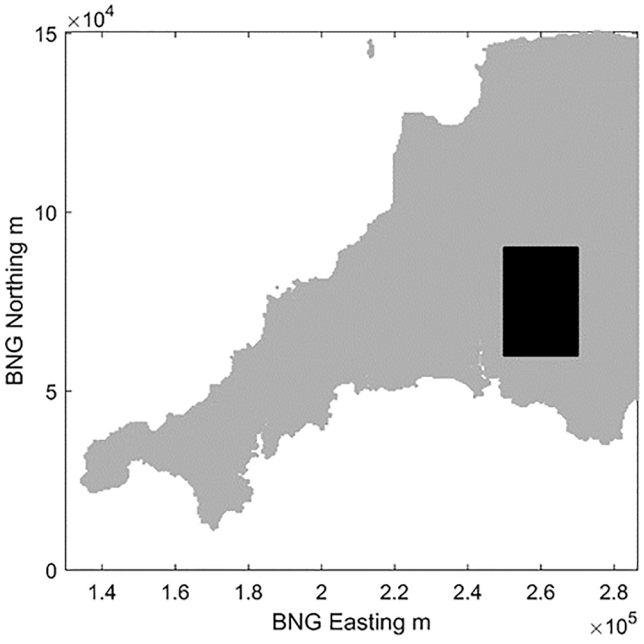

Fig. 1. The south west of England. Black rectangle corresponds to the area covered by the radiometric potassium map shown in Fig. 3 and includes the Dartmoor National Park. BNG refers to British National Grid.

measurements used in this paper were made by Parry et al. (2012) using a metal probe and have been widely studied. Parry et al. (2012) divided the Dartmoor National Park peatland into nine carbon unit areas based on soil and vegetation. They estimated a linear model for peat depth in each carbon unit area using slope and elevation as explanatory covariates and then used these models to predict peat depths across the national park. Subsequently, Parry and Charman (2013) combined the peat depth measurements with measurements of organic carbon concentration and bulk density to predict soil organic carbon stocks across Dartmoor. The peat depth measurements of Parry et al. (2012) were also used by Fyfe et al. (2013) to inform a ground penetrating radar survey in a portion of Dartmoor and to assess the importance of sub-peat carbon storage. Young et al. (2018) noted that the linear models used by Parry et al. (2012) required the assumption that the model residuals were independent whereas there was evidence of spatial autocorrelation amongst these residuals. Therefore, Young et al. (2018) estimated linear mixed models of the same peat depth measurements that included slope and elevation as explanatory covariates and accounted for this spatial autocorrelation. This led to improved validation results. Finlayson et al. (In press) estimated similar linear mixed models of peat depths in the Loch Lomond and the Trossachs National Park, Scotland. Young et al. (2018) noted that their model predictions in Dartmoor were most uncertain on plateaus and in depressions where deep peats were predicted. They considered how future surveys of peat depth should be designed and recommended that plateaus and depressions should be sampled sufficiently densely that spatial autocorrelation between the measurements is evident and therefore the prediction accuracy is improved beyond that which can be achieved by the standard linear model. Gatis et al. (2019) integrated the radiometric data from the Tellus South West survey with the peat depth measurements of Parry et al. (2012). They estimated linear models of the natural logarithm of peat depth and found that the best predictions resulted from using the total radiometric dose and slope as explanatory covariates.

2. Theory

2.1. Linear mixed models

We summarise the underlying theory of the spatial linear mixed models used in this paper. More details can be obtained from Webster and Oliver (2007) and Marchant (2018). The spatial variation of measurements of an environmental property can be represented by a linear mixed model. This model splits the variation into fixed effects which reflect underlying trends in the property that are related to covariates and random effects which reflect the residual variation that cannot be explained by these trends. We denote the measured value of the property at location $\mathbf{x}_i$ by $z(\mathbf{x}_i)$. The linear mixed model is written:

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{z} = [z(\mathbf{x}_1), z(\mathbf{x}_2), \cdots z(\mathbf{x}_n)]^{\mathrm{T}}$ is a vector containing the measured values from n locations. The $\mathbf{M}\boldsymbol{\beta}$ are the fixed effects. Often, the first column of the $n \times q$ design matrix $\mathbf{M}$ contains 1 s and the remaining columns contain covariates that are linearly related to the measured values. The length q vector $\boldsymbol{\beta}$ contains estimated coefficients and thus the fixed effects are equal to a constant plus a linear sum reflecting the relationship between the measured values and a series of covariates. The length n vector $\boldsymbol{\varepsilon}$ contains the random effect or residual at each measurement location. The $\boldsymbol{\varepsilon}$ vector is assumed to have been realised from a multivariate Gaussian distribution with mean zero and $n \times n$ covariance matrix $\mathbf{C}$. For many soil properties this assumption might not be plausible. For example, the vector of measured values might include a small number of large outlying values leading to the distribution being highly skewed. A transformation might be applied to such variables so that they comply with the Gaussian assumption. For highly skewed properties, a linear mixed model of the logarithm of the measured values is often estimated (Webster and Oliver, 2007).

If the distribution of the random effects (possibly following a transformation) is assumed to be second order stationary then the variances on the main diagonal of $\mathbf{C}$ are identical. The off-diagonal terms indicate the degree of spatial correlation between pairs of random effects. In a linear model these off-diagonal elements are zero. The spatial correlation is assumed to decrease as the lag or distance between pairs of measurement locations become larger according to an authorized covariance function (Webster and Oliver, 2007). One authorised function is the nested nugget and exponential:

$$C(h) = \begin{cases} c_0 + c_1 & \text{if } h = 0 \\ c_1 \exp\left(\dfrac{-h}{a}\right) & \text{if } h > 0 \end{cases} \tag{2}$$

where h is the lag distance separating two measurement locations, c_0 the nugget variance, c_1 the partial sill variance (the variance of the spatially correlated component) and a is a spatial parameter. The spatial parameter must be greater than zero and the variances greater than or equal to zero. We refer to the sum of the nugget and partial sill variances as the sill variance.

Both the fixed effects coefficients $\boldsymbol{\beta}$ and the random effects parameters $\boldsymbol{\alpha} = [c_0, c_1, a]$ must be estimated from the available data. This can be achieved by maximum likelihood. The likelihood or probability that the measured data would have been realised from a proposed model can be calculated for any admissible values of $\boldsymbol{\alpha}$. The $\boldsymbol{\beta}$ values that will maximise the likelihood for those $\boldsymbol{\alpha}$ are:

$$\boldsymbol{\beta} = \left(\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{M}\right)^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{z}. \tag{3}$$

A numerical optimization routine can be used to find the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters which lead to the largest value of this likelihood. Further details of maximum likelihood estimation for spatial data are provided by Lark et al. (2006).

One challenge when proposing linear mixed models or linear models of a particular property is deciding how many and which covariates should be included in the fixed effects design matrix $\mathbf{M}$. The addition of an additional covariate to this matrix will not lead to a decrease in the maximised likelihood because an unaltered model and likelihood results if the additional element of $\boldsymbol{\beta}$ is zero. Therefore, it might appear that a model is being improved by the inclusion of additional covariates since the maximised likelihood tends to increase. However, the additional covariates might not be significantly related to the property of interest and the improvements in likelihood might only reflect patterns in the property and covariates that occur by chance. Such a model is said to be overfitted. The model accurately represents the variation in the data that are used to estimate it but is less accurate when applied to other data. The problem of overfitting in spatial models is often addressed using the Akaike Information Criterion (AIC; Akaike, 1973):

$$\mathrm{AIC} = 2k - 2L, \tag{4}$$

where k is the number of estimated parameters or coefficients in a model and L is the natural logarithm of the maximised likelihood. The design matrix $\mathbf{M}$ which leads to the smallest AIC is thought to contain the most appropriate covariates to represent the variation of the property of interest. This criterion favours models which have a large likelihood but penalizes complex models with many parameters.

2.2. Basis splines

In the linear mixed model described in the previous section the fixed effects relationship between a covariate and the measured values of the property of interest is linear. This assumption of linearity is rather restrictive. It can be relaxed by expressing the fixed effects as a linear sum of spline basis functions defined according to the value of the covariate. Spline functions (Wood, 2017) are piecewise polynomials which

means that they consist of a series of smooth sub-functions. The points where these sub-functions meet are referred to as knots.

Fig. 2 illustrates nonlinear fixed effects related to a covariate x. This covariate has been scaled so that its smallest value is zero and largest value is one. Examples of B or Basis splines are shown in the panels on the left of the figure. In each panel a single B-spline sub-function is highlighted in red. These sub-functions have compact support – they are only non-zero within a continuous subset of values of x. All of the sets of B-splines have five equally spaced knots between $x = 0$ and $x = 1$. In the top plot, the ith B-spline is equal to one between the $(i-1)$th and ith knot and is zero otherwise. This is a first order B-spline. A fixed effect design matrix $\mathbf{M}$ could be defined such that each column contains the values of a first order B-spline sub-function corresponding to the covariate value for each measurement. When this design matrix is multiplied by a set of coefficients,$\boldsymbol{\beta}$, a nonlinear fixed effect function results. This fixed effect function has a discontinuity or jump at each knot as can be seen in the top right panel of Fig. 2.

The second order B-spline consists of a triangular sub-functions surrounding each knot. When this basis function is included in a design matrix, fixed effects with discontinuous derivatives at the knots result (Fig. 2, second row). Similarly a third order B-spline leads to fixed effects with discontinuous second derivatives at each knot.

For a B-spline of a given order, the number of knots control the degree of smoothness of the resultant fixed effects. If the number of knots is almost as large as the number of measurements then the fixed effects might follow the measured relationships between the covariate and the property of interest almost exactly. However, it is unlikely that such a detailed relationship would be applicable to data not used in the calibration and the model is overfitted. When estimating B-spline models in a non-spatial context the objective might be to minimise the sum of the squared differences between the measured data and the results of the model. It is common to control for overfitting by introducing a term in the objective function that penalizes rough or non-smooth models (Wood, 2017). In a spatial context the model fitting procedure is more complex since it must account for any spatial autocorrelation amongst the data. This correlation implies that the errors cannot simply be summed within an objective function.

The B-splines described here can be included in the spatial linear mixed model (Eqn. (1)) by using an appropriate $\mathbf{M}$ matrix. One might choose to use third order B-splines to ensure that the resultant fixed effects have continuous derivatives. Different numbers of uniform knots could be applied. The number of knots that leads to the lowest AIC might be considered to have the appropriate degree of smoothness.

2.3. Spatial prediction

Once the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters have been estimated the linear mixed model and measurements of the property of interest can be used to predict the expected value of the property at locations where it has not been measured. The values of any covariates within the fixed effects design matrix must be known at these locations. The $n_p \times q$ fixed effects design matrix containing these covariates at the prediction locations is denoted $\mathbf{M}_p$.

The estimated $\boldsymbol{\alpha}$ and Eqn. (2) are used to calculate $\mathbf{C}_{po}$, the length $n_p \times n$ matrix of covariances between potential measurements at the locations where predictions are required and the actual measurements. Similarly, the $n_p \times n_p$ covariance matrix of potential measurements at the prediction locations is calculated and denoted $\mathbf{C}_{pp}$. The predicted expectation of the possibly transformed property at the prediction locations is:

$$\widehat{\boldsymbol{Z}}(\mathbf{x}_p) = \mathrm{E}[\mathbf{z}(\mathbf{x}_p)] = \mathbf{M}_p\boldsymbol{\beta} + \mathbf{C}_{po}\mathbf{C}^{-1}(\mathbf{z} - \mathbf{M}\boldsymbol{\beta}), \tag{5}$$

and the corresponding prediction covariance matrix is:

$$\mathbf{V} = (\mathbf{M}_p - \mathbf{C}_{po}\mathbf{C}^{-1}\mathbf{M})(\mathbf{M}^T\mathbf{C}^{-1}\mathbf{M})^{-1}(\mathbf{M}_p - \mathbf{C}_{po}\mathbf{C}^{-1}\mathbf{M})^T + (\mathbf{C}_{pp} - \mathbf{C}_{po}\mathbf{C}^{-1}\mathbf{C}_{po}^T) \tag{6}$$

If the autocorrelation of the property of interest is not utilised when making predictions then the matrix $\mathbf{C}_{po}$ is treated as if all of its entries were zeros and the expected value and prediction variance are:

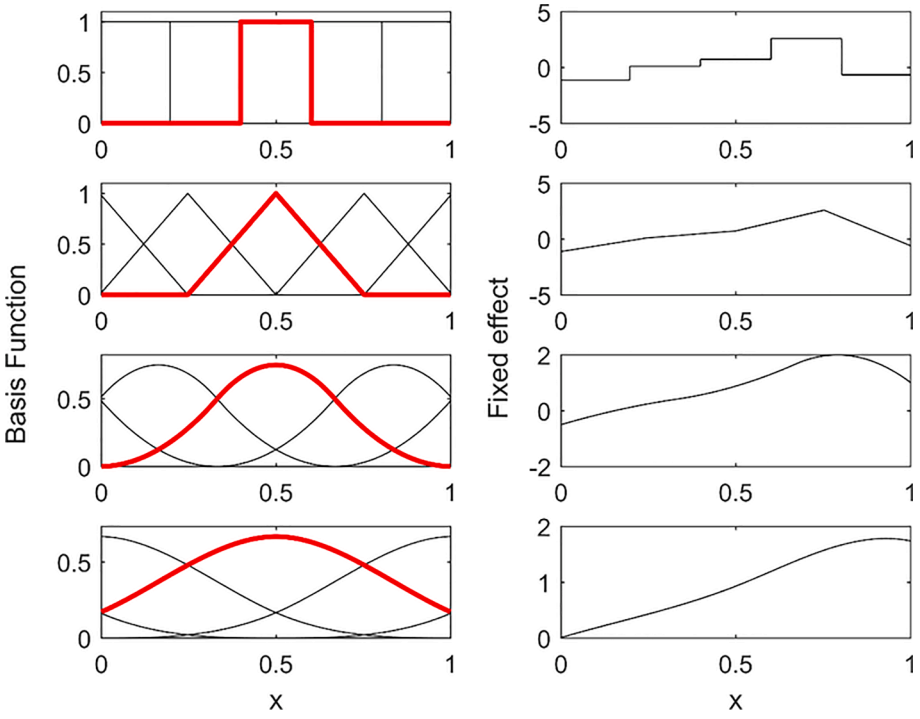

Fig. 2. Examples of B-spline basis functions for covariate (x) values between zero and one (left) and of fixed effects that result by multiplying each set of basis functions by the same randomly selected regression coefficients (right). An individual basis sub-function is highlighted in red. The order of the B-splines increases from two to five upon moving down through the plots. All plots have five uniformly spaced knots.

$$\widehat{\boldsymbol{Z}}(\mathbf{x}_\mathrm{p}) = \mathrm{E}\left[\boldsymbol{z}(\mathbf{x}_\mathrm{p})\right] = \mathbf{M}_\mathrm{p}\boldsymbol{\beta}, \tag{7}$$

and:

$$\mathbf{V} = \mathbf{M}_\mathrm{p}\left(\mathbf{M}^\mathrm{T}\mathbf{C}^{-1}\mathbf{M}\right)^{-1}\mathbf{M}_\mathrm{p}{}^\mathrm{T} + \mathbf{C}_\mathrm{pp}. \tag{8}$$

For these linear model predictions the main diagonal entries of $\mathbf{C}$ and $\mathbf{C}_\mathrm{pp}$ are equal to the sill variance and the off-diagonal entries are zero.

For both the linear mixed model and linear model predictors the elements of the main diagonal of $\mathbf{V}$ correspond to the variance or uncertainty of the predictions. The first term on the right hand side of Eqs. (6) and (8) account for the uncertainty in the fixed effects. The second term accounts from the uncertainty that results from the residual spatial variation of the property. In Eq. (6) the $\mathbf{C}_\mathrm{po}\mathbf{C}^{-1}\mathbf{C}_\mathrm{po}^\mathrm{T}$ is the amount by which the spatial autocorrelation has reduced the prediction variances and covariances. Note that these uncertainties are related to the covariance matrices of the property of interest but not directly to the measured values. These matrices can be calculated if the $\boldsymbol{\alpha}$ parameters are known. The uncertainty of the fixed effects can be calculated without knowing the value of the $\boldsymbol{\beta}$ coefficients.

The assumption of Gaussian random effects in the linear mixed model and the linear model imply that the distribution of the predictions is multivariate Gaussian and therefore the expected values and covariance matrix are sufficient information to determine the probability density function for each prediction location. This also implies that it is possible to sample plausible realisations of the property by using the Cholesky decomposition (Webster and Oliver, 2007) to sample from the multivariate Gaussian distribution with expectation $\widehat{\boldsymbol{Z}}(\mathbf{x}_\mathrm{p})$ and covariance matrix $\mathbf{V}$.

If the property of interest was transformed prior to modelling then a backtransform must be applied so that the predictions can be presented in their original units. It is not generally possible to simply apply the inverse of the transform to the predicted values of the property. For example, if the exponential of the prediction of a log transformed property is calculated then the result is equal to the median of the predicted distribution in the original units rather than the mean. The backtransformation can be easily achieved using simulated values. If 1000 simulated values of a log-transformed property at a particular location are produced then the exponential of each simulated value can be calculated. The mean of these values is an estimate of the mean of the predicted distribution of the property in original units.

The effectiveness of different linear mixed models or linear models can be quantified by validation of the model predictions. Ideally, some measurements would be held back and not used in the model calibration so that they can be used for validation. Often, a scarcity of measurements means that this is not practical. Instead, a cross-validation procedure is used where the model is re-estimated without a proportion of randomly selected measurements and then the property of interest is predicted at the locations of the omitted measurements. This procedure might be repeated, selecting different measurements to be omitted each time until each measurement has been omitted once. Upon cross-validation, the mean error (ME):

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left\{z_i - \widehat{Z}_i\right\}, \tag{9}$$

indicates whether the spatial predictions are biased and the root mean squared error (RMSE):

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}\left\{z_i - \widehat{Z}_i\right\}^2\right]^{\frac{1}{2}}, \tag{10}$$

quantifies the accuracy of the predictions. Here, z_i and $\widehat{Z}_i$ are the measured and cross-validation predicted value at $\mathbf{x}_i$. The correlation between measured and predicted values might be considered as an additional cross-validation metric. However, standard correlation coefficients only consider the extent to which pairs of variables are linearly related to each other rather than the similarity between the two variables. Lin's concordance correlation (Lin, 1989) which does consider this similarity is therefore a better cross validation statistic.

The prediction variances might be validated by calculating the standardised squared prediction errors (SSPEs):

$$\theta_i = \frac{\left\{z_i - \widehat{Z}_i\right\}^2}{V_i}, \tag{11}$$

where V_i is the variance of the prediction at location $\mathbf{x}_i$. The assumptions of Gaussian random effects in the linear mixed model and linear model imply that the θ_i should be realised from a chi-squared distribution with one degree of freedom. The mean $\overline{\theta}$ and median $\widetilde{\theta}$ of such a distribution have expected values of 1.0 and 0.45 respectively and these quantities are often quoted (e.g. Minasny and McBratney, 2007) as measures of the accuracy with which the uncertainty in model predictions have been quantified.

2.4. Optimized spatial sampling

The data collected in a survey of a spatial variable must be suitable for both estimation of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters of the model and spatial prediction of the variable. The number of measurements that can be made is often limited according to the resources available for the survey. Therefore it is important to select the measurement locations carefully. Van Groenigen et al. (1999) suggested the SSA optimization procedure to select the locations of a fixed number of measurements that constitute a spatial survey. The SSA procedure converges to a set of locations that minimize an objective function. The objective function is likely to reflect the level of uncertainty that will result from the survey. For example, if the $\boldsymbol{\alpha}$ were known for a property that was represented by a linear mixed model (Eqn. (1)) the objective function could be the average variance of predictions (Eqn. (6)) at a number of specified prediction locations. This objective function could be calculated prior to sampling without knowledge of the measurement values or the entries of $\boldsymbol{\beta}$.

The SSA algorithm starts with a random selection of measurement locations. Then a location is perturbed randomly. The perturbation is accepted if it causes the objective function to decrease. A perturbation that increases the objective function might be accepted. The probability of acceptance decreases with the magnitude of the increase in the objective function. If the perturbation is not accepted then the measurement returns to its previous location. The algorithm continues iteratively, perturbing each measurement location in turn and then repeating the process across a number of cycles. The probability of accepting a perturbation that increases the objective function is decreased upon the start of each cycle and the locations converge to a pattern which minimizes the objective function. The stochastic nature of the algorithm means it is likely to converge to a global rather than local minima.

If the linear mixed model has constant fixed effects and the objective function is the average prediction variance (Eqn. (6)) then the measurement locations are likely to be evenly dispersed across the study region. Such a design ensures that no prediction location is a large distance from a measurement and therefore the measurements are suitable for interpolation. Brus and Heuvelink (2007) explored sample designs that minimised the average prediction variance when the fixed effects were linear functions of covariates. They found that as well as dispersing the measurement locations across the study region these designs tended to include both large and small values of the covariate in order to accurately estimate the gradient of the relationships with the covariates.

Marchant and Lark (2007) and Zhu and Stein (2006) added additional terms to the objective function which approximated the effects of uncertainty of the estimated $\boldsymbol{\alpha}$ parameters. If the additional variance resulting from $\boldsymbol{\alpha}$ parameter uncertainty in the prediction at $\mathbf{x}_i$ is denoted

τ_i^2 then the prediction variance at this location is:

$$\mathbf{W}_{ii} = \mathbf{V}_{ii} + \tau_i^2 \quad (12)$$

where $\mathbf{V}_{ii}$ is the ith entry of the main diagonal of $\mathbf{V}$. Both sets of authors found that their objective functions led to sample designs where the measurement locations were generally evenly dispersed across the study region but there was a proportion of close pairs of locations which were particularly useful in estimating the nugget parameter. Lark and Marchant (2018) optimized survey designs for variables with different spatial model parameters and determined that if a tenth of observations are close pairs then the design is generally suitable for both covariance model estimation and geostatistical prediction. Wadoux et al. (2017) optimized survey designs for non-stationary geostatistical models where the variance of the property of interest varied according to a covariate. They found that sampling was particularly focussed where the variance was largest.

3. Methods

3.1. Peat depth measurements

We consider the 425 peat depth measurements from the Dartmoor National Park (Fig. 3) which, as described in the Introduction, were studied by Young et al. (2018). These measurements were a subset of the measurements obtained by Parry et al. (2012) with the permission of the National Trust and Dartmoor National Park Authority. These two papers give full details of how the peat depths were measured, the sample design and of the study region and Young et al. (2018) describe how the subset of the data can be accessed. An extended dataset has since been studied by Gatis et al. (2019).

Dartmoor has high average rainfall of 1974 mm (Met Office, 2010) and underlying geology of impermeable granite. This has led to the formation of extensive blanket peatland. The study region includes two blanket peat soil series - Crowdy and Winter Hill. The peat depths were measured with an extendable metal probe. The quoted depths are the average of five replicates. One measurement is made at a central point and the other four are made 4 m from this point to form a cross. Two different sampling approaches were employed. Peat depths were measured on a regular grid with 250 m intervals in the south of the study region. A stratified sample approach was used in both the north and south of the region to ensure representative sampling of slope and elevation.

3.2. Covariate data

The covariate data used in this study were originally compiled by Kirkwood et al. (2016) and used to map geochemical properties across the south west of England. We focus upon radiometric data obtained from the Tellus South West airborne survey (Beamish et al., 2014). The survey flew 61,000 km of lines across the south west of England with each line separated by 200 m. The radiometric data were sampled at 1 Hz leading to a mean distance between measurements along the line of 71 m. The horizontal support or footprint of each measurement varied according to the altitude and speed of the aircraft. Beamish (2013) determined that 90% of the response for a measurement from the Tellus Northern Ireland airborne survey flown at an altitude of 60 m would have an elliptical footprint of area 109 000 m^2. The greatest contribution came from directly beneath the aircraft and it would fall off rapidly with lateral distance from the flight line. The Tellus South West survey was flown at an average altitude of 91.6 m. The radiometric counts of potassium, thorium and uranium and the total counts were interpolated to a 100 m grid using bicubic splines (Fig. 3). Beamish et al. (2014) provide full details of the radiometric component of the Tellus South West survey and the pre-processing of the data.

Kirkwood et al. (2016) considered a further 22 covariates derived from the NextMap aerial elevation survey (Intermap Technologies, 2007), a land gravity survey (British Geological Survey, 1968), the magnetic component of the Tellus South West survey (Beamish et al., 2014) and Landsat-8 satellite imagery (Roy et al., 2014). All of these covariates were resampled to the same 100 m grid as the radiometric survey data.

3.3. Statistical analyses

The values of the covariates corresponding to each of the 425 locations where peat depths were measured were determined by finding the nearest neighbour to the measurement locations amongst the covariate grid. Linear mixed models were then used to investigate the

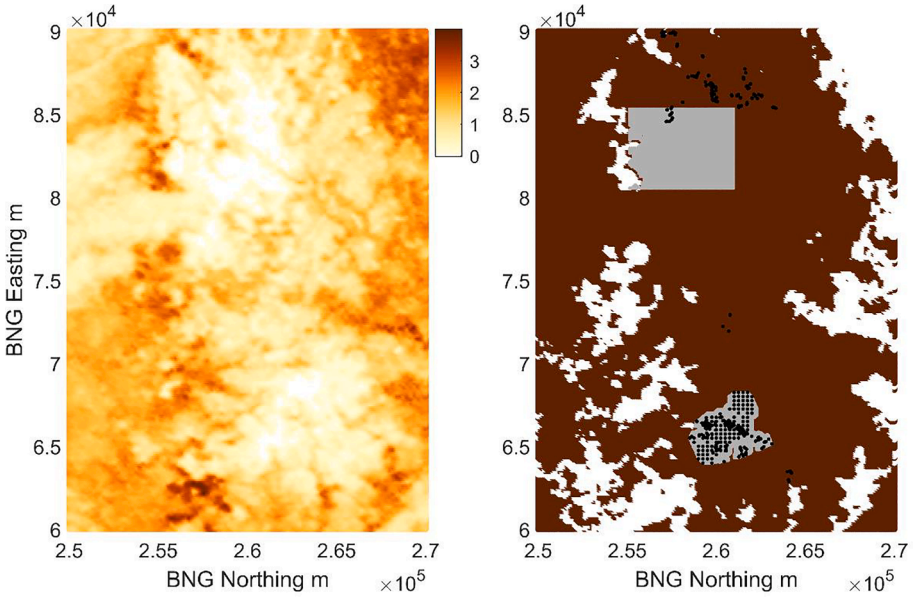

Fig. 3. (left) Interpolated map of BGS Tellus Survey radiometric potassium (%) for portion of south west England highlighted in Fig. 1; (right) locations where interpolated potassium is less than 1.9% (brown shading) with annotated measurement locations (black dots), area of predicted maps (southerly grey shading) and sub-region where optimized sampling is explored (northerly grey shading). Contains data sourced from Young et al. (2018).

relationships between peat depth and these covariates. An initial linear mixed model with constant fixed effects was estimated by maximum likelihood. Then a series of linear mixed models which included one of the covariates compiled by Kirkwood et al. (2016) were estimated. The AIC of each model was recorded. Nonlinear relationships between the peat depths and covariates were accommodated by using third order B-spline basis functions of the covariates in the fixed effects design matrix. The knots of these basis functions were uniformly spaced between the smallest and largest covariate value. Such models were estimated with different numbers of knots varying between four and eight. The AIC was used as a criterion to decide upon the optimal number of knots for each covariate. Ten-fold cross validation was applied to the estimated models and the MEs, RMSEs, Lin's concordance correlation, and the mean and median SSPEs were calculated and compared.

The standard cross-validation results use the linear mixed model prediction equations (Eqs. (5) and (6)) and reflect the impact of both the fixed and random effects upon the predictions of peat depth. We also considered how effective the fixed effects were in isolation by using the linear model predictor (Eqs. (7) and (8)).

Subsequent analyses focussed on the best fitting linear mixed model. This model used radiometric potassium measurements in the fixed effects. The estimated model was used to predict maps of peat depths in the area covered by the gridded peat depth measurements in the south of the study region. Two sets of maps were produced. The first set were purely based upon the linear model predictor. The second set utilised both the fixed and the random effects in interpolating peat depth across the study region. Plots of the estimated peat depth and the width of the 90% confidence interval were produced.

Finally, we considered how the model incorporating radiometric potassium measurements could be used to optimize the design of a survey of peat depths. A 277 000 m^2 sub-region in the north of the study region was selected. This sub-region was based on a 5800 m $\times$ 4800 m rectangle but locations where radiometric potassium was greater than 1.9% (the largest measured value at the location of the peat depth measurements) were omitted. The peat depth measurements were assumed to be realized from the best fitting linear mixed or linear models. Spatial simulated annealing was used to design peat depth surveys for this sub-region that minimized the mean width of the 90% prediction intervals for both the linear mixed model and linear model predictions on a regular 100 m spaced grid of n_p prediction locations covering the sub-region. For Gaussian variables this objective function O was calculated using:

$$O = \frac{1}{n_p}\sum_{i=1}^{n_p} 3.29\sqrt{\mathbf{W}_{ii}}, \tag{13}$$

where $\mathbf{W}_{ii}$ was the ith main diagonal entry of the covariance matrix $\mathbf{W}$ defined in Eqs. (6) and (12) for the linear mixed model and Eqs. (8) and (12) for the linear model. The best fitting $\boldsymbol{\alpha}$ values were assumed to be known and used to calculate the $\mathbf{V}$ and $\mathbf{W}$ covariance matrices. In the case of a variable that had undergone a log-transform then the objective function was calculated in the original units using:

$$O = \frac{1}{n_p}\sum_{i=1}^{n_p} \exp\left(\mathbf{M}_{\mathbf{p}(i)}\boldsymbol{\beta} + 1.64\sqrt{\mathbf{W}_{ii}}\right) - \exp\left(\mathbf{M}_{\mathbf{p}(i)}\boldsymbol{\beta} - 1.64\sqrt{\mathbf{W}_{ii}}\right), \tag{14}$$

where $\mathbf{M}_{\mathbf{p}(i)}$ is the ith row of the $\mathbf{M}_{\mathbf{p}}$matrix. Note that Eq. (14) requires that the $\boldsymbol{\beta}$ are known. A survey was also designed where the fixed effects were assumed to be constant in order to quantify the improvement in prediction accuracy that resulted from the radiometric information.

The number of measurements in these surveys was sequentially increased in increments of 25 from 25 to 200. The effectiveness of these sample designs was tested using simulated data. For each design, the linear mixed model was used to simulate peat depth at the suggested measurement locations and at the nodes of a prediction grid with interval 100 m covering the 277 000 m^2 sub-region. A linear mixed model or a linear mixed model with the same fixed effects design matrix as that assumed in the survey design was estimated using the data from the proposed measurement locations. This model and the data from the proposed locations were then used to predict peat depth on the grid. This process was repeated 100 times and the mean of root mean squared difference between the predicted and simulated data for each realisation was recorded.

4. Results

4.1. Estimation of linear mixed models of peat depths

The measured peat depths varied between 5 and 330 cm (Fig. 4). The distribution of these measurements had a positive skewness coefficient of 1.56. Webster and Oliver (2007) suggest that a variable with skewness coefficient greater than 1.0 cannot be considered to be consistent with a Gaussian distribution. The logarithm of peat depth plus one had a skewness of 0.25. We therefore used linear mixed models to represent the spatial variation of this transformed property. The estimated covariance function for the transformed peat depths with constant fixed effects is shown in Fig. 5. The sill variance is 0.72 (log cm)^2 with a relatively small nugget variance of 0.04 (log cm)^2. Spatial autocorrelation is evident up to 2000 m.

The correlation between the shifted and log-transformed peat measurements and the radiometric potassium data was −0.66. This was the largest magnitude correlation amongst the 22 covariates considered by Kirkwood et al. (2016). The fourth band (red, 0.64–0.67 μm) of the Landsat-8 imagery had a correlation of −0.65 with the shifted and log-transformed peat measurements. The largest correlation between these transformed measurements and a topographic property was 0.45 with elevation. The estimated fixed effects for the linear mixed model of transformed peat depth with linear fixed effects of these three covariates are shown in Fig. 6. In these three examples, the AIC is improved beyond the value of the model with constant fixed effects. The lowest AIC is achieved by the model that includes radiometric potassium (Table 1).

The AIC of the model that includes radiometric potassium is further improved by specifying the fixed effects to be B-spline nonlinear functions of radiometric potassium. The smallest AIC occurs for the B-splines with five knots (Table 1). The corresponding fixed effects are shown in Fig. 7. Nonlinear functions of Landsat-8 Band 4 and elevation do not improve the AIC beyond the values achieved when the fixed effects were linear functions of these variables. The covariance function of the random effects corresponding to these fixed effects is shown in Fig. 5 (right). The inclusion of the fixed effects has reduced the sill to 0.28 (log cm)^2. The nugget variance remains a small proportion of this sill but spatial autocorrelation is only evident up to distances of around 500 m.

4.2. Validation of linear mixed models

When a constant is used to predict the transformed peat depth measurements in 10-fold cross-validation the RMSE is 0.86 log(cm) and Lin's concordance coefficient between the transformed measurements and predictions is 0.00 (Table 2). For the linear model predictions based on the nonlinear function of radiometric potassium or the linear functions of Landsat-8 Band 4 and elevation the RMSEs decrease to 0.53, 0.73 and 0.76 log(cm) respectively and the Lin's concordance coefficients increase to 0.75, 0.31 and 0.31. The cross-validation results in Table 1 further indicate that all of these models are approximately unbiased and the uncertainty of the predictions is adequately modelled since the mean SSPE is close to 1.0 and the median SSPE close to 0.45. When these predictions are back-transformed to the original units the RMSE for the constant is 75.6 cm, for the nonlinear radiometric potassium model it is 45.7 cm, for the linear Landsat-8 Band 4 model it is 67.8 cm and for the elevation model it is 67.7 cm.

A linear mixed model with constant fixed effects achieves superior cross-validation results to any of the linear model predictions (Table 3).

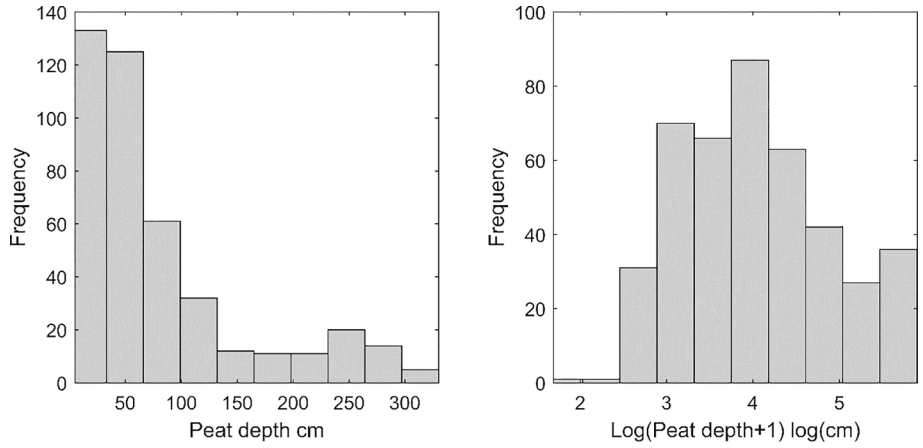

Fig. 4. Histograms of measured peat depth (left) and of shifted and log transformed measured peat depth (right) sourced from Young et al. (2018).

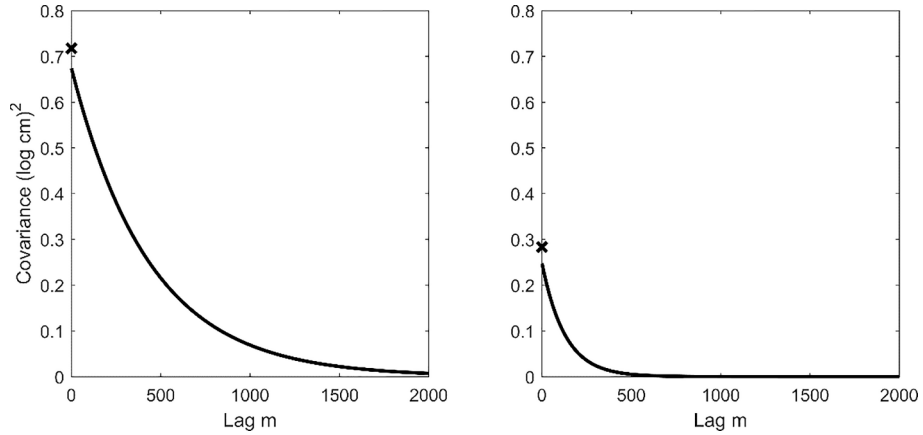

Fig. 5. Estimated covariance function for log(peat depth + 1) with constant fixed effects (left) and with fixed effects that are nonlinear B-spline function of radiometric potassium (right). The crosses indicate the covariance for zero lag.

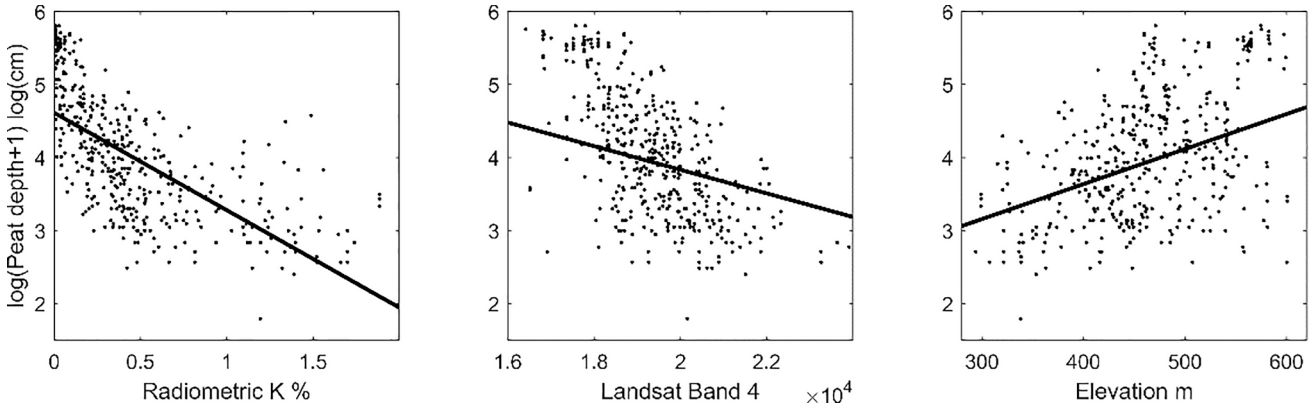

Fig. 6. Fixed effects of estimated linear mixed models relating named covariate to log(peat depth + 1). Measured shifted and log transformed peat depths are denoted by black dots and the estimated linear fixed effects by black line.

The RMSE for the transformed peat depths is 0.46 log(cm) and Lin's concordance coefficient is 0.83. There are small improvements to these cross-validation criteria when the fixed effects are specified to be the nonlinear function of radiometric potassium. Fixed effects that are linear functions of Landsat-8 Band 4 or elevation have almost identical cross-validation results to the constant fixed effects model.

4.3. Spatial prediction of peat depths

The predicted maps of peat depth covering the area where gridded measurements were made (Figs. 8 and 9) indicate that the depths are greatest in the south east of this region. This underlying trend can be seen when using the linear model predictions with nonlinear radiometric potassium fixed effects. More detail is introduced to the maps when the linear mixed model predictor is applied and the random effects

Table 1
AIC values upon maximum likelihood estimation of linear mixed models with different fixed effects.

Fixed effect	Covariate		
	Rad K	LandSat-8	Altitude
Constant	317.25	317.25	317.25
Constant + Linear	286.47	302.14	313.26
B-spline 4 knots	263.80	303.62	314.35
B-spline 5 knots	262.65	304.75	315.73
B-spline 6 knots	263.45	304.99	316.33
B-spline 7 knots	264.50	303.87	317.33
B-spline 8 knots	264.62	302.53	316.78

are included in the predictions. In transformed units (Fig. 8), the width of the prediction interval when using the linear model predictor is relatively constant across the study region. When random effects are also included, smaller uncertainties are apparent in the vicinity of measurement locations. When the predictions are back-transformed to cm (Fig. 9) the degree of uncertainty is primarily controlled by the magnitude of the prediction. Peat depth is more uncertain where it is predicted to be large. Some evidence of the uncertainty being reduced in the vicinity of measurement locations is apparent when the linear mixed model predictor is applied.

4.4. Optimization of sample designs

When a linear model is used to predict peat depths then the objective functions for optimal survey design (Eqns (13) and (14)) are only related to the distribution of covariate values within the sample rather than the spatial configuration of the observation locations. For example, when the model was specified as a linear function of radiometric potassium then the optimal design for prediction of transformed peat depths focussed sampling at the locations where radiometric potassium was either large or small (Fig. 10, left). The resultant measurements would be suitable to accurately estimate the gradient of the linear relationship. If the nonlinear B-spline fixed effects are specified then the measurement locations must have better coverage of the entire range of radiometric potassium values (Fig. 10, centre). If these fixed effects are to be used to predict peat depths back-transformed to cm (Eq. (14)) then there is a greater focus on measuring at locations with small radiometric potassium values (Fig. 10, right). At these sites the peat depth is particularly uncertain.

When the spatially correlated random effects are also included in the predictions, the configuration of the measurement locations in space also controls the effectiveness of sample designs. In each of the examples presented in Fig. 11, the fixed effects are the nonlinear B-spline function of radiometric potassium. When the objective function requires accurate predictions of transformed peat depths (Eq. (13)) the measurement locations are relatively evenly distributed across the study region. When the objective function reflects the uncertainty in back-transformed units (Eq. (14)) the measurement locations are focussed where radiometric potassium is small since this is where peat depth is most uncertain. Some measurements are still required at locations with large radiometric

Table 2
10-fold cross validation results for maximum likelihood estimated models with different fixed effects. Predictions include estimated fixed effects but ignore impacts of spatial correlation. For each covariate, the model achieving the lowest AIC is used.

Criterion and unit	Covariate			
	None	Rad. K	LandSat-8	Altitude
Unit: Log cm				
Bias	−0.11	−0.01	−0.08	−0.09
Concordance	0.00	0.75	0.31	0.31
RMSE	0.86	0.53	0.73	0.76
Mean $\overline{\theta}$	1.00	0.98	0.98	0.92
Median $\widetilde{\theta}$	0.53	0.52	0.53	0.56
Unit: cm				
Bias	−8.06	−1.26	−11.30	−7.31
Concordance	0.00	0.76	0.23	0.27
RMSE	75.56	45.68	67.72	67.65

Table 3
10-fold cross validation results for maximum likelihood estimated models with different fixed effects. Predictions include estimated fixed effects and impacts of spatial correlation. For each covariate, the model achieving the lowest AIC is used.

Criterion and unit	Covariate			
	None	Rad. K	LandSat-8	Altitude
Unit: Log cm				
Bias	0.01	0.01	0.01	0.01
Concordance	0.83	0.85	0.83	0.83
RMSE	0.46	0.44	0.46	0.47
Mean $\overline{\theta}$	1.08	1.12	1.10	1.09
Median $\widetilde{\theta}$	0.40	0.38	0.40	0.37
Unit: cm				
Bias	2.04	1.90	1.28	1.72
Concordance	0.88	0.90	0.89	0.88
RMSE	34.15	32.50	33.24	34.32

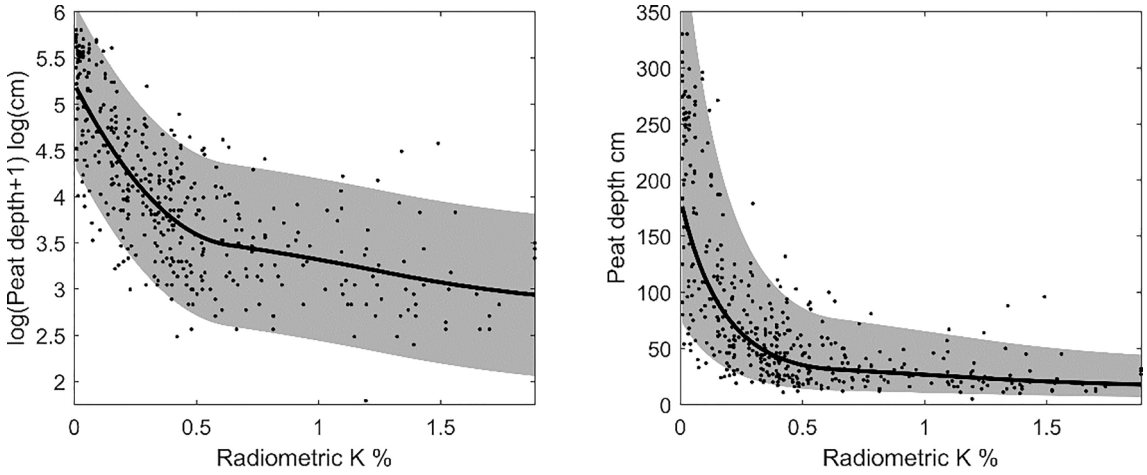

Fig. 7. The estimated third order B-spline fixed effects with five knots relating radiometric potassium to log (peat depth + 1) (left) and the back transformed relationship (right). The measured values are denoted by black dots and the 90% confidence interval by grey shading.

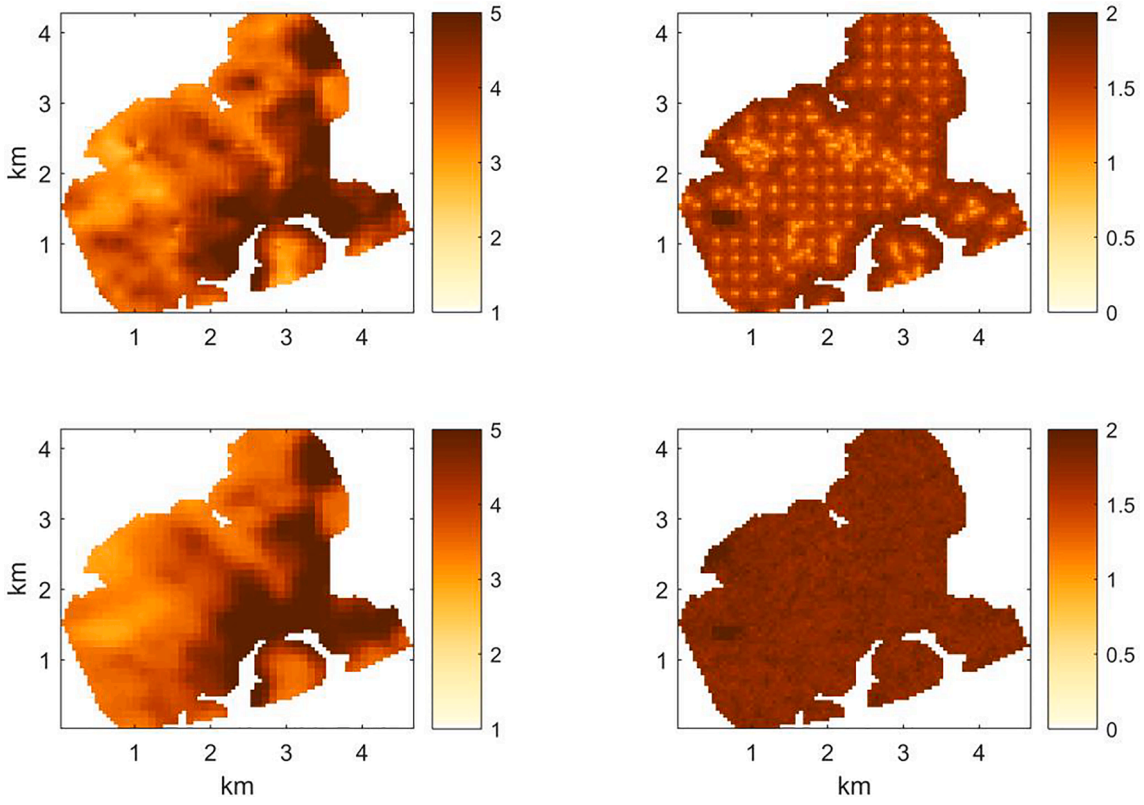

Fig. 8. Predictions of shifted and log transformed peat depth (log cm) in the south of the Dartmoor National Park. Top row uses linear mixed model predictor whereas bottom row uses linear model. Plots on the left are expected values and plots on the right are width of 90% prediction interval. Fixed effects are B-spline functions of radiometric potassium.

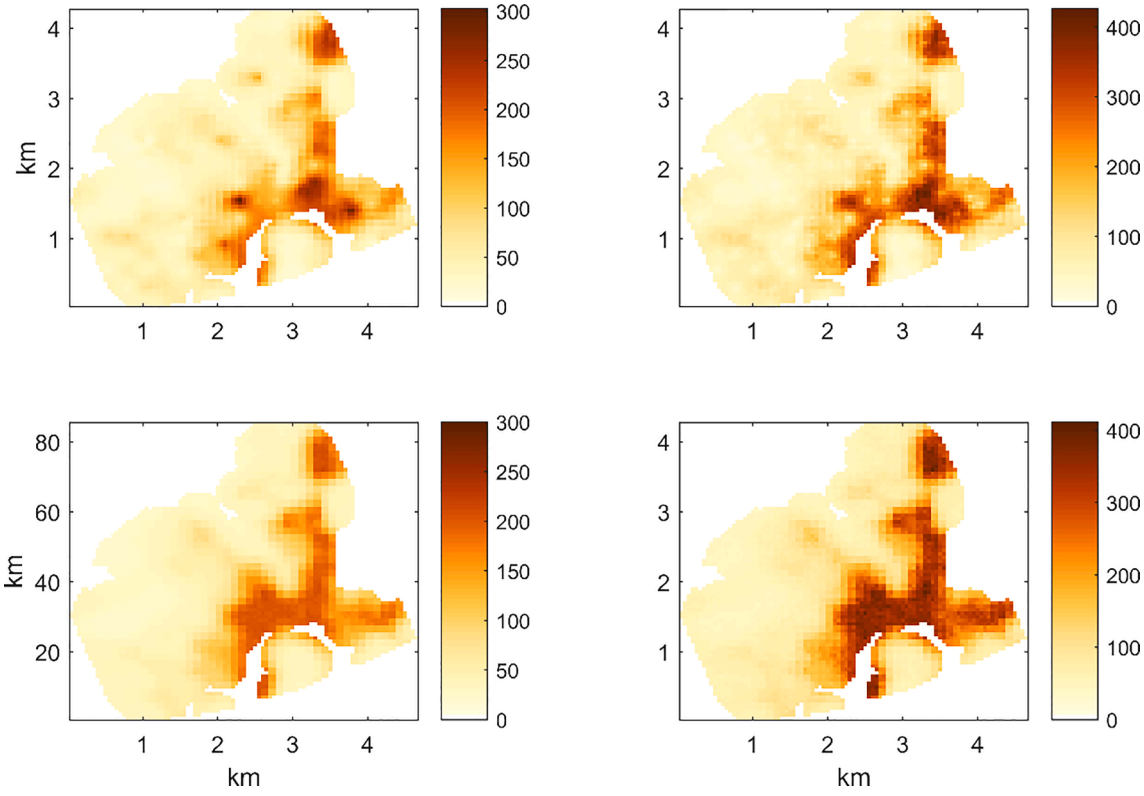

Fig. 9. Predictions of peat depth (cm) in the south of the Dartmoor National Park. Top row uses linear mixed model predictor whereas bottom row uses linear model. Plots on the left are expected values and plots on the right are width of 90% prediction interval. Fixed effects are B-spline functions of radiometric potassium.

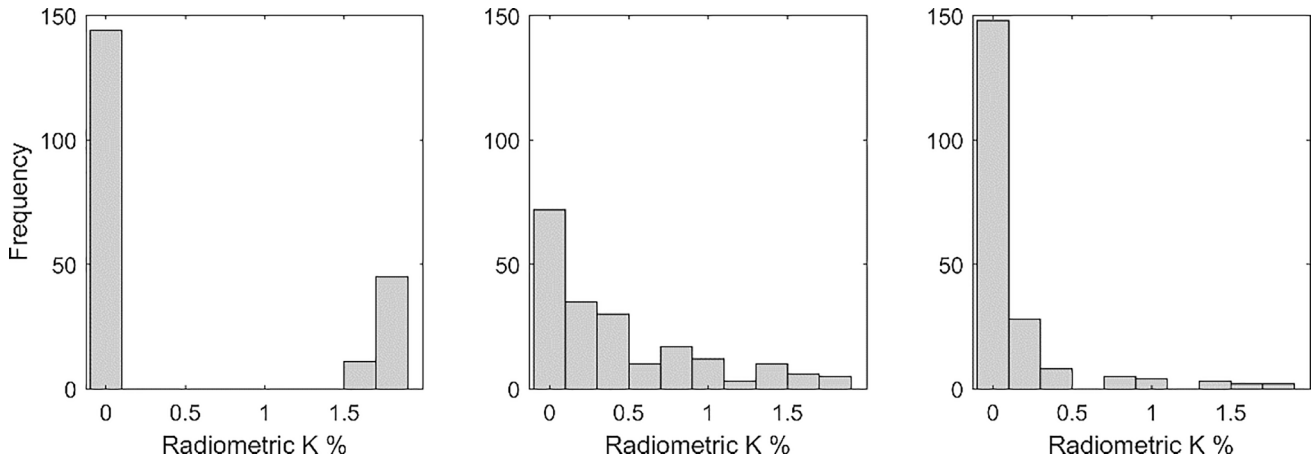

Fig. 10. Histograms of radiometric potassium values in optimized sample designs for linear model prediction of peat depth. Objective functions are (left) mean prediction variances in log cm when fixed effects are linear function of radiometric potassium; (centre) mean prediction variances in log cm when fixed effects are B-spline function of radiometric potassium; (right) mean prediction variances in cm when fixed effects are B-spline function of radiometric potassium.

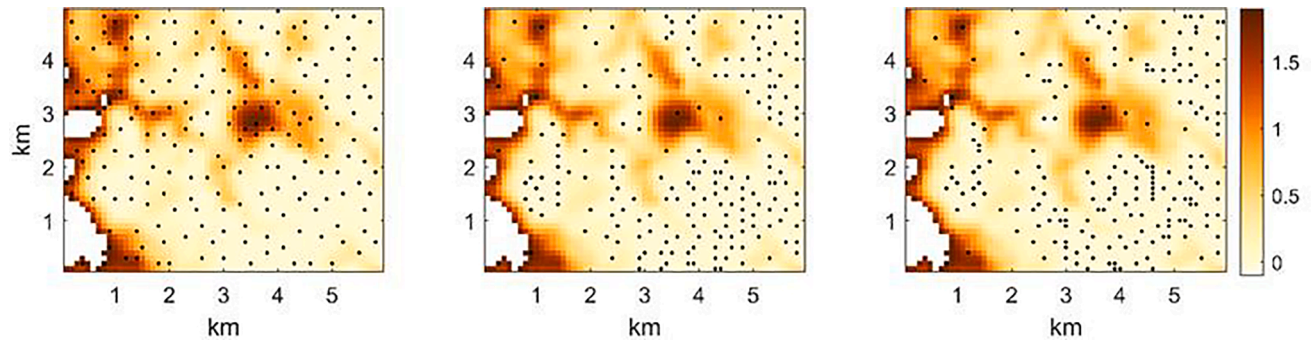

Fig. 11. Optimized 200 location sample design when objective function is (left) width of 90% prediction interval in log cm; (centre) width of 90% prediction interval in cm and (right) width of 90% prediction interval in cm including the impact of model parameter uncertainty. Sample designs are superimposed on BGS Tellus Survey interpolated radiometric potassium values (%).

potassium values to ensure that the fixed effects coefficients are accurately estimated. The optimized designs considered thus far ignore the effects of uncertainty in estimating the $\boldsymbol{\alpha}$ covariance function parameters. When this uncertainty is included in the objective function more close pairs of measurement locations are introduced (Fig. 11, right). These are required to accurately estimate the nugget parameter of the covariance function.

The effectiveness of sample designs optimized to minimize the uncertainty of predictions in back-transformed units and including covariance parameter uncertainty was tested using simulated data. Optimal designs based on different objective functions were used for each model as detailed in Section 4.3. Fig. 12 indicates that for optimized surveys consisting of 25 points, a linear mixed model where the fixed effects are a nonlinear function of radiometric potassium had an average RMSE of 76.9 cm. The average RMSE of the linear model with the same fixed effects was 75.6 cm for the same sample size. The linear mixed model has a marginally smaller RMSE because it is a simpler model and therefore the parameters can be estimated from this small sample with more certainty. Predictions from a linear mixed model with constant fixed effects had a mean RMSE of 92.6 cm compared to an RMSE of 106.0 cm for a linear model with constant fixed effects.

The average RMSEs for the linear mixed model decrease as the sample size is increased and are smaller than those for the linear model for all sample sizes greater than or equal to 50. This is an indication of the benefit of the spatially autocorrelated random effects. The mean RMSEs for the linear mixed model with constant fixed effects decrease more quickly than those for the nonlinear linear mixed model. This is

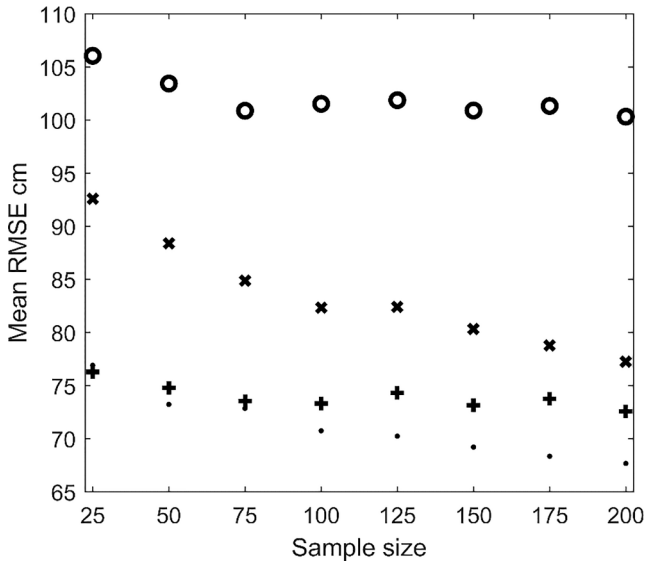

Fig. 12. Mean based on 100 sets of simulated data of RMSE for predictions of peat depth (cm) using optimized designs of between 25 and 200 samples. Peat depths are simulated using the linear mixed model where fixed effects are B-spline function of radiometric potassium. Results are shown for linear mixed model (dots) and linear model (+) predictions where fixed effects are B-spline function of radiometric potassium, for linear mixed model (x) and linear model (o) predictions where fixed effects are constant.

likely to be because of the greater autocorrelation amongst the random effects for this model (see Fig. 5). The mean RMSEs for the linear model with constant fixed effects decreased slowly to 100 cm for a sample of size 200.

5. Discussion

5.1. A framework for integrating sensor data into predictions of soil properties

The results illustrate how the integration of sensor data can lead to improved accuracy in spatial predictions of ground measured soil properties such as peat depth. The framework which delivered these improvements included a flexible linear mixed model where the fixed effects were a nonlinear function of the recorded sensor data. In contrast to the standard application of universal and regression kriging techniques (Webster and Oliver, 2007) the regression and covariance parameters were estimated simultaneously by maximizing the likelihood of the observed data arising according to the proposed model. This meant that the AIC values from different models could be meaningfully compared and used to determine the optimal level of smoothing in the fixed effects. Further flexibility could be added in the model by, for example, including outputs from multiple sensors or by allowing the variance of the ground measurements to vary, perhaps nonlinearly, according to covariates. The cost of this flexibility will be the additional computation resources required to estimate a large number of linear mixed models and determine the appropriate level of model complexity. This could be a major disadvantage of the approach if the models were to be estimated sequentially on a single computer but modern high performance computing facilities permit all the models to be estimated simultaneously.

The framework also included the use of the SSA algorithm to optimize the locations of the required ground measurements. A novel objective function was used which accommodated non-Gaussian random effects, nonlinear fixed effects and ensured that the survey was suitable to estimate both the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters of the linear mixed model. This lead to intuitively sensible survey designs which included close pairs of measurements to quantify short range variation, measurements being distributed over the range of covariate values to estimate the fixed effects and measurements being focussed in regions of the study area where the property was most uncertain.

A common drawback of optimized sampling approaches is that the objective functions require *a priori* knowledge of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters which are to be estimated from the eventual survey. Approximate values of these parameters must be assumed, perhaps based on surveys of the same soil property in similar circumstances. These assumed values should not greatly influence the final predictions since they will be re-estimated once the data have been collected. The optimal sampling approach is particularly efficient for adaptive (Marchant and Lark, 2006) or multiphase (Marchant et al., 2013) sampling where the design is adjusted after initial phases of measurements have been used to estimate the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters. Practitioners often assume that the fixed effects of spatial models are linear functions of a covariate. This assumption can lead to focussing sampling effort where the covariate values are particularly large or small. Our optimized designs based on nonlinear fixed effects lead to more even sampling across the range of values of the covariate so that deviations from the linear model can be identified.

The measurements of peat depth considered in this paper were heteroscedastic, their uncertainty or variance was larger for larger peat depths. This heteroscedasticity was accommodated in the model via the logarithmic transform and backtransform. When surveys were designed to minimise errors in the original units, these transforms led to a greater density of measurements where the peat depth were expected to be large. A similar effect was observed by Wadoux et al. (2017), when they accommodated heteroscedasticity of rainfall measurements by permitting the variance of the rainfall to be related to topographic covariates.

5.2. The usefulness of radiometric potassium data in predicting peat depths

The results of this study confirm the usefulness of including radiometric information when mapping peat depths. Linear mixed models that include radiometric potassium values in the fixed effects better fit the measured peat depths from Dartmoor than models which assume constant fixed effects or fixed effects that are related to the other covariates compiled by Kirkwood et al. (2016). The fixed effects based on radiometric potassium also lead to superior cross-validation statistics relative to the other models.

Linear mixed models consistently have smaller RMSEs than linear models because of the benefit of including the spatially autocorrelated random effects in the prediction. However, the benefit of including the radiometric potassium information is smaller for the linear mixed models in comparison to the linear models. Also, the range of the spatial autocorrelation amongst the random effects is much smaller for the linear mixed model including radiometric potassium than for the linear mixed model with constant fixed effects. These points indicate that the radiometric potassium is explaining a substantial proportion of the variation that could otherwise have been explained by collecting additional ground measurements of peat depth and then using the spatial autocorrelation to predict at other locations. Inclusion of the radiometric information therefore reduces the number of ground based measurements of peat depth that are required to achieve a specified degree of accuracy in the maps.

In the experiments using simulated data to predict peat depth on a 277 000 m^2 sub-region of Dartmoor National Park, the linear mixed and linear models including fixed effects which were nonlinear functions of radiometric potassium had similar RMSEs when 50 or fewer ground measurements were made. For these small sampling densities there is little benefit from spatial autocorrelation in the predictions and the linear mixed model requires some of the sampling effort to be spent on estimating the random effect parameters of the model. As the sample size is increased the benefit of the linear mixed model becomes more evident.

The linear mixed model with constant fixed effects has larger RMSEs than the linear mixed model that includes radiometric potassium in the fixed effects for all numbers of ground based samples. When 25 ground measurements are made then RMSEs for the constant fixed effects are 20% larger. For 200 ground measurements the RMSEs for constant fixed effects are only 14% larger than the corresponding results including radiometric potassium but still larger than those achieved from making 25 ground measurements and including radiometric potassium in the model.

The uncertainty in the estimated relationship between radiometric potassium and peat depth is largest for the deepest peats when the radiometric signal is smallest (Fig. 7). This finding reflects the approximate log-Gaussian distribution of the measured data. It is consistent with the conclusions of Beamish (2014) who suggested that 60 cm of 80% saturated peat was sufficient to absorb 90% of the radioactive signal from the underlying signal. This implies that a small radiometric potassium value could indicate any peat depth greater than 60 cm. It does appear that radiometric potassium data can be used to classify where peat depth is likely to be greater or less than 60 cm and is consistent with the application of radiometric potassium data by the Finnish Geological survey (Airo et al., 2014).

The survey designs optimized to predict peat depth in cm focus measurements where radiometric potassium is smallest and peat depth most uncertain. Young et al. (2018) also suggested that peat depth measurements should be focussed where the peats were deep although they inferred the locations of the deepest peats from topographical parameters. The relationship between peat depth and radiometric

potassium is likely to vary across the wider landscape. For example, the magnitude of the radioactive source signal is likely to vary according to the underlying geology and the rate of attenuation of this radioactive signal with peat depth is likely to vary according to the moisture content or other physical characteristics of the peat. Therefore, it is advisable to re-estimate the parameters of the linear mixed model for different case studies. The optimized sampling approach described in this paper ensures that the correct proportions of the ground measurements are located to aid the estimation of the different model parameters and to interpolate from.

6. Conclusions

The relationships between sensor data and soil properties can be represented within linear mixed models leading to accurate spatial predictions of the soil property that require fewer ground measurements. Greater flexibility can be included in these models by using B-splines to permit nonlinear fixed effect relationships. The locations of the ground measurements can be chosen using an optimized sampling algorithm leading to greater efficiencies. Simulated data, based on the pattern of peat depth variation in Dartmoor, are used to demonstrate that this analysis and survey design framework can lead to more cost-effective surveys of peat depth.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the NERC project NE/T004169/1 as part of the Landscapes Decisions Programme and by EPSRC as part of the Senior Fellowship in the Role of Digital Technology in Understanding, Mitigating and Adapting to Environmental Change grant no: EP/P002285/1. This paper is published with the permission of the Executive Director, BGS. The author is grateful to David Beamish (BGS), Christoph Kratz (Natural England) and Dylan Young (Leeds University) for informative discussions.

References

Airo, M.-L., Hyvönen, E., Lerssi, J., Leväniemi, H., Ruotsalainen, A., 2014. Tips and tools for the application of GTK's airborne geophysical data. In: Geological Survey of Finland, Report of Investigation. vol. 215. pp. 33.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Budapest, Akadémiai Kiadó, pp. 267–281.

Beamish, D., 2013. Gamma ray attenuation in the soils of Northern Ireland, with special reference to peat. J. Environ. Radioact. 115, 13–27.

Beamish, D., 2014. Peat mapping associations of airborne radiometric survey data. Remote Sens. 6, 521–539.

Beamish, D., Howard, A.S., Ward, E.K., White, J., Young, M.E., 2014. Tellus SouthWest Airborne Geophysical Data. Natural Environment Research Council, British Geological Survey.

Beamish, D., 2015. Relationships between gamma-ray attenuation and soils in SW England. Geoderma 259–260, 174–186.

Survey, B.G., et al., 1968. GB Land Gravity Survey. Br. Geol. Surv.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86–95.

Dissanska, M., Bernier, M., Payette, S., 2009. Object-based classification of very high resolution panchromatic images for evaluating recent change in the structure of patterned peatlands. Can. J. Remote Sens. 35, 189–215.

Finlayson, A., Marchant, B.P., Whitbread, K., Hughes, L., Barron, H.F., In press. Soil Use and Management, https://doi.org/10.1111/sum.12596.

Fyfe, R.M., Coombe, R., Davies, H., Parry, L., 2013. The importance of sub-peat carbon storage as shown by data from Dartmoor, UK. Soil Use Manag. 30, 23–31.

Gatis, N., Luscombe, D.J., Carless, D., Parry, L.E., Fyfe, R.M., Harrod, T.R., Brazier, R.E., Anderson, K., 2019. Mapping upland peat depth using airborne radiometric and lidar survey data. Geoderma 335, 78–87.

van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

Technologies, I., 2007. NEXTMap British Digital Terrain Model Dataset Produced by Intermap. NERC Earth Measurement Data Centre.

Keaney, A., McKinley, J., Graham, C., Robinson, M., Ruffell, A., 2013. Spatial statistics to estimate peat thickness using airborne radiometric data. Spatial Statistics 5, 3–24.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61.

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, 787–799.

Lark, R.M., Marchant, B.P., 2018. How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters? Geoderma 319, 89–99.

Lin, L.-I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.

Met Office, 2010. In: Princetown 1971-2000 averages. http://www.metoffice.gov.uk/climate/uk/averages/19712000/sites/princetown.html.

Marchant, B.P., 2018. Model-based geostatistics. In: McBratney, A.B., Minasny, B., Stockmann, U. (Eds.), Pedometrics: A system of quantitative soil information. Springer.

Marchant, B.P., Lark, R.M., 2006. Adaptive sampling and reconnaissance surveys for geostatistical mapping of the soil. Eur. J. Soil Sci. 57, 831–845.

Marchant, B.P., Lark, R.M., 2007. Optimized sample schemes for geostatistical surveys. Math. Geol. 39, 113–134.

Marchant, B.P., McBratney, A.B., Lark, R.M., Minasny, B., 2013. Optimized multi-phase sampling for soil remediation surveys. Spatial Statistics 4, 1–13.

Minasny, B., Berglund, O., Connolly, J., Hedley, C., de Vries, F., Gimona, A., Kempen, B., Kidd, D., Lilja, H., Malone, B., McBratney, A., Roudier, P., S., O'Rourke, Rudiyanto, Padarian, J., Poggio, L., ten Caten, A., Thompson, D., Tuve, C., Widyatmanti, W. (2019). Digital mapping of peatlands a critical review. Earth Science Reviews, 196, 102870.

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

Parry, L.E., Charman, D.J., 2013. Modelling soil organic carbon distribution in blanket peatlands at a landscape scale. Geoderma 211–212, 75–84.

Parry, L.E., Charman, D.J., Noades, J.P.W., 2012. A method for modelling peat depth in blanket peatlands. Soil Use Manag. 28, 614–624.

Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation—an example from Scotland. Geoderma 232, 284–299.

Ravi Shanker, D., 2017. Remote Sensing of Soils. Springer, Berlin/Heidelberg.

Roy, D.P., Wulder, M., Loveland, T., Woodcock, C., Allen, R., Anderson, M., Helder, D., Irons, J., Johnson, D., Kennedy, R., 2014. Landsat-8: science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.

Stolte, J., Tesfai, M., Øygarden, L., Kværnø, S., Keizer, J., Verheijen, F., Panagos, P., Ballabio, C., Hessel, R., 2015. Soil threats in Europe: Status, Methods, Drivers and Effects on Ecosystem Services. A Review Report, Deliverable 2.1 of the RECARE Project; Office for Official Publications of the European Community: Luxembourg, Vol. EUR 27607, 69–78.

Wadoux, A.M.J.C., Brus, D.J., Rico-Ramirez, M.A., Heuvelink, G.B.M., 2017. Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Adv. Water Resour. 107, 126–138.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. Wiley, Chichester, UK.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, 2nd Edition. Chapman & Hall/CRC, Boca Raton.

Young, D.M., Parry, L.E., Lee, D., Ray, S., 2018. Spatial models with covariates improve estimates of peat depth in blanket peatlands. PLoS ONE 13 (9), e0202691.

Zhu, Z., Stein, M.L., 2006. Spatial Sampling Design for Prediction with Estimated Parameters. J. Agric. Biol. Environ. Stat. 11, 24–44.